npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01416-x

Scaling convolutional neural networks achieves expert level seizure detection in neonatal EEG

Robert Hogan[1] ✉, Sean R. Mathieson[1,2], Aurel Luca[1], Soraia Ventura[1,2], Sean Griffin[1], Geraldine B. Boylan[1,2,3] & John M. O'Toole[1] ✉

Neonatal seizures require urgent treatment, but often go undetected without expert EEG monitoring. We have developed and validated a seizure detection model using retrospective EEG data from 332 neonates. A convolutional neural network was trained and tested on over 50,000 hours ($n = 202$) of annotated single-channel EEG containing 12,402 seizure events. This model was then validated on two independent multi-reviewer datasets ($n = 51$ and $n = 79$). Increasing data and model size improved performance: Matthews correlation coefficient (MCC) and Pearson's correlation (r) increased by up to 50% (15%) with data (model) scaling. The largest model (21m parameters) achieved state-of-the-art on an open-access dataset (MCC = 0.764, $r = 0.824$, and AUC = 0.982). This model also attained expert-level performance on both validation sets, a first in this field, with no significant difference in inter-rater agreement when the model replaces an expert ($|\Delta\kappa| < 0.094$, $p > 0.05$).

The most frequent cause of neonatal seizures is acute brain injury in the early postnatal period. Seizures typically emerge over the first 72 postnatal hours in term neonates, primarily caused by hypoxic-ischaemic encephalopathy (HIE) or cerebrovascular injury[1–3]. More than half of neonates with moderate or severe HIE develop seizures[1,3,4]. For those neonates who do develop seizures, approximately 7% to 10% are at risk of death and 23% to 50% are at risk of poor outcome[1,5,6].

Seizures can be subtle, often without clinical correlate, and often remain undetected[7]. Continuous electroencephalogram (EEG) monitoring is the gold standard for neonatal seizure surveillance. Yet real-time interpretation of the EEG requires specialised expertise that is not always available, limiting the capacity for continuous review of EEGs. A recent multi-centre study found that, even with continuous EEG or amplitude-integrated EEG readily available, only 11% of seizures were treated within 1 hour of onset[4]. Prompt treatment can reduce seizure burden and therefore may reduce seizure-mediated neuronal damage and improve outcomes[8].

Automated review of the EEG, with expert oversight, would allow for increased monitoring of at-risk neonates. A recent clinical trial of an automated algorithm to detect EEG seizures demonstrated the potential clinical utility[9]. Yet this seizure detection algorithm, which was developed in 2011[10], has been comprehensively surpassed in performance by a range of newer methods[11–19]. Most of these contemporary seizure detection methods use deep neural networks. These powerful tools offer increased performance over feature-based machine learning methods by enabling end-to-end learning, from raw EEG to label class, and by scaling performance with increasing model size and training data. We have identified 4 key challenges in the current literature that may be constraining performance.

First, many deep-learning models are trained with small datasets, a significant limitation in this field[11,13–17,19]. A widely used open-access dataset contains 112 h of EEG recordings from 79 neonates[20], although in many cases, only a subset of 39 neonates with seizures is used[15–17,19]. Second, most methods use global, and not per-channel, seizure annotations[11–19]. This enables faster annotation of the EEG but provides less detailed information for model training. Additionally, this can make the models susceptible to variations in the EEG montage. Third, most methods use a relatively small network architecture, with fewer than 50k parameters[11–19]. This may limit the extent to which a model can capture the complexity of the data. Fourth, validation of models on held-out datasets is frequently omitted, making it difficult to determine how the models would perform on new, unseen data[11,15–17,19].

In this study, we aim to address these limitations. Our primary goal is to develop a deep-learning model capable of detecting seizures in neonatal EEG with accuracy suitable for clinical application. To this end, we test the hypothesis that increasing both model size and training data will improve performance. Our models are based on a modern convolutional neural network architecture and are trained to detect seizures on a per-channel basis. Additionally, we validate our models on independent, held-out datasets from Cork and Helsinki to determine efficacy on unseen data.

[1]CergenX Ltd, Dublin, Ireland. [2]INFANT Research Centre, University College Cork, Cork, Ireland. [3]Department of Paediatrics and Child Health, University College Cork, Cork, Ireland. ✉e-mail: rhogan@cergenx.com; jotoole@cergenx.com

Table 1 | Model variants explored in this work

model	depth (*D*)	width (*W*)	parameters (count)	computation (FLOP)
Nano	1	1	38.7 k	1.9 m
Small	2	2	289.2 k	14.4 m
Medium	3	4	1.7 m	84.3 m
Large	3	8	6.7 m	335.4 m
Extra Large	6	10	20.6 m	1 G

Depth and width parameters (*D*, *W*) were selected to generate models with an approximate logarithmic scale in parameter count across 3 orders of magnitude.
Key: FLOP, floating point operations.

Results

Model and data scaling

We evaluated a wide range of model scales, as described in Table 1, from the 39k parameter Nano variant up to the >500 times larger 21m parameter Extra-Large (XL) model and find significant performance improvements. Figure 1b illustrates these improvement gains in Matthews correlation coefficient (MCC), correlation, and error rate suggesting that model scaling is indeed a viable path to better models for neonatal seizure detection. A representative sample of the model output in Fig. 2b gives a qualitative sense of this performance improvement.

One notable feature is a large drop in performance for the first 10-times scaling from the Nano to Small model. This phenomenon has been observed repeatedly in other applications and is known as *deep double descent*[21]. We also see some evidence of this in data scaling in Fig. 1a from 1k to 10k hours of EEG.

Figure 1c presents results for held-out validation sets across different model scales. As these datasets only have global annotations across all channels, we take the maximum over the per-channel outputs to produce a global prediction. This simplification may obscure some of the per-channel performance differences in the models. Nevertheless, the power-law trend of improvement with model scale is clear across many metrics, displaying a strong validation of the scaling hypothesis. We also find the appearance of the double-descent dip here again, although less pronounced, across several metrics on both datasets.

We do, however, see indications of diminishing returns for some metrics for the Large and XL models on the Helsinki dataset. We investigate this further later in this section.

We also quantify the effect of increasing dataset size with random subsampling by (1) EEG segment and (2) neonate (keep all segments for the sampled neonates, and drop all others). We do so by training a Medium model with (sub-samples of) 80% of the development dataset and test on the same left-out 20%. We find that in both cases there is significant performance gains, up to 50%, from scaling the data—illustrated in Fig. 1a. Adding more EEG segments improves performance, even for datasets >20 times larger than the nearest published work, indicating that scaling data remains a powerful lever for improving models.

Model performance

Table 2 presents a comprehensive evaluation of the XL model across the 3 datasets. The results of the test set are evaluated per channel. Combining across all channels to form a global annotation increases detection performance: for example, AUC increases from 0.978 to 0.988 and MCC increases from 0.648 to 0.703. In Table 3 we also include the limited set of metrics available for direct comparison to the literature. Despite our relatively simplistic approach to translating from per-channel to global predictions, we still find that our models compare quite favourably to those published in the literature. This is true even for models that have been trained on the Helsinki dataset and report a cross-validation result.

Additionally, to assess performance on a per-neonate level we analyse the XL model's ability to estimate seizure burden. Table 4 shows that on the Cork validation set, the model's estimate of seizure burden has no statistically significant difference to that determined by the consensus of experts. On the Helsinki dataset the seizure burden was underestimated, but on both datasets the median difference is small enough that it is unlikely to have clinical relevance. Finally, we also assess performance on neonates without seizure and find median values of ≤0.01 mins/h on both datasets, confirming the low false detection rates of the model.

The XL model attains expert-level equivalence on both Cork and Helsinki validation datasets. In both cases, the change in agreement by replacing a human expert with the AI model predictions was consistent with 0: $\Delta\kappa = -0.094$ (95% CI: −0.189, 0.005) for Cork and $\Delta\kappa = -0.082$ (−0.156, 0.002) for Helsinki. For the Helsinki dataset, the Medium model also reaches this benchmark and the Large model is just narrowly rejected but neither model achieves this benchmark on the Cork validation dataset. For the smaller models, such as the Nano and Small, this benchmark is well out of reach ($p < 0.001$). Results for all models are presented in Table 5.

Event duration analysis

Figure 3 presents the distribution of model performance for increasing seizure event durations. We find that for long seizures (>300 s) the model performs well, with a detection rate of 100%. Most of the missed events are for short seizures (<30 s). The difficulty with short seizures has a more pronounced effect on the Helsinki dataset where they were more commonly annotated.

Distribution shift

Although we find strong scaling performance with model size for the Cork validation set, Fig. 1c indicates diminishing returns for the Large and XL models on the Helsinki validation set. For those models, we find that an optimal classification threshold shifts down from 0.5 to 0.4 and 0.3 respectively. This may indicate that the larger, more capable models are learning some features that are useful on training sets but may not generalise to all settings.

One hypothesis for why we observe this effect in the Helsinki dataset but not in our training data or the Cork validation set could be due to differences in clinical protocols applied in different centres. The most obvious difference is that almost 50% (38/78) of the neonates have had EEG recorded ≥1 week after birth, in contrast to the Cork validation set which were all within a week of birth. If we divide the Helsinki dataset into two groups, those with EEGs recorded within a week (early-EEG group) and those with EEG recorded after the first week of life (late-EEG group), we find significant differences in primary diagnosis. A primary diagnosis of either asphyxia (including HIE) or stroke accounts for 92% (32/37) in the early-EEG group compared with just 32% (10/31) in the late-EEG group, $p < 0.001$ ($n = 68$; Fisher exact test). This may not be unexpected as the suspected diagnosis would likely be the main driver for EEG monitoring.

With this division of the dataset, we find a remarkable concordance between this explanation of the distribution shift and the scaling behaviour for these cohorts. In Fig. 4 we show that for the early-EEG group the same scaling behaviour observed in both Cork datasets is recovered. In contrast, for the late-EEG group, we see that the performance peaks at the Medium model and starts to degrade progressively for the Large and XL models. This is suggestive that these more capable models are indeed learning something specific about EEG, which may be related to the primary diagnosis or to postnatal age.

Montage robustness

A feature of our seizure detection model is its independence to channel montage, both the number of channels and the type of montage. To investigate this robustness, we take our predictions on the Helsinki dataset and simulate data loss or montage changes by randomly inserting contiguous sections of zeros in the per-channel model output. The final prediction is still calculated as the maximum over all channels so this dropped data will not contribute to the global estimate. We drop 10%, 25%, 50%, and

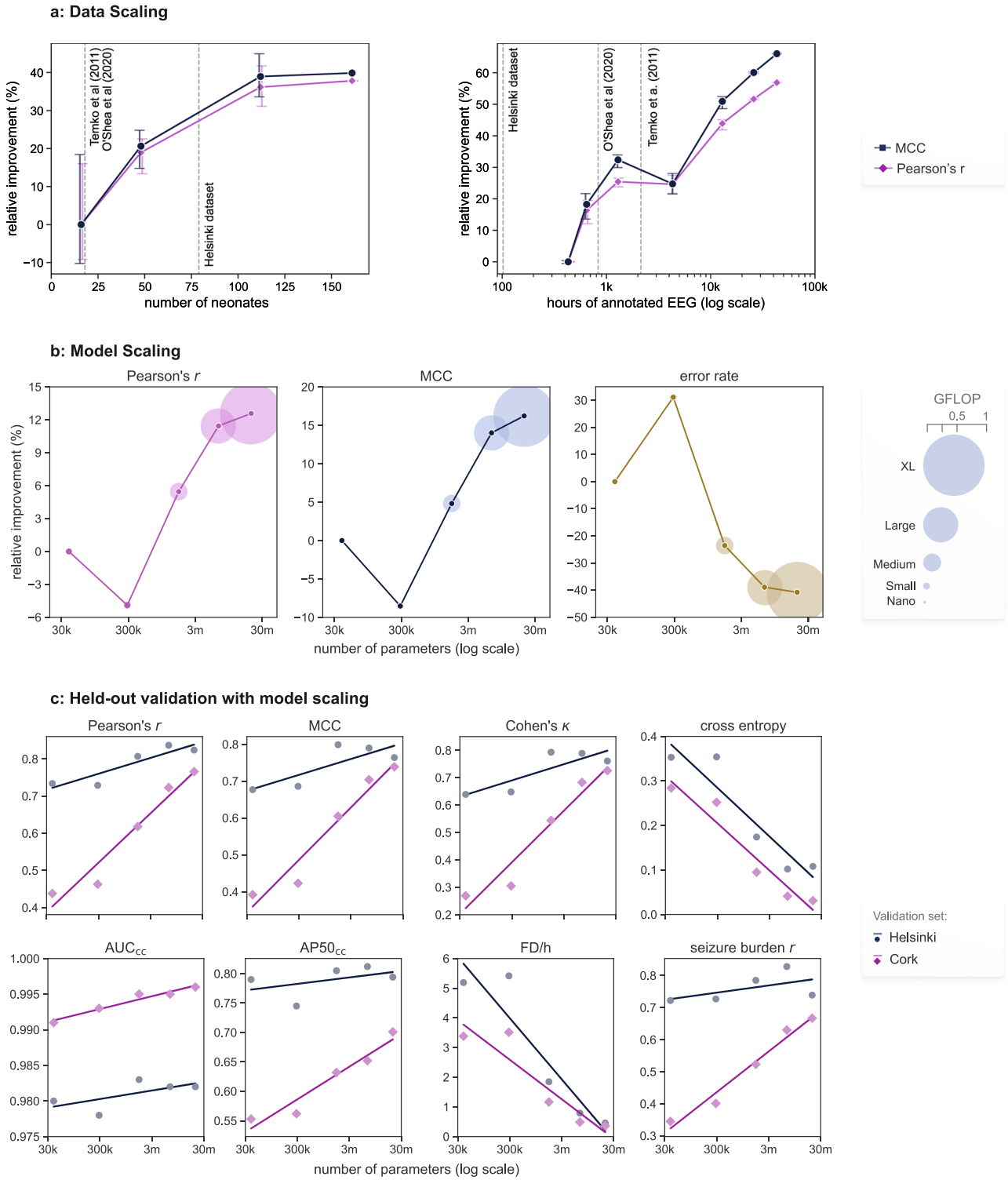

Fig. 1 | Scaling training data and model size yields approximate power-law performance gains. Metrics calculated at the segment level for 20% (41/202 neonates) of the development dataset in (**a**) and (**b**) and neonate level for held-out validation datasets in (**c**). **a** Performance improves with increasing number of neonates and hours of annotated EEG (counted per channel) in training set; error bars denote min/max over 3 trials. Prominent datasets from the literature are included for comparison[10,13,20]. **b** Scaling model size over 3 orders of magnitude reveals typical deep-double descent pattern with a performance dip for the Small model before recovering for larger models. Marker size indicates computational cost in giga floating- point operations (GFLOPs). **c** Scaling model size on the held-out datasets from Cork and Helsinki. We include a linear fit to illustrate the predictability of performance increase. See Table 8 for description of metrics used.

100% of the channel data at random in contiguous segments; here 100% is equivalent to dropping the channel. This was applied across increasing numbers of channels until all but 1 were affected. This procedure was repeated for 20 trials.

The result of this experiment is shown in Fig. 5, where we summarise the impact as % degradation relative to zero data loss for both the AUC and MCC metrics. We find that the model is remarkably robust: by dropping one-half of the channels the AUC (MCC) degrades by only 1.4% (7.0%). If

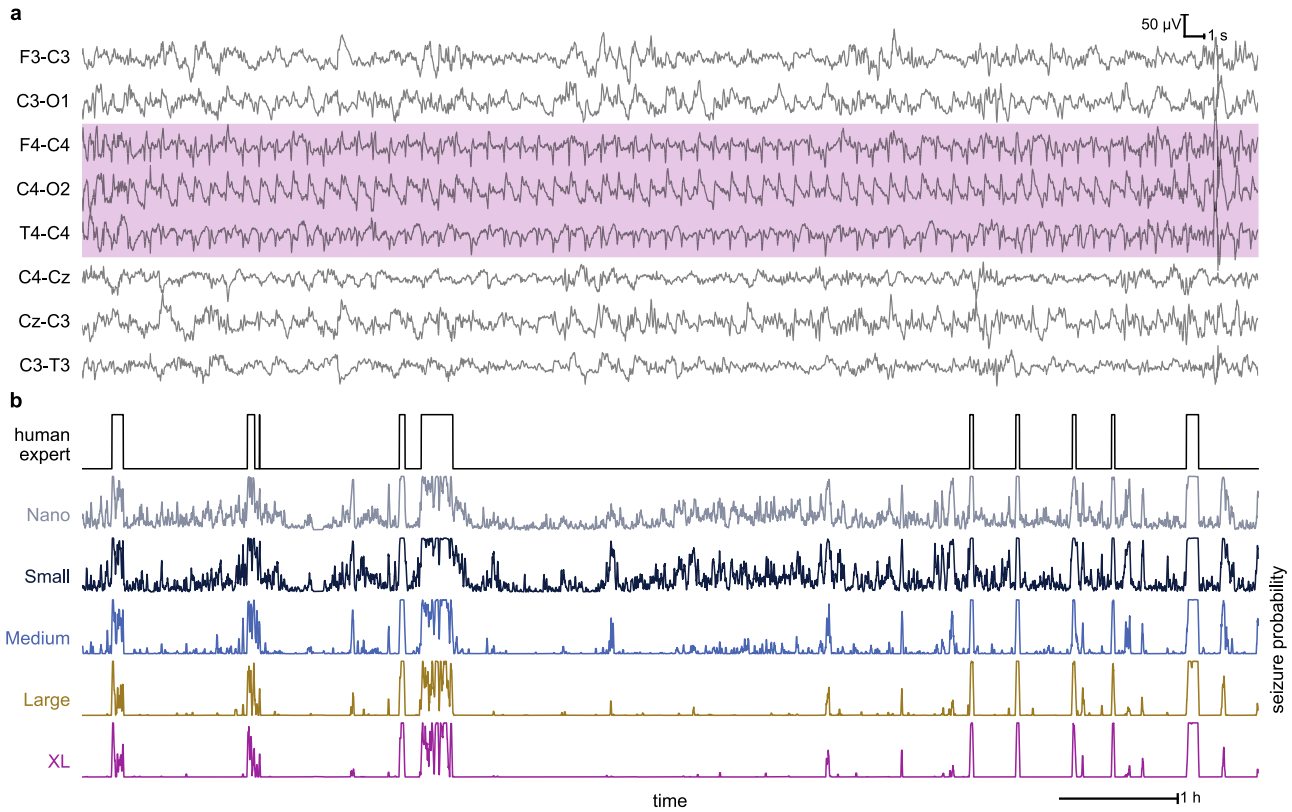

Fig. 2 | Segment of EEG with seizure and comparison of different model outputs. **a** Sixty second sample of EEG, from the development dataset, with per-channel seizure annotations shaded. In this example, only 3/8 channels contain seizure. **b** Annotation and model outputs for 10 h from C4-O2 of the same EEG recording. The EEG sample in (**a**) corresponds to the first 60 s of the first seizure event in (**b**). Models of different scales---namely the Nano, Small, Medium, Large, and Extra Large (XL) models---become more confident, suppressing the output for non-seizure periods while maintaining high agreement in seizure periods. This ease of interpretation would be beneficial to clinical implementation that use a real-time model output trace.

Table 2 | Performance of the XL model on 3 datasets

	Test Set	Validation Sets	
	Cork (n = 41) per-channel	Cork (n = 51) global channel	Helsinki (n = 79) global channel
AUC	0.978	0.996	0.982
AP / AP50	0.694 / 0.533	0.833 / 0.701	0.891 / 0.794
Pearson's r	0.723	0.766	0.824
MCC	0.648	0.739	0.764
Cohen's κ	0.630	0.726	0.761
Sensitivity/ Specificity (%)	51.5 / 99.9	88.9 / 99.2	72.9 / 98.5
PPV/NPV (%)	82.0 / 99.6	62.0 / 99.8	84.9 / 96.8
FD/h	0.053	0.363	0.459
Seizure Burden, r	0.902	0.667	0.739

Testing results are from 20% of the development dataset. Validation results are from held-out datasets from Cork and Helsinki (described in Table 6.) Performance is assessed per-channel on the test dataset and globally (across all channels) on the validation datasets. All metrics are calculated by concatenating all EEG recordings. Metrics for the held-out (validation) multi-annotator sets are based on unanimous consensus annotations.

Key: AUC, area under the receiver-operator-characteristic curve (AUC); AP, average precision; AP50, average precision with recall > 50%; MCC, Matthews correlation coefficient; PPV, positive predictive value; NPV, negative predictive value; FD/h, false detections per hour; r represents correlation; κ represents kappa.

the data loss is partial, we see even stronger results; for example, dropping 25% from 17/18 channels we see only a 0.5% (3.0%) drop in AUC (MCC). The upper bound on performance here is of course determined by whether there is sufficient information remaining in the data to recover the global annotation even in principle, a dependence of the spatial distribution of the seizure event.

Discussion

We have developed a state-of-the-art convolutional neural network for neonatal seizure detection, improving substantially upon previously published results. We also have verified our hypothesis that scaling is a hitherto underutilised lever for performance improvement in neonatal EEG analysis. Scaling both the dataset size, by neonate and by duration of EEG, yielded up to 50% increases in MCC. Scaling model sizes similarly delivered significant performance improvements of up to 15% in MCC. The result of these improvements is that our best model, the 21m parameter XL variant of the ConvNeXt architecture, attains expert-level equivalence with the EEG experts on two independent, fully held-out validation sets ($\Delta\kappa \neq 0$ rejected with $p > 0.05$).

Much of the literature focuses on methodological improvements, with specialised architectures trained on very small datasets yielding incremental gains[15–17,19]. Our work challenges this approach and suggests a more promising path to expert-level models is through data and model scale. A key part of the model scaling strategy is designing an architecture with computational efficient scaling. Failure to do so can lead to prohibitively expensive training iterations. Scaling the fully-convolutional neural network model[13], for example, to an equivalent size of the XL model would require >6 times the computational load.

Our scaling results also challenge the conventional wisdom that increasing model size will eventually lead to overfitting and decreased generalisation performance. Indeed to date, most research in neonatal EEG has focused on relatively small models, with <50k parameters[11–19]. Despite this, model scaling well past the point of over-parameterisation has been a key feature of recent AI progress[21–23]. This observation that performance will

Table 3 | Comparison of proposed model and other published models tested on the Helsinki dataset

	all data ($n = 79$)	seizure only ($n = 39$)		
	AUC_{cc}	AUC (median [IQR])	Cohen's κ	FD/h
XL ConvNeXt (ours)	**0.982**	**0.996 (0.975 - 1.000)**	0.800	**0.34**
FcCNN[13]	0.956	-	-	-
ResNet[14]	0.964	-	-	-
SVM–Cork[10]	-	0.961 (0.869 – 0.990)	-	1.00
SVM–Helsinki[40a]	0.955	0.988 (0.931 – 0.998)	-	0.86
GAT[19a]	-	0.993 (0.964 – 0.995)	**0.880**	0.86

Our extra-large (XL) model achieves a new state-of-the-art on most metrics, even outperforming models trained directly on the Helsinki data. In keeping with published methods, all but the concatenated AUC (AUC_{cc}) is evaluated on the subset for neonates ($n = 39$) with seizures. A more complete set of metrics on all 79 neonates is shown in Table 2.

[a] Leave-one-out (LOO) testing result using the Helsinki dataset.

Key: IQR, interquartile range; AUC, area under the receiver-operator-characteristic curve; FcCNN, fully convolutional neural network; SVM, support vector machine; GAT, graph attention network; FD/h, false detections per hour.

Table 4 | Performance of extra-large (XL) model estimating seizure burden

		Seizure burden (mins/h)	
		Cork	Helsinki
Seizure	Expert	0.94 (0.13, 5.93)	6.94 (0.16, 52.13)
	XL model	1.46 (0.11, 6.90)	4.55 (0.0, 42.83)
	Δ (model - expert)	0.03 (−0.45, 5.67)	−1.32 (−22.86, 0.83)[a]
Non-seizure	XL model	0.01 (0.00, 1.27)	0.00 (0.00, 6.12)

For each neonate seizure burden is calculated as the mean minutes of seizure per hour of EEG. The table presents median (interquartile range) seizure burden over neonates in both the Cork and Helsinki held-out validation sets using the experts' annotations and the model predictions. Δ shows the model's predicted seizure burden minus experts' seizure burden for neonates with seizures. Experts' annotation is derived from the consensus annotation. We also include the predicted seizure burden for babies with no consensus seizures.

[a] Denotes statistically significant difference using Wilcoxon signed-rank test.

initially decline before improving with scaling is known as deep-double descent and was found to occur across a range of tasks, model architectures, and optimisation methods[21]. Figure 1b illustrates this finding in all metrics with a decrease in performance for the Small model comparative to the smaller Nano model. We also see indications of this in data scaling (Fig. 1a), where increasing the size of the training dataset actually decreases performance before improving again with more data. This surprising finding is a corollary of the deep-double descent effect on model scale and was also observed elsewhere[21]. If operating in a narrow scale range, on the left-hand side of the double-descent dip, it is understandable that smaller models and datasets would seem optimal (as found in other studies[13]). However, an exploration of a much larger scale range, as we show here, yields substantial benefits by moving past the double-descent trap.

A limitation in the neonatal seizure detection literature is that AUC is almost always presented as the lead—and often only—performance metric[10–19]. This metric can be misleading for many reasons[24–26]. For example, with large class imbalance, as is the case for electrographic seizures, false positives are obscured. To illustrate this, our worst performing model (Nano) has an AUC of 0.980 on the Helsinki set, exceeding the best reported value of 0.964[14]. Our XL model improves on this only slightly to 0.982 but has approximately 10 times fewer FD/h and achieves expert-level agreement on both held out datasets. The Nano model, in contrast, is far from achieving expert-level agreement: $\Delta\kappa$ is approximately 5-times (2-times) larger on Cork (Helsinki) validation datasets.

Addressing this limitation, we present a comprehensive set of metrics for continuous and binary variables, including more balanced measures of performance, such as MCC, Pearson's r, and Cohen's κ[24–27], in addition to metrics with more clinical relevance, such as FD/h, correlation with seizure burden, and expert-level equivalence testing. We have developed an open-source framework for metric calculation to assist with transparency in reporting of performance for this field.

We have also highlighted the utility of developing models with per-channel annotations, making the algorithm adaptable to different clinical montage requirements or protocols. Figure 6 illustrates the heterogeneous time-varying nature of seizure focus among EEG channels. As a result, global labels will obfuscate important channel differences, similar to injecting noise into the training data. Although global labels, or weak labels[13], are easier to annotate, they present only summary information without detail and therefore fail to maximise the full potential of the valuable EEG data. By providing a strong training label, Fig. 5 shows that per-channel models are flexible to different montages and even robust to large amounts of data loss, as is likely to occur in a clinical environment.

We found evidence of a distribution shift on the Helsinki validation set. Returns on model scaling appears to diminish after the Medium model with the best model becoming metric dependent, indicating that the gains for the Large and XL models don't transfer as well to this dataset (see Fig. 1c). Analysis in Fig. 4 indicates that the Large and XL models are learning something specific to the early-EEG group (postnatal age <1 week) compared to the late-EEG group (>1 week). We speculate that this could be related to subtle differences in the EEG waveforms associated with either postnatal age or, more likely, with primary diagnosis such as HIE or stroke versus other primary diagnoses such as sepsis, meningitis, or recovery post cardiac surgery[20]. This suggests that future development of seizure detectors could benefit from more diverse training data, recorded from neonates at different postnatal ages and with more varied pathologies and seizure aetiologies other than HIE and stroke.

The key result of this work is—for the first time—a thorough demonstration of an expert-level neonatal EEG seizure detector. Although this claim has been made before[28] it was accompanied by some important caveats. First, it was a cross-validation result and not a held-out dataset. Second, this model failed to reach expert-level equivalence when validated on a held-out set[29]. Third, statistical equivalence was found for only one $\Delta\kappa_a$, when replacing one expert, and not for the overall $\Delta\kappa$, an average over the 3 annotators, as our test finds. In our work, in contrast, we report statistical equivalence to experts on two different fully held-out datasets with a combined number of 130 neonates with over 2.7k hours of EEG. For these reasons, we believe that our claim of expert-level equivalence is the first of its kind for neonatal seizure detection.

This study is not without limitations. The observed distribution shift on the Helsinki validation set suggests the XL model works best within the first week of life. Although seizures are most common during this period[1–3], we should not assume that this covers all possible use cases. Another possible limitation is that our development dataset is from one centre. A promising direction for improvement on both counts is to train on a more diverse multi-centre dataset of EEG with

recordings from a larger postnatal time range. And lastly, although we show that the proposed model attains expert-level agreement on our retrospective validation sets, a clinical investigation of the algorithm cotside is the best way to evaluate utility.

In conclusion, we find strong evidence that scaling training data and model size improves performance for neonatal EEG seizure detection. Held-out validation, on datasets with a combined total 2.7-k hours of multi-channel EEG from 130 neonates, found accurate and reliable generalisation performance. Achieving expert-level performance demonstrates readiness for clinical validation. Automated analysis of long-duration EEG facilitates increased seizure surveillance for at-risk neonates. This, in turn, can assist in timely neuroprotective strategies to help improve long-term outcomes for vulnerable neonates in critical care.

Methods

Development dataset

EEG records from 202 term neonates were obtained via a fully-anonymised database of EEG recordings from the Cork University Maternity Hospital (CUMH), Ireland. EEG was recorded as part of ongoing clinical research studies. Informed consent was obtained from the parents or guardians and ethical approval was obtained from the Clinical Research Ethics Committee

Table 5 | Estimates of $\Delta\kappa$, the change in level of agreement by replacing a human expert with the AI model predictions

	Cork ($n = 51$, 2.5k h)		Helsinki ($n = 79{,}102$ h)	
model	**$\Delta\kappa$ (95% CI)**	***p*-value**	**$\Delta\kappa$ (95% CI)**	***p*-value**
Nano	−0.424 (−0.572 to −0.275)	<0.001	−0.132 (−0.206 to −0.060)	<0.001
Small	−0.387 (−0.521 to −0.252)	<0.001	−0.126 (−0.193 to −0.058)	<0.001
Medium	−0.195 (−0.324 to −0.066)	0.003	**−0.052 (−0.118 to 0.010)**	0.099
Large	−0.114 (−0.221 to −0.008)	0.035	−0.066 (−0.126 to −0.001)	0.047
XL	**−0.094 (−0.189 to 0.005)**	0.063	**−0.082 (−0.156 to 0.002)**	0.055

Results are described for both Cork and Helsinki held-out datasets for all model sizes. Distributions are estimated from 1000 bootstrap samples and confidence intervals (CI) including zero and a *p*-value > 0.05 indicate no difference in inter-rater agreement (highlighted by bold font). The Extra-Large (XL) model passes the test for expert equivalence on both datasets.

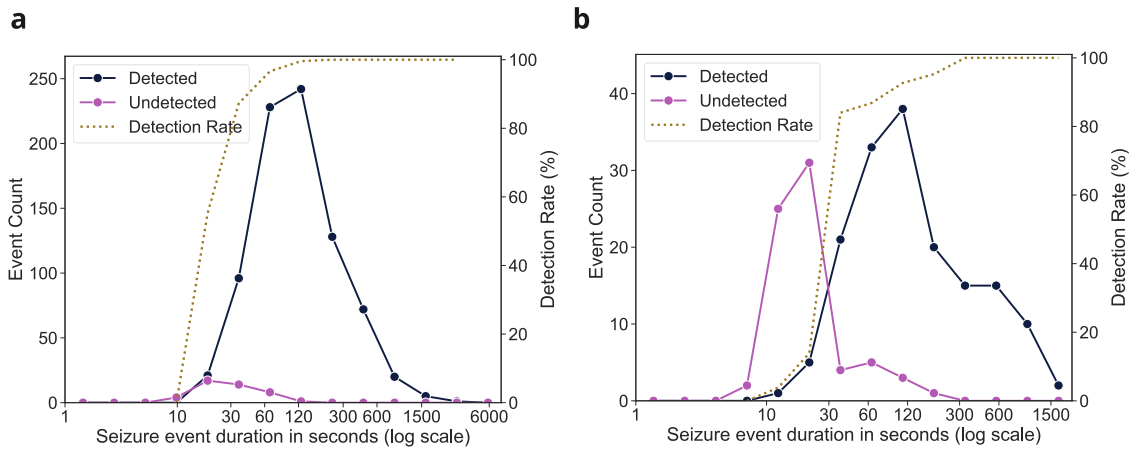

Fig. 3 | Influence of seizure event duration on detection performance. Extra-Large (XL) model performance by event duration of consensus seizures for (**a**) Cork and (**b**) Helsinki validation datasets. A notable finding is that most of the model errors are for short seizures (<30 s), where perhaps the 16 s input segment size limits detection resolution.

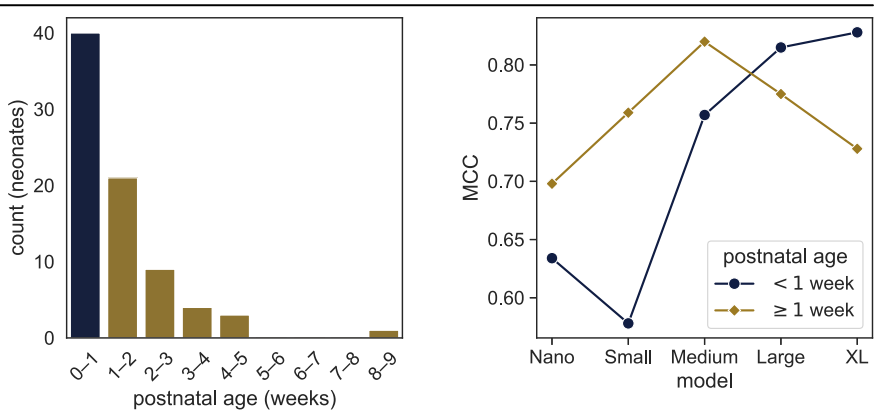

Fig. 4 | Divergence of scaling behaviour between 2 groups in the Helsinki validation dataset. **a** Distribution of postnatal age in weeks. **b** Matthews correlation coefficient (MCC) for both groups. The scaling for <1 week postnatal age tracks closely with that observed in both Cork datasets, even matching the double-descent dip for the Small model. At ≥1 week however we see progressive degradation for the Large and Extra-Large (XL) models comparative to the Medium model.

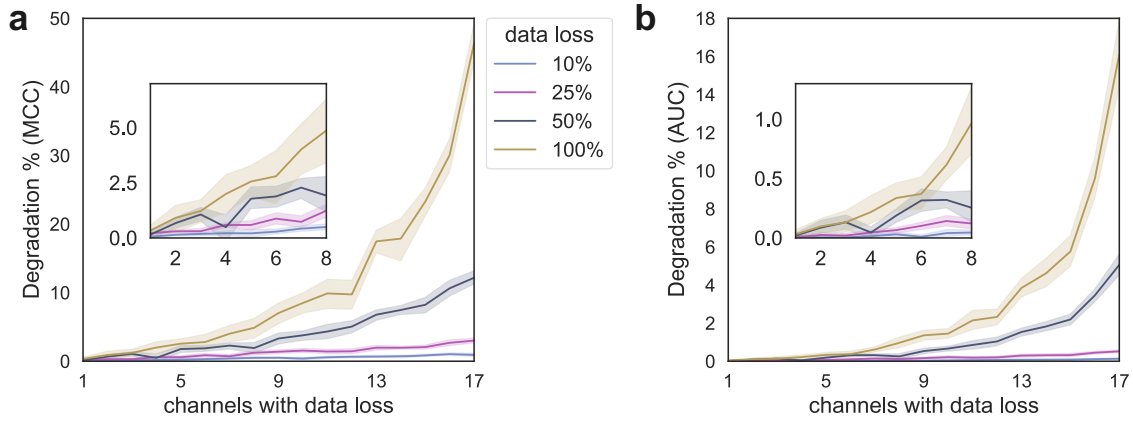

Fig. 5 | Summary of the effect of data loss on model performance on the Helsinki dataset. Degradation is measured relative to zero data loss for Matthews correlation coefficient (MCC) in (**a**) and area under the receiver-operator-characteristic curve (AUC) in (**b**) using XL model. Inset figures illustrate data loss for up to 8 channels.

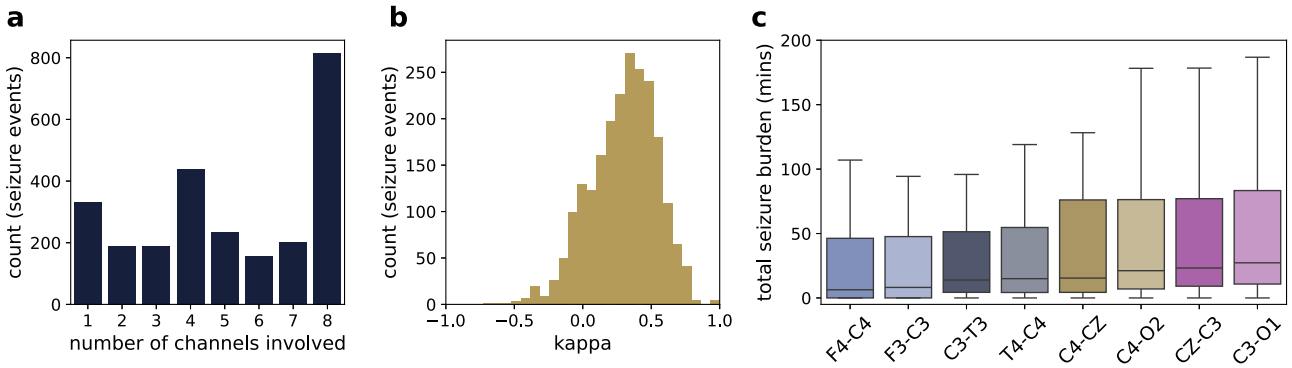

Fig. 6 | Summary of per-channel EEG seizure annotations for 77 neonates. a: number of channels involved in each seizure event. **b**: agreement among seizure annotations across channels for each seizure event, as quantified by Fleiss κ. **c**: total seizure duration for each neonates' EEG estimated from each channel separately. For a small number of EEGs, F3 is replaced by Fp1 or Fp3; and likewise, F4 is replaced by Fp2 or Fp4.

of the Cork Teaching Hospitals. EEG recording commenced as soon as possible after birth and continued for hours or days. EEGs were recorded from term neonates with mixed aetiologies at risk of seizures in the neonatal intensive care unit (NICU) in most cases. We also include a control subset of healthy term newborns recorded in the postnatal wards (≤2 h of EEG per neonate) to use as part of the training data[30–33].

The Neurofax EEG-1200 (Nihon Kohden), NicoletOne ICU Monitor (Natus, USA), or the Lifelines EEG (iEEG Lifelines, Stockbridge, United Kingdom) machines were used to record the EEG. Sampling frequencies were set at 200, 256, or 500 Hz depending on the machine. EEG signals were recorded from the frontal (F3/F4, Fp1/Fp2, or Fp3/Fp4), temporal (T3/T4), central (C3/C4 and CZ), and occipital (O1/O2) or parietal (P3/P4) regions.

A total of 6487 h of multi-channel EEG was reviewed for seizure by two neonatal neurophysiologists (authors SRM and SV). A bipolar montage of 8 channels was used to review seizures, as shown in Fig. 2a. For the control cohort of healthy newborns, the montage was set to F4–T4, T4–P4, P4–CZ, CZ–P3, F3–T3, T3–P3 as these records did not include C3/C4 electrodes[31].

Each channel was reviewed and annotated separately, resulting in 50,299 h of annotated EEG. Seizures were identified in 77 neonates. A total of 12,402 individual per-channel seizure events were annotated (see Fig. 2a for example of per-channel annotations), with a median (interquartile range, IQR) of 48 (19 to 144) distinct seizures events per neonate. Demographic and clinical data are presented in Table 6.

To estimate inter-rater agreement, EEG from 13 neonates was reviewed by both neurophysiologists. Cohen's κ indicated high inter-rater agreement, with a median κ of 0.808 (IQR: 0.702–0.874; range: 0.548–0.990). Although this is calculated on a per-channel rather than global annotation, agreement is in keeping with the previously reported estimates of inter-rater agreement: $\kappa = 0.767$ for the Helsinki dataset[20] and $\kappa = 0.827$ for a Cork/London dataset[34]; both assessments used Fleiss κ to account for the 3 reviewers.

Analysis of the per-channel annotations indicate a high degree of variability in the number of EEG channels involved in each seizure event and in the variability of the time-synchronization of seizures across channels, as illustrated in Fig. 6. This figure also indicates that seizure burden is approximately independent of EEG channel, although the frontal channels (F3–C3 and F4–C4) appear to have a slightly lower burden compared to the other channels.

The per-channel annotations were used to develop a channel-independent algorithm. Different centres will use different protocols when recording EEG, ranging from a 1-channel amplitude-integrated EEG (aEEG) to a full 10:20 electrode array of 19 channels[20]. Developing an AI model on a specific number of channels and a specific montage leads to models that are sensitive to that montage only. Electrodes may detach or become unusable due to artefact during recording. Sustaining a long-duration EEG recording, as is needed for seizure surveillance, without degradation of signal quality on some channels may be unrealistic, given the challenging recording environment of the NICU.

Held-out EEG validation sets

To validate the performance of our algorithms we tested on two held-out, unseen datasets. The first dataset is a cohort consisting of EEG from 51 term neonates with mixed aetiologies at risk of seizures[34]. EEGs were reviewed

Table 6 | Cohort demographics according to the EEG datasets

	Development Dataset	Validation Datasets	
	Cork (n = 202)	Cork (n = 51)	Helsinki (n = 79)
gestational age, weeks+days	40+0 (39+2 to 40+5) [n = 197]	40+4 (39+3 to 41+2)	40 to 41[c] [n = 78]
birth weight, kg	3.50 (3.27 to 3.76) [n = 77]	3.50 (3.13 to 3.91)	3.50 to 4.00[c] [n = 64]
sex, female	86 (42.8%) [n = 201]	22 (43.1%)	35 (45.5%) [n = 77]
Clinical demographics:			
normal cohort[a]	73 (36.2%)	0 (0%)	0 (0%)
hypothermia	28 (18.7%) [n = 150]	13 (25.5%)	
HIE,	63 (31.2%)	27 (13.4%)	29 (14.4%)
mild/moderate/severe	23 / 26 / 14	8 / 13 / 6	3 / 8[d] / 18
birth asphyxia	7 (3.5%)	10 (5.0%)	4 (2.0%)
stroke	9 (4.5%)	6 (3.0%)	11 (5.4%)
EEG characteristics:			
total duration, h	6487	2548	112
duration[b], h	7.4 (1.1 to 51.6)	36.4 (21.4 to 74.4)	1.24 (1.06 to 1.59)
start[b], h	13.8 (10.3 to 27.8)	4.6 (3.0 to 17.9)	<168
neonates with seizures	77 (38.1%)	24 (47.1%)	39 (49.4%)
annotated seizure events	12,402	1572[e]	450[e]

Data are presented as median (interquartile range) or number (percentage) unless otherwise stated. Number of neonates (n) is indicated when data not complete. The development set is used for training and testing the models and the validation sets are used for held-out testing. For the Helsinki dataset, clinical information is extracted from the associated metadata. Common primary diagnoses of HIE, birth asphyxia, and stroke are included in the clinical demographics.
Key: HIE, hypoxic-ischaemic encephalopathy.
[a] EEG collected from healthy neonates in the postnatal ward[31].
[b] per neonate; EEG start is time from birth to first EEG recording.
[c] mode, as data are categorical.
[d] category includes moderate and mild/moderate as defined in the metadata[20].
[e] average number for the 3 reviewers.

independently by three international EEG experts, with a high level of agreement[34]. Although the EEG data is collected in the same location as the development dataset (CUMH), there is no reviewer overlap between this and the development dataset.

The second validation dataset is an open-access neonatal EEG dataset with seizure annotations[20]. Again, this was reviewed by three EEG experts. The dataset consists of EEG from 79 term neonates with mixed aetiologies.

For both validation datasets, seizure annotations were global, a single label used to indicate seizure in one or more channels. We refer to the datasets according to geographic origin: the Cork and Helsinki validation sets. Table 6 includes demographic information on both datasets.

Seizure detection model

We develop a modern convolutional neural network, based on the ConvNeXt architecture[35], for our seizure detection model. In order to test our hypothesis of increasing model scale leading to improved performance we implement several variants of the model related by a simple width and depth scaling paramaterisation. All models are trained to maximise classification performance on 16 s segments of EEG. The hyperparameters and pre- and post-processing are the same for each model (these were fixed via experiments using the smallest model). The development of these models is described in more detail in the following.

We adapt the ConvNeXt architecture[35], originally designed for 2D computer vision applications, to our 1D time-series EEG data. This

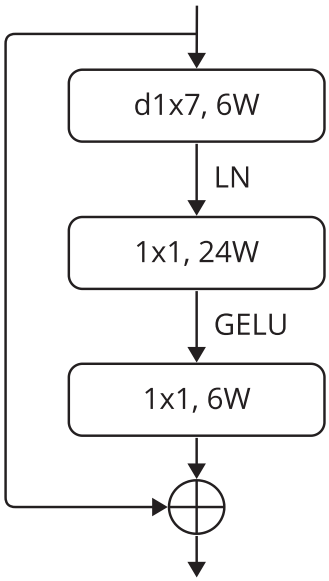

Fig. 7 | ConvNeXt block. Here the W is an integer parameter we use to control the width of our models. The block includes three convolutional layers, one with depth-wise convolutions (indicated by the d) and one-dimensional kernel of length 7 samples, followed by two 1 × 1 convolutional layers, an equivalent implementation of a multi-layer perceptron. Notable features are the use of layer normalisation (LN) rather than batch normalisation and a Gaussian error linear unit (GELU) instead of the rectified linear unit (ReLU).

architecture was systematically designed for efficiency and performance. Taking inspiration from the recent success of vision transformer architectures it was designed with purely convolutional components and achieved state-of-the-art performance across several computer-vision tasks[35]. The basic building block of the model is shown in Fig. 7. Notably, the use of depth-wise convolution and stacked 1 × 1 convolutional layers contribute to increased computational efficiency without sacrificing accuracy.

The detailed architecture is described in Table 7. Our parameterisation defines the network by 2 parameters: D for depth and W for width. Due to the residual structure, simply varying these two integer values allows for easy creation of model variants at different scales without any further adjustments. In this work, we explore models ranging in scales from 38.7k – 20.6m parameters; see Table 1 for the depth–width parameter settings for each model.

Training methods

For long-duration continuous recordings, seizure events typically occupy a small fraction of recording time, with the majority of the EEG being seizure free. A study by Rennie and colleagues described a median (IQR) total seizure burden of 69 (28 to 118) minutes over a median (IQR) of 70 (31 to 97) hours of EEG recording[4]. Additionally, not all neonates with EEG monitoring will have seizures: the same study found that 139 from 214 neonates did not have recorded electrographic seizures, despite the long duration of monitoring. Our development dataset reflects this imbalance, with an approximate class imbalance of 50:1. This imbalance can present a challenge for training machine-learning models, as the models can become biased towards the majority class.

The most common ways to deal with this are (a) oversampling the minority class, (b) undersampling the majority class, and (c) re-weighting the loss function. Oversampling is computational demanding, and for large datasets such as long-duration EEG recordings, unappealing and wasteful of expensive computational resources. Undersampling is also wasteful, as a large proportion of the diverse EEG records are discarded. Loss re-weighting is usually a good option but in our case such a large imbalance can result in large loss values which, even with gradient clipping, can de-stabilise the learning.

Table 7 | Model architecture for the proposed ConvNeXt model

Component	Description	Input tensor (dimension)	Output tensor (dimension)
stem	1 × 4, stride 4	1 × 1 × 1024	6*W* × 1 × 256
stage 1	$\begin{bmatrix} d1 \times 7, & 6W \\ 1 \times 1, & 24W \\ 1 \times 1, & 6W \end{bmatrix} \times D$	6*W* × 1 × 256	6*W* × 1 × 256
downsample	LN, 1 × 2, stride 2	6W × 1 × 256	6W × 1 × 128
stage 2	$\begin{bmatrix} d1 \times 7, & 12W \\ 1 \times 1, & 48W \\ 1 \times 1, & 12W \end{bmatrix} \times D$	6*W* × 1 × 128	12*W* × 1 × 128
downsample	LN, 1 × 2, stride 2	12W × 1 × 128	12W × 1 × 64
stage 3	$\begin{bmatrix} d1 \times 7, & 24W \\ 1 \times 1, & 96W \\ 1 \times 1, & 24W \end{bmatrix} \times 3D$	12*W* × 1 × 64	24*W* × 1 × 64
downsample	LN, 1 × 2, stride 2	24W × 1 × 64	24W × 1 × 32
stage 4	$\begin{bmatrix} d1 \times 7, & 48W \\ 1 \times 1, & 192W \\ 1 \times 1, & 48W \end{bmatrix} \times D$	24*W* × 1 × 32	48*W* × 1 × 32
average pool	1 × 32	48*W* × 1 × 32	48*W* × 1 × 1
linear layer	48*W* × 1	48*W* × 1 × 1	1

Parameters (*D*, *W*) define the depth (*D*) and width (*W*) of the network. The network starts with the stem, followed by multiple stages and downsampling layers, and finishes with average pooling and a linear layer to combine features. The input tensor is an array of 1024 samples for 16 s of EEG sampled at 64 Hz.

Table 8 | Description of metrics used in this work

Variable type	Name	Description
continuous	AP	average precision (area-under the precision–recall curve)
	AP50	AP for recall > 0.5; × 2 to normalise to range (0,1)
	Pearson's *r*	Pearson's correlation coefficient for (p, y)
	cross entropy	mean($y \log(p) + (1-y)\log(1-p)$)
	AUC	area under the receiver-operator-characteristic curve
binary	PPV	positive predictive values (precision); $TP/(TP+FP)$
	NPV	negative predictive values; $TN/(TN+FN)$
	sensitivity	$TP/(TP+FN)$ (recall)
	specificity	$TN/(TN+FP)$
	error rate	$(FP+FN)/N$; 1 − accuracy
	MCC	Matthews correlation coefficient
	Cohen's κ	measure of pairwise agreement accounting for chance
	Fleiss' κ	generalisation of Cohen's κ to >2 annotators
	FD/h	false event detection per hour
	seizure burden, *r*	correlation coefficient for hourly estimate of predicted versus true seizure burden in mins/h

Note whenever *cc* subscript is used it means the value was computed by concatenating all recordings.
Key: y, true label; p, model prediction probability; TP, true positives; FP, false positives; TN, true negatives; FN, false negatives; N, total predictions.

Instead, we use stratified mini-batch sampling: we keep all data and dynamically undersample the non-seizure examples at random during training. From one training epoch to the next the model will see a different sample of the non-seizure data but the same seizure data. By selecting a different random sample of non-seizure data per training epoch, all of the non-seizure data will be exposed during training with a sufficient number of training epochs. In practice, we found that dynamically undersampling to a ratio of 5:1 and combining loss re-weighting to account for this imbalance was the most stable and efficient implementation.

All models are trained with the same hyperparameters using AdamW on a learning-rate schedule. The learning-rate schedule follows a variant of the 1-cycle policy[36] with 4 phases: warmup, freeze-at-max, cooldown, then freeze-at-min. The learning rate changes logarithmically during the warmup and cooldown phases. In our experiments we found this schedule reliably led to training convergence: 10 random initialisations of the Medium model resulted in a mean (standard deviation) relative change in MCC of just −0.083% (0.614%). This eliminated the need for early-stopping based on monitoring of validation loss. Although common practice in many machine-learning applications, we have found this to be unreliable. The large variability among neonates resulted in the early stopping condition being highly sensitive to the choice of babies in the validation set. One approach to mitigate this is to use more than one *k*-fold[13], but this results in several models that need to be ensembled somehow. Problematically, this sensitivity of the model to the validation set raises questions about generalisation to unseen data when using this method. Additionally, a consequence of the deep-double descent phenomenon, which we observe in the Results section (Fig. 1), is that early stopping will only select the best model in the special case of when the model size and dataset size are critically balanced[21].

To improve model robustness we developed and experimented with several data augmentation techniques. This consisted of several signal processing transformations: magnitude scaling, magnitude warping, jitter, time warping, and spectral-phase randomisation. In addition, generic transformations such as flip, cutmix[37], cutout[38], and mixup[39] were applied. The parameters of each augmentation were manually adjusted to ensure all transformations were label preserving. Different probabilities were assigned to each transformation for a given batch. From our experimentation, only flip and cutout gave consistent improvements in performance and were therefore included in the model development presented here. Improvement varies somewhat with scale but inclusion of augmentation gives ~5% relative improvement on MCC.

Pre- and post-processing

Pre-processing of the EEG consisted of bandpass filtering within the 0.3–30 Hz passband, downsampling to 64 Hz, and removal of some artefacts. These artefacts were either periods of contiguous zeros, caused by checking the impedance of electrode scalp contact, or periods of excessive high-amplitude activity, defined by a standard deviation greater than 1 mV for each segment. EEG was divided into 16 s segments with a step size of 4 s. These segments were labelled as seizure if ≥8 s of the segment was annotated as seizure and non-seizure otherwise. Each channel of the segment was then used as a separate training example and was assigned a positive label if that specific channel contained a seizure annotation. This results in ~42m segments (10.5m without overlap) with a negative:positive ratio of ~ 50: 1.

When testing the model with a full EEG recording, we processed 16 s segments with a step size of 0.25 s. The continuous-valued output of the model is then smoothed with a 32 s rectangular window. From this probability-like output, we apply the standard threshold of 0.5 to generate the binary decision mask. Very short segments (<10 s) of seizure (non-seizure) are deleted (filled) in the final mask. We deliberately restrict our post-processing to be simple and limited in contrast to some more involved schemes in previous work[10,13,40]. While approaches like adding a collar to detected events or optimising the threshold can help with some metrics on some datasets[10,13,40], we believe the best way to generalise well to other datasets is to rely on the model to learn the start and end of seizure events directly from the data.

All models were designed and built using the development dataset. The set was divided with a random 80:20 split of neonates: 80% of neonates' EEG used for training and 20% for testing. When development was finalised the models were then trained on all the

development dataset and tested on the held-out validation sets. There was no back-and-forth between model development and testing on the held-out datasets.

Evaluating performance

We conduct a comprehensive evaluation of the model using two complementary approaches: (1) performance metrics using human annotations as the gold standard and (2) human-expert equivalence testing. To enable reproducible research, we developed an open-source Python framework to run the evaluations (including both metrics and statistical tests) used in the study (available at https://github.com/CergenX/SPEED, commit c09f60a).

We include a range of performance metrics to avoid reliance on a single metric. Because of the many limitations associated with the area under the receiver-operating-characteristic curve (AUC)[24–26], we opt to include more transparent measures such Pearson's correlation and Matthews correlation coefficient (MCC)[26,27]. We also include clinically-relevant measures such as false detections per hour (FD/h) and seizure burden per hour. A complete list of metrics is presented in Table 8. With multiple annotators, as we have for both our validation sets, we follow the convention of using a consensus annotation[13,19,40].

The metrics presented in Table 8 use annotations from a single expert or a consensus of experts as a gold standard. This approach is useful for comparing models but is often hard to interpret for clinical adoption. It also fails to capture the level of agreement among experts or quantify performance relative to that.

We evaluate performance relative to inter-rater agreement using a test developed for neonatal EEG seizure detection[28,29,40]. The method measures the impact of replacing each expert with the AI model predictions and quantifying the difference in inter-rater agreement using Fleiss κ to account for agreements by random chance. We define this difference in agreement for our 3 annotator held-out datasets as

$$\Delta\kappa_a = \kappa_{\text{experts}} - \kappa_{\text{AI},a} \quad \text{for } a = 1, 2, 3 \tag{1}$$

where κ_{experts} is the inter-rater agreement among the 3 experts and $\kappa_{\text{AI},a}$ is the agreement with 2 experts and the AI for the 3 possible combinations. An overall difference in agreement, $\Delta\kappa$, is estimated as the mean value of $\Delta\kappa_a$ over the 3 experts. The condition of $\Delta\kappa = 0$ indicates that the AI predictions do not change inter-rater agreement and therefore can be considered equivalent[29,40]. To test whether $\Delta\kappa = 0$, we follow the process of generating a distribution of $\Delta\kappa$ by bootstrapping with 1000 iterations randomly resampling by neonate, computing $\Delta\kappa$ for each resample. This allows us to estimate the variability in $\Delta\kappa$ introduced by variability in inter-rater agreement as well as model performance. From this distribution, if the 95% confidence interval (CI) includes 0 then we accept the null hypothesis that the model predictions do not significantly alter inter-rater agreement. Adherence to this condition establishes expert-level performance for the AI model.

Data availability

The EEG datasets are not publicly available because under the terms of the data licensing agreement and the consent obtained from the ethics committee we do not have permission to share the raw data. The Helsinki EEG dataset used for validation is freely available at https://doi.org/10.5281/zenodo.4940267.

Code availability

A PyTorch implementation of the model is publicly available at https://github.com/cergenx/ConvNeXt-Seizure (commit fd7e48f). The code use for all evaluation metrics is publicly available at https://github.com/cergenx/SPEED/ (commit c09f60a).

Received: 3 July 2024; Accepted: 21 December 2024;
Published online: 08 January 2025

References

1. Tekgul, H. et al. The current etiologic profile and neurodevelopmental outcome of seizures in term newborn infants. *Pediatrics* **117**, 1270–1280 (2006).
2. Soul, J. S. Acute symptomatic seizures in term neonates: etiologies and treatments. *Semin. Fetal. Neonatal. Med.* **23**, 183–190 (2018).
3. Pisani, F., Spagnoli, C., Falsaperla, R., Nagarajan, L. & Ramantani, G. Seizures in the neonate: a review of etiologies and outcomes. *Seizure* **85**, 48–56 (2021).
4. Rennie, J. M. et al. Characterisation of neonatal seizures and their treatment using continuous EEG monitoring: a multicentre experience. *Arch. Dis. Child Fetal. Neonatal. Ed.* **104**, F493–F501 (2018).
5. Uria-Avellanal, C., Marlow, N. & Rennie, J. M. Outcome following neonatal seizures. *Semin. Fetal. Neonatal. Med.* **18**, 224–232 (2013).
6. Srinivasakumar, P. et al. Treating EEG seizures in hypoxic ischemic encephalopathy: a randomized controlled trial. *Pediatrics* **136**, e1302–e1309 (2015).
7. Murray, D. M. et al. Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures. *Arch. Dis. Child Fetal. Neonatal. Ed.* **93**, F187–F191 (2008).
8. Pavel, A. M. et al. Neonatal seizure management: is the timing of treatment critical? *J. Pediatr.* **243**, 61–68.e2 (2022).
9. Pavel, A. M. et al. A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial. *Lancet Child Adolesc. Health* **4**, 740–749 (2020).
10. Temko, A., Thomas, E., Marnane, W., Lightbody, G. & Boylan, G. B. EEG-based neonatal seizure detection with support vector machines. *Clin. Neurophysiol.* **122**, 464–473 (2011).
11. Ansari, A. H. et al. Neonatal seizure detection using deep convolutional neural networks. *Int. J Neural. Syst.* **29**, 1850011 (2019).
12. Isaev, D.Y. et al. Attention-based network for weak labels in neonatal seizure detection. *Proc. Mach. Learn. Healthcare Conf.* **126**, 479–507 (PMLR, 2020).
13. O'Shea, A., Lightbody, G., Boylan, G. B. & Temko, A. Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture. *Neural Netw.* **123**, 12–25 (2020).
14. Daly, A, O'Shea, A, Lightbody, G & Temko, A. Towards deeper neural networks for neonatal seizure detection. In *Int. Conf. Eng. Med. Bio. Soc. (EMBC)* 920–923 (IEEE, 2021).
15. Caliskan, A & Rencuzogullari, S. Transfer learning to detect neonatal seizure from electroencephalography signals. *Neural Comput. Appl.* **33**, 2087–12101 (2021).
16. Tanveer, M. A., Khan, M. J., Sajid, H. & Naseer, N. Convolutional neural networks ensemble model for neonatal seizure detection. *J. Neurosci. Methods* **358**, 109197 (2021).
17. Raeisi, K. et al. A graph convolutional neural network for the automated detection of seizures in the neonatal EEG. *Comput. Methods Prog. Biomed.* **222**, 106950 (2022).
18. Daly, A, Lightbody, G & Temko, A. Bridging the source-target mismatch with pseudo labeling for neonatal seizure detection. In *Eur. Signal Proc. Conf. (EUSIPCO)* 1100–1104 (IEEE, 2023).
19. Raeisi, K. et al. A class-imbalance aware and explainable spatio-temporal graph attention network for neonatal seizure detection. *Int. J. Neural Syst.* **33**, 2350046 (2023).
20. Stevenson, N. J., Tapani, K., Lauronen, L. & Vanhatalo, S. A dataset of neonatal EEG recordings with seizure annotations. *Sci. Data* **6**, 1–8 (2019).
21. Nakkiran, P. et al. Deep double descent: where bigger models and more data hurt. *J. Stat. Mech. Theory Exp.* **2021**, 124003 (2021).
22. Kaplan, J. et al. Scaling laws for neural language models. *arXiv* https://doi.org/10.48550/arXiv.2001.08361 (2020).
23. Hoffmann, J. et al. Training compute-optimal large language models. *arXiv* https://doi.org/10.48550/arXiv.2203.15556 (2022).

24. Lobo, J. M., iménez-Valverde, A. & Real, R. AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **17**, 145–151 (2007).
25. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432 (2015).
26. Chicco, D & Jurman, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min.* **16**, 4 (2023).
27. Chicco, D & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
28. Stevenson, N.J., Tapani, K & Vanhatalo, S. Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert. *Int. Conf. Eng. Med. Bio. Soc*. **2019**, 5991–5994 (2019).
29. Tapani, K. T., Nevalainen, P., Vanhatalo, S. & Stevenson, N. J. Validating an SVM-based neonatal seizure detection algorithm for generalizability, non-inferiority and clinical efficacy. *Comput. Biol. Med.* **145**, 105399 (2022).
30. Korotchikova, I. et al. EEG in the healthy term newborn within 12 hours of birth. *Clin. Neurophysiol.* **120**, 1046–53 (2009).
31. Korotchikova, I., Stevenson, N. J., Livingstone, V., Ryan, C. A. & Boylan, G. B. Sleep–wake cycle of the healthy term newborn infant in the immediate postnatal period. *Clin. Neurophysiol.* **127**, 2095–2101 (2016).
32. Raurale, S. A., Boylan, G. B., Lightbody, G & O'Toole, J. M. Identifying tracé alternant activity in neonatal EEG using an inter-burst detection approach. *Int. Conf. Eng. Med. Bio. Soc.* **2020**, 5984–5987 (2020).
33. Garvey, A. A. et al. Multichannel EEG abnormalities during the first 6 hours in infants with mild hypoxic-ischaemic encephalopathy. *Pediatr. Res.* **90**, 117–124 (2021).
34. Stevenson, N. J. et al. Interobserver agreement for neonatal seizure detection using multichannel EEG. *Ann. Clin. Transl. Neurol.* **2**, 1002–1011 (2015).
35. Liu, Z. et al. A ConvNet for the 2020s. In *Proc IEEE/CVF Int Conf Computer Vision (CVPR)*, 11966–11976 (IEEE, 2022).
36. Smith, L.N. & Topin, N Super-convergence: very fast training of neural networks using large learning rates. *AI Mach. Learn. Multidomain Oper. Appl*. **11006**, 369–386 (SPIE, 2019).
37. Yun, S. et al. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proc. IEEE/CVF Int. Conf. Computer Vision (CVPR)*, 6023–6032 (IEEE, 2019).
38. DeVries, T & Taylor, G.W. Improved regularization of convolutional neural networks with cutout. *arXiv* https://doi.org/10.48550/arXiv.1708.04552 (2017).
39. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* https://doi.org/10.48550/arXiv.1710.09412 (2018).
40. Tapani, K. T., Vanhatalo, S. & Stevenson, N. J. Time-varying EEG correlations improve automated neonatal seizure detection. *Int. J. Neural Syst.* **29**, 1850030 (2019).

Acknowledgements

GB was supported by a Wellcome Trust Innovator Award (209325/Z/17/Z). This research was funded by an Enterprise Ireland Disruptive Technology Innovation Fund Award to CergenX (DT 2023 435). Theses funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

R.H., J.O.T., and S.G. were responsible for the study conception. R.H. and J.O.T. were responsible for study design, model training and validation, visualisations, and drafted the initial version of the manuscript. S.M. and S.V. provided annotations of the training and test data. A.L. was responsible for pre-processing and management of the EEG data and annotations. G.B. was responsible for data curation and provided advisory support for the project. All authors read and approved the final manuscript.

Competing interests

All authors are affiliated with CergenX Ltd, a company developing neuromonitoring technologies for newborns: G.B. and S.G. are co-founders; R.H., A.L., and J.O.T. are employees; and S.M. and S.V. were paid contractors.

Additional information

Correspondence and requests for materials should be addressed to Robert Hogan or John M. O'Toole.