Titre: Title:	Diagnostiquer l'épilepsie à l'aide de l'apprentissage machine appliqué à l'électroencéphalogramme
Auteur: Author:	Emile Lemoine
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Lemoine, E. (2025). Diagnostiquer l'épilepsie à l'aide de l'apprentissage machine appliqué à l'électroencéphalogramme [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/68171/

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Diagnostiquer l'épilepsie à l'aide de l'apprentissage machine appliqué à l'électroencéphalogramme

EMILE LEMOINE

Institut de génie biomédical

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*

Génie biomédical

Août 2025

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Cette thèse intitulée :

Diagnostiquer l'épilepsie à l'aide de l'apprentissage machine appliqué à l'électroencéphalogramme

présentée par **Emile LEMOINE**

en vue de l'obtention du diplôme de *Philosophiæ Doctor*

a été dûment acceptée par le jury d'examen constitué de :

Benjamin DE LEENER, président
Frédéric LESAGE, membre et directeur de recherche
Elie BOU ASSI, membre et codirecteur de recherche
Dang Khoa NGUYEN, membre et codirecteur de recherche
Guillaume DUMAS, membre
Maxime BAUD, membre externe

REMERCIEMENTS

Je souhaite d'abord remercier mes directeurs de thèse, Elie Bou Assi, Frédéric Lesage et Dang Nguyen. Merci pour votre encadrement, vos conseils et vos révisions minutieuses (même celles de dernière minute!). Vous avez su équilibrer votre supervision en me donnant une grande liberté pour explorer mes idées, et ainsi me permettre de bâtir ma confiance et mon autonomie comme chercheur. Je suis extrêmement reconnaissant pour les opportunités que vous avez mises sur mon chemin. Plus particulièrement, Elie, je suis fier d'avoir pu participer à tes débuts comme chercheur indépendant. Ta disponibilité et ton enthousiasme m'ont aidé à traverser les hauts et les bas de cette aventure, et travailler à tes côtés a rendu excitants mêmes les moments les plus stressants. J'ai hâte de devenir votre collègue et de continuer à bâtir de nouveaux projets ambitieux à vos côtés.

Je tiens à remercier sincèrement les membres du jury, Benjamin de Leener, Guillaume Dumas et Maxime Baud, d'avoir accepté d'évaluer ce travail. Votre expertise et le temps consacré à la lecture de cette thèse contribueront grandement à en améliorer sa qualité.

Cette aventure n'aurait pas été possible sans le soutien financier des Instituts de Recherche en Santé du Canada et du Fonds de Recherche Québec-Santé, auxquels j'exprime ma reconnaissance.

Je souhaite remercier chaleureusement tous les membres du laboratoire, passés et présents, qui ont contribué à rendre cette expérience si enrichissante. Une pensée particulière pour Véronique et Manon, dont le soutien a été précieux. Merci également à Amal et à la Polytechnique.

Merci au programme de résidence en neurologie, particulièrement aux directeurs Nicolas Jodoin et Antoine Duquette. Votre flexibilité permet aux résidents de concrétiser leurs ambitions, et mon parcours n'aurait pas été possible sans votre support. Merci à Catherine Larochelle pour son mentorat, et à mes corésidents pour leurs encouragements et leur soutien pendant ces années intenses.

À Justine, merci pour ton amour et ton soutien inconditionnel à toute heure du jour ou de la nuit. Ta présence est essentielle à tout ce que j'entreprends. Et Roméo, merci pour ta chaleur et aussi ta patience quand les promenades se faisaient attendre.

Enfin, merci à mes parents, à ma sœur Juliette et mon frère Victor, ainsi qu'à toute ma famille, pour votre amour et votre compréhension pendant ces années chargées. C'est grâce à vous que j'ai la confiance et l'ambition de m'embarquer dans des projets comme celui-ci.

RÉSUMÉ

L'épilepsie est une maladie neurologique chronique invalidante qui affecte 1% de la population mondiale. Elle est caractérisée par la survenue imprévisible de crises épileptiques, des décharges électriques cérébrales anormales et synchronisées qui sont parfois associées à des symptômes neurologiques. L'électroencéphalogramme (EEG), un examen qui enregistre l'activité électrique corticale via des électrodes positionnées sur le scalp, est central au diagnostic et à la prise en charge de l'épilepsie. Son utilité diagnostique repose principalement sur l'identification visuelle de décharges épileptiformes interictales (DÉI), des anomalies sporadiques et asymptomatiques qui reflètent une irritabilité corticale anormale et donc un risque de crise accru. Malgré son importance, l'utilité de l'EEG en épilepsie pourrait être bonifiée. D'une part, les DÉI sont détectées chez une minorité de patients lors d'un EEG de routine de 30 minutes. D'autre part, leur identification visuelle comporte une part de subjectivité, menant parfois à un surdiagnostic. Des études préliminaires suggèrent qu'il existe d'autres différences subtiles dans l'EEG de patients avec épilepsie qui, bien qu'invisibles à l'œil nu, pourrait être captées par des méthodes computationnelles. Couplés au DÉI, ces marqueurs pourraient augmenter l'utilité diagnostique et prognostique de l'EEG. Cependant, le développement et la validation de tels biomarqueurs sont freinés par le manque de bases de données cliniques appropriées.

L'objectif principal de cette thèse était de développer des méthodes computationnelles pour extraire de l'EEG des marqueurs quantifiables du risque de crise, indépendamment de la présence de DÉI. Plus spécifiquement, nous visions à: 1) développer une base de données d'EEG de routine avec données cliniques détaillées provenant de patients consécutifs, permettant la découverte et la validation rigoureuse de biomarqueurs; 2) valider les performances des biomarqueurs neurophysiologiques précédemment décrits et explorer de nouvelles caractéristiques du signal EEG associées à l'épilepsie; et 3) concevoir et optimiser un modèle d'apprentissage profond interprétable pour la détection de l'épilepsie et la prédiction du risque de crise à partir de l'EEG de routine.

Pour le premier objectif, nous avons mis sur pied une cohorte de patients consécutifs ayant eu un EEG de routine au Centre hospitalier de l'Université de Montréal (CHUM) entre 2018 et 2020. Cette base de données unique contient présentement plus de 1 000 EEG consécutifs effectués chez plus de 900 patients. Chaque EEG est accompagné de variables cliniques détaillées provenant de leur suivi clinique, qui s'étend en moyenne sur plus de 2 ans après l'EEG. Pour chaque étude, de

nouveaux patients sont utilisés comme échantillon test, avec un décalage temporel mimant le déploiement de la technologie dans la vraie pratique.

Le deuxième objectif a été adressé dans un premier article qui évaluait des caractéristiques computationnelles classiques comme la puissance de bande et l'entropie, extraites automatiquement de segments d'EEG de 10 secondes, pour prédire la récurrence de crise à un an. Le modèle développé a atteint une aire sous la courbe ROC (AUROC) de 0.63 (IC 95%: 0.55–0.71) dans la cohorte de test, démontrant des performances significatives même en l'absence de DÉI. Ceci validait ainsi l'hypothèse que l'EEG contient des biomarqueurs du risque de crise indépendants des DÉI, mais avec des performances diagnostiques modestes.

Pour le troisième objectif, nous avons développé deux architectures d'apprentissage profond. Le premier modèle visait à démontrer l'applicabilité de ce type d'approche pour améliorer la précision diagnostique de l'EEG pour l'épilepsie. DeepEpilepsy, un *Vision Transformer* qui modélise directement le signal EEG brut, a surpassé les marqueurs computationnels ainsi que l'interprétation basée sur les DÉI avec une AUROC de 0.76 (0.69–0.83), atteignant 0.83 lorsque combiné aux DÉI. Une analyse de l'espace latent a révélé que DeepEpilepsy semblait dépendre des variabilités dans la puissance de bande des hautes fréquences (50–100 Hz).

Bien que DeepEpilepsy démontrait la puissance des modèles profonds, son utilité clinique était limitée par la nature binaire de sa classification qui ne prenait pas en compte la variabilité clinique au sein des patients avec et sans épilepsie. Pour résoudre ce problème, nous avons développé EEGSurvNet, un modèle de survie profond qui permet de prédire le risque de crise à travers le temps. Entraîné sur 917 EEG et testé sur 135 enregistrement indépendants, EEGSurvNet atteint une discrimination (AUROC intégré à deux ans = 0.69, AUROC à deux mois = 0.80) et une calibration (score de Brier intégré à deux mois = 0.18) supérieures aux prédicteurs traditionnels. Comme DeepEpilepsy, EEGSurvNet ne dépend pas de la présence de DÉI ou de ralentissement anormal, et celui-ci obtient même de meilleures performances en leur absence. Cependant, les patrons captés semblent plutôt situés dans la bande de fréquence 6–15 Hz et évoluent sur une échelle temporelle d'au moins une minute.

Cette thèse établit de nouveaux standards méthodologiques pour le développement d'algorithmes en épilepsie, notamment par l'utilisation d'une cohorte consécutive de patients et une validation temporellement décalée qui mime le déploiement clinique réel. Sur le plan scientifique, nos travaux

ouvrent une nouvelle fenêtre pour investiguer la neurophysiologie de l'épilepsie: les modèles profonds ont révélé des patrons EEG distincts des marqueurs traditionnels, évoluant sur une échelle temporelle plus longue. Ces caractéristiques, indépendantes des DÉI, pourraient refléter des altérations plus subtiles des réseaux neuronaux. Sur le plan clinique, la quantification précise du risque de crise pourrait transformer la prise en charge des patients en améliorant la certitude diagnostique, en guidant l'ajustement thérapeutique et en optimisant la sélection des patients pour des interventions plus agressives. Bien que la principale limitation soit l'utilisation de données d'un seul centre, une validation multicentrique est en cours. L'impact réel de cette technologie dépendra non seulement de cette validation robuste, mais aussi d'une réflexion approfondie sur son intégration dans la pratique clinique et son interaction avec le jugement médical.

ABSTRACT

Epilepsy is a chronic disabling neurological disease affecting 1% of the world's population. It is characterized by unpredictable seizures—abnormal and synchronized electrical discharges in the brain that can provoke neurological symptoms. Electroencephalogram (EEG), a test that records cortical synaptic electrical activity via scalp electrodes, is central to the diagnosis and management of epilepsy. Its diagnostic utility in epilepsy relies primarily on the visual identification of interictal epileptiform discharges (IEDs), sporadic and asymptomatic abnormalities that reflect abnormal cortical irritability and thus an increased seizure risk. Despite its central role in epilepsy, the utility of EEG could be enhanced. On the one hand, interictal epileptiform discharges (IEDs) are detected in only a minority of patients during a routine 30-minute EEG. On the other hand, their visual identification remains somewhat subjective, which may occasionally lead to overdiagnosis. Preliminary studies suggest that there are other subtle differences in the EEG of patients with epilepsy that, although invisible to the naked eye, could be captured by computational methods. Coupled with IEDs, these markers could increase the diagnostic and prognostic yield of the EEG. However, the development and validation of such biomarkers are restrained by the lack of appropriate clinical databases.

The main objective of this thesis was to develop computational methods to extract quantifiable markers of seizure risk from EEG, independently of the presence of IEDs. Specifically, we aimed to: 1) develop a routine EEG database with detailed clinical data from consecutive patients, enabling the discovery and rigorous validation of biomarkers; 2) validate the performance of previously described neurophysiological biomarkers and explore new EEG signal characteristics associated with epilepsy; and 3) design and optimize an interpretable deep learning model for epilepsy detection and seizure risk prediction from routine EEG.

For the first objective, we established a cohort of consecutive patients who underwent routine EEG at the Centre hospitalier de l'Université de Montréal between 2018 and 2020. This unique database currently contains over 1,000 consecutive EEGs performed on more than 900 patients. Each EEG is accompanied by detailed clinical variables from their clinical follow-up, which extends on average over 2 years after the EEG. For each study, new patients are used as test samples, with a temporal shift mimicking real-world technology deployment.

The second objective was addressed in a first article that evaluated classical computational features such as band power and entropy, automatically extracted from 10-second EEG segments, to predict one-year seizure recurrence. The developed model achieved an area under the receiver operating characteristic curve (AUROC) of 0.63 (95% CI: 0.55-0.71) in the test cohort, demonstrating significant performance even in the absence of IEDs. This validated the hypothesis that EEG contains seizure risk biomarkers independent of IEDs, albeit with modest diagnostic performance.

For the third objective, we developed two deep learning architectures. The first model, presented in the second article of the thesis, aimed to demonstrate the applicability of this approach to improve the diagnostic accuracy of EEG for epilepsy. DeepEpilepsy, a Vision Transformer that directly models raw EEG signal, outperformed computational markers and IED-based interpretation with an area under the ROC curve of 0.76 (95% CI: 0.69-0.83), reaching 0.83 when combined with IEDs. Latent space analysis revealed that DeepEpilepsy seemed to depend on variabilities in high-frequency band power (50–100 Hz).

Although DeepEpilepsy demonstrated the power of deep models, its clinical utility was limited by the binary nature of its classification, which did not account for clinical variability among patients with and without epilepsy. To address this issue, we developed EEGSurvNet, a deep survival model that predicts seizure risk over time. In a third study, we trained EEGSurvNet on 917 EEGs and tested on 135 independent recordings, and showed that this model achieves discrimination (integrated AUROC at two years = 0.69, AUROC at two months = 0.80) and calibration (integrated Brier Score at two months = 0.18) superior to traditional predictors. Like DeepEpilepsy, EEGSurvNet does not depend on the presence of IEDs or abnormal slowing, and even performs better in their absence. However, the captured patterns appear to be located in the 6–15 Hz frequency band and evolve over a time scale of at least one minute.

This thesis establishes new methodological standards for the development of algorithms in epilepsy, notably through the use of a consecutive patient cohort and temporally shifted validation that mimics real clinical deployment. From a scientific perspective, our work opens a new window to investigate the neurophysiology of epilepsy: deep models have revealed EEG patterns distinct from traditional markers, evolving over a longer time scale. These characteristics, independent of IEDs, could reflect more subtle alterations in neural networks. From a clinical perspective, precise quantification of seizure risk could transform patient care by improving diagnostic certainty,

guiding therapeutic adjustment, and optimizing patient selection for invasive interventions. Although the main limitation is the use of data from a single center, multicenter validation is ongoing. The real impact of this technology will depend not only on this robust validation but also on careful consideration of its integration into clinical practice and its interaction with medical judgment.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

aHR	Adjuted hazards ratio
ApEn	Approximate entropy
ASM	Antiseizure medication
AUC	Area under the curve
AUROC	Aire sous la courbe ROC ou Area under the receiver operating characteristic curve
AUPRC	Area under the precision recall curve
BP	Band power
CD	Correlation dimension
CHUM	Centre hospitalier de l'Université de Montréal
CI	Confidence interval
CNN	Réseau de neurones à convolution ou Convolutional neural network
CRCHUM	Centre de recherche du Centre hospitalier de l'Université de Montréal
CV	Cross-validation
CvT	Compact Vision Transformer
DÉI	Décharge épileptiforme interictale
DL	Deep learning
ECG	Électrocardiogramme
EDF	European data format
EEG	Électroencéphalogramme
EMG	Électromyogramme
FIR	Finite impulse response
FuzzEn	Fuzzy entropy
GLM	Generalized linear model

GPU	Processeur graphique ou Graphical processing unit
HE	Hurst exponent
HR	Hazards ratio
iAUROC	Aire sous la courbe ROC intégrée
iBS	Score de Brier intégré
IC	Intervalle de confiance
IED	Interictal epileptiform discharge
IoC	Improvement over chance
IPCW	Inverse de la probabilité de censure ou Inverse probability of censoring weight
IQR	Interquartile range
IRM	Imagerie par resonance magnétique
LL	Line length
LSTM	Long short term memory
ML	Machine learning
MLP	Multilayer perceptron
MRI	Magnetic resonance imaging
NPV	Negative predictive value
PAF	Peak alpha frequency
PermEn	Permutation entropy
PPV	Positive predictive value
PSWE	Paroxysmal slow wave event
PWE	Patient with epilepsy
REB	Research ethics board
ReLU	Rectified Linear Unit

RF	Random forest
ROC	Receiver operating characteristic
RNN	Réseau de neurones récurrent ou Recurrent neural network
SampEn	Sample entropy
SpecEn	Spectral entropy
SVM	Support vector machine
TIRDA	Ralentissement temporal rythmique intermittent ou Temporal intermittent rhythmic delta activity
TRIPOD	Transparent reporting of multivariate prediction model for individual prognosis or diagnosis
ViT	Vision Transformer
ViT1d	Vision Transformer with 1-dimensional convolutional tokenizer

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

L'épilepsie affecte 1% de la population mondiale [1]. Il s'agit d'une maladie chronique invalidante caractérisée par la survenue sporadique et imprévisible d'une activité électrique anormale et synchronisée dans le cerveau, appelée crise épileptique [2], [3]. Ces crises peuvent entraîner des symptômes neurologiques incontrôlables, avec ou sans altération de l'état de conscience, mettant les patients en danger de blessure grave, d'accident automobile et même de décès [4], [5].

L'électroencéphalogramme (EEG), dont le premier enregistrement chez un humain est attribué à Hans Berger en 1929 [6], est un examen central au diagnostic et à la prise en charge de l'épilepsie. L'EEG consiste à enregistrer l'activité électrique corticale via des électrodes apposées sur le scalp, offrant une fenêtre sur les processus mêmes qui sous-tendent l'épilepsie. Son interprétation requiert une analyse visuelle approfondie des oscillations cérébrales à la recherche de changements subtils dans leur fréquence, amplitude et symétrie. Quelques années après les premiers EEG de Berger, on décrit de brèves ondes pointues, sporadiques et asymptomatiques, émanant du cerveau de patients avec épilepsie [6]. Ces décharges, maintenant appelées pointes épileptiformes ou décharges épileptiformes interictales (DÉI), sont le reflet d'une irritabilité corticale anormale et sont fortement associées à l'épilepsie [7]. Elles se distinguent des crises par une occurrence plus fréquente, une durée brève (< 200 ms) et l'absence de symptômes associés. Malheureusement, l'utilisation de DÉI comme biomarqueur d'épilepsie est limitée. Premièrement, de par leur nature sporadique, elles apparaissent chez moins de la moitié des patients avec épilepsie sur un enregistrement standard de 30 minutes [8], [9], [10]. Deuxièmement, elles peuvent survenir chez des patients sans épilepsie [11]. Troisièmement, elles reposent sur une analyse subjective et sont fréquemment sur-identifiées: plusieurs ondes pointues peuvent s'apparenter à des DÉI, et la mauvaise identification des DÉI est une cause importante du surdiagnostic d'épilepsie [12], [13], [14].

L'apprentissage machine, particulièrement l'apprentissage profond, a révolutionné plusieurs sphères de notre société. Dans les dernières années, des modèles mathématiques contenant plusieurs milliards de paramètres, optimisés sur des jeux de données massifs, se sont inscrits comme l'état de l'art dans plusieurs domaines tels que l'analyse d'image [15], l'interprétation du son [16] et la génération de texte [17]. En sciences, ces modèles ont mené à la découverte de médicaments [18] et ont résolu des problèmes jugés « impossibles » comme la modélisation de

protéines [19], [20]. L'EEG est un signal complexe avec des dynamiques temporelles, spatiales et fréquentielles encore largement incomprises [21], [22]. L'apprentissage machine est un outil idéal pour en extraire une représentation utile sur le plan clinique. L'application de l'apprentissage machine à l'EEG pourrait augmenter l'utilité diagnostique et prognostique de cet examen, offrir une alternative quantitative et automatisée à l'interprétation visuelle et possiblement mener à une meilleure compréhension de la neurophysiologie de l'épilepsie.

Cette thèse adresse la question suivante: peut-on utiliser l'apprentissage machine pour modéliser le signal EEG en vue d'estimer le risque de crise d'un patient, sans dépendre de la présence de DÉI? Le projet se décline en trois étapes. Premièrement, nous mesurons les performances d'un modèle d'apprentissage machine basé sur l'extraction de caractéristiques classiques issues de la littérature. Par la suite, nous développons et évaluons des modèles profonds pour le diagnostic d'épilepsie. Troisièmement, nous améliorons l'architecture du modèle profond pour apprendre explicitement le risque de crise à travers le temps. Ces trois étapes sont rendues possibles grâce à une riche base d'EEG accompagnées de données cliniques bâtie sur mesure pour répondre à la question clinique au centre de cette thèse.

CHAPITRE 2 REVUE DE LITTÉRATURE

2.1 Le défi diagnostic de l'épilepsie

L'épilepsie est caractérisée par un risque élevé et persistant de crises épileptiques [2]. Ces crises correspondent à des épisodes transitoires d'activité neuronale excessive et anormalement synchrone, pouvant entraîner des manifestations neurologiques [3]. Environ 10% de la population subira une crise au cours de sa vie, sans nécessairement être atteint d'épilepsie [1], [23]. Après une seule crise, le risque de récidive est en moyenne de 40–50% [24]. Après deux crises non-provoquées espacées de plus de 24h, le risque de récidive s'élève à 73% à 4 ans [25]. Lorsque ce niveau de risque est atteint, le diagnostic d'épilepsie peut être posé, et un traitement est généralement indiqué. Certains facteurs indiquent un risque plus élevé après une crise unique et permettent de diagnostiquer l'épilepsie plus précocement. Ces facteurs incluent la présence de certaines lésions épileptogènes à l'imagerie cérébrale, un examen neurologique anormal et la présence d'anomalies épileptiformes à l'EEG [2].

L'EEG de routine est l'évaluation paraclinique primaire la plus importante chez les patients avec suspicion d'épilepsie [26], [27], [28]. Cet examen de 30 à 60 minutes consiste à enregistrer l'activité électrique du cerveau, principalement la sommation de potentiels synaptiques corticaux, à travers des électrodes apposées à la surface du scalp. Cette activité est amplifiée et digitalisée par un appareil d'enregistrement, puis analysé à l'œil par un neurologue certifié dans la lecture d'EEG (Figure 2.1). L'activité électrique est scrutée à la recherche de patrons anormaux comme un ralentissement des oscillations ou des pointes épileptiformes ou décharges épileptiformes intericales (DÉI; Figure 2.2). Les DÉI sont des décharges abruptes, brèves (20–200 ms), de haute amplitude, souvent suivies d'une onde lente, qui interrompent le rythme de fond [28], [29]. Ces décharges sont fortement associées à l'épilepsie: leur présence indique un risque de crise futur environ deux fois plus élevé après une première crise [30], [31], [32]. Elles permettent donc de poser un diagnostic chez un patient ayant eu une seule crise [2] et de caractériser le type et la localisation de l'épilepsie [33]. De plus, elles sont fréquemment utilisées comme outil diagnostic chez des patients avec une probabilité pré-test faible, pour lesquels il existe une incertitude quant à la nature de leurs épisodes [34], [35]. La présence de pointes viendra appuyer la suspicion d'épilepsie, et leur absence pourrait encourager la poursuite d'autres hypothèses diagnostiques.

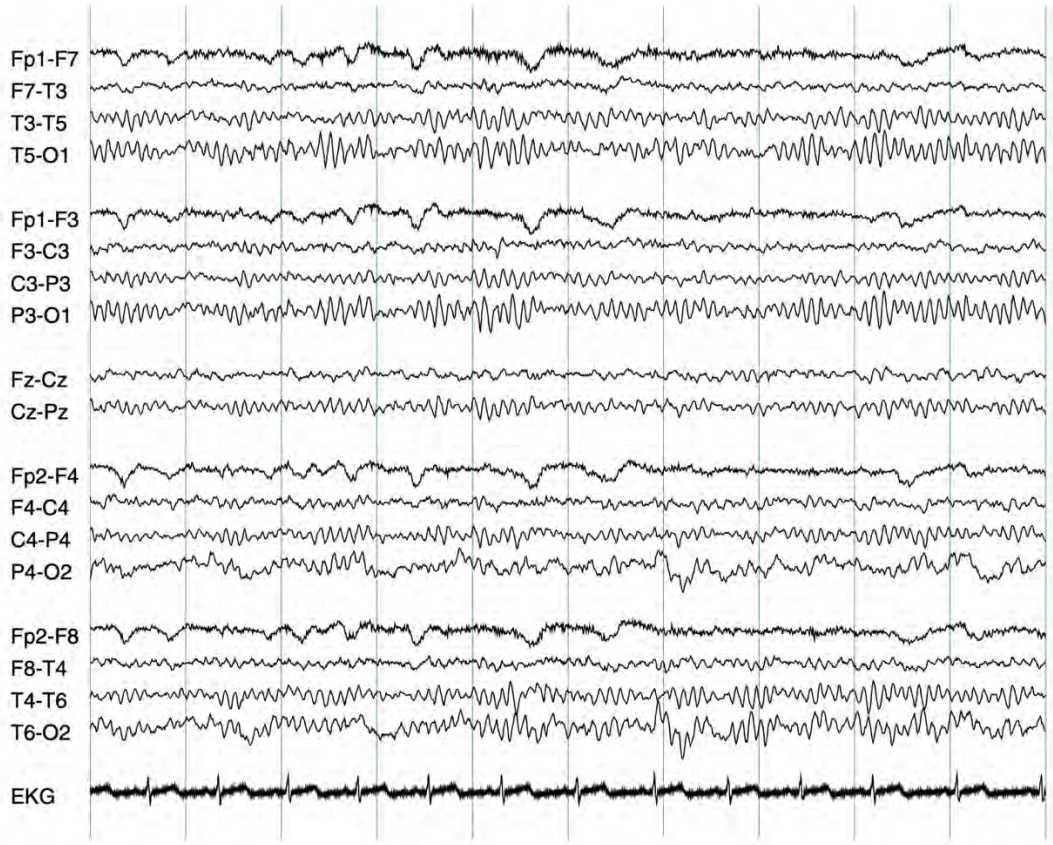


Figure 2.1 : Électroencéphalogramme à l'état d'éveil chez un patient sain. Image tirée de [36], © 2020, Elsevier, avec permission d'Elsevier.

La prédiction du risque de récidive de crise par l'EEG de routine est limitée par deux facteurs. Premièrement, sa faible sensibilité: des EEG de routine chez des patients avec épilepsie n'identifient des DÉI que dans 29–55% des cas [10], [28]. D'autre part, l'identification des pointes repose sur une analyse subjective de leur morphologie, localisation et fréquence, avec une fiabilité inter-observateur qui est au plus modérée [37], [38]. Par conséquence, des ondes pointues mais physiologiques peuvent être surinterprétées comme étant des anomalies épileptiformes [14]. La mauvaise interprétation de l'EEG est considérée comme un facteur majeur de surdiagnostic de l'épilepsie [39]. À cela s'ajoute une certaine prévalence des pointes épileptiformes chez les individus sans épilepsie, estimée à 1.74% (et allant jusqu'à 5.96% chez la personne âgée) [11]. De plus, la précision diagnostique de l'EEG dépend de la certitude clinique quant à la survenue d'une crise d'épilepsie, alors que pour beaucoup de patients, la nature même des épisodes est incertaine. Plusieurs causes alternatives peuvent expliquer des pertes de conscience ou symptômes neurologiques transitoires, et en l'absence de témoin ou de souvenirs francs des épisodes, il est

difficile d'être certain que ceux-ci soient des crises épileptiques. Dans les cliniques de « première crise », on conclut que l'épisode qui a amené le patient à consulter était non-épileptique dans 16–60% des cas, avec des diagnostics alternatifs qui incluent des syncopes, des accidents ischémiques transitoires, ou des crises psychogènes non-épileptiques [40], [41], [42], [43]. Pour ces raisons, le taux de mauvais diagnostic est élevé: le taux d'erreur diagnostic approche 20% dans la communauté [13], et près de 25% des patients référés pour épilepsie réfractaire (i.e., ayant eu un échec à plus de deux médicaments anticrises appropriés), après une évaluation approfondie, n'ont jamais eu d'épilepsie [44].

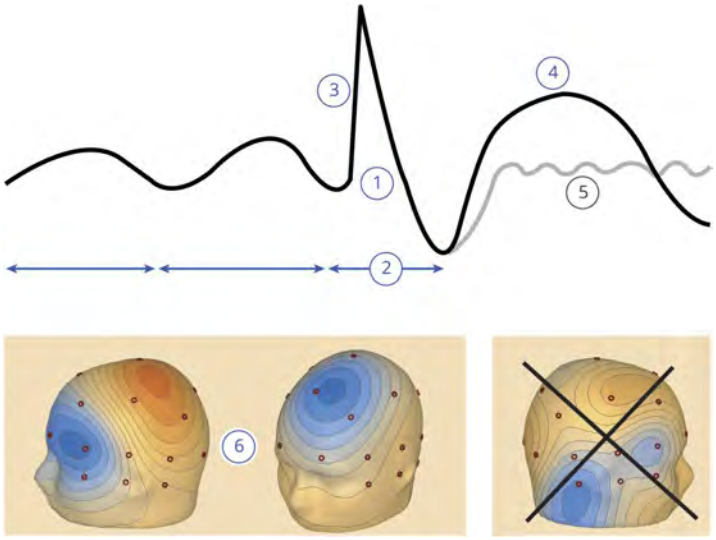


Figure 2.2 : Décharge épileptiforme intericale, ou pointe. Les annotations numériques dénotent les critères proposés par l'*International Federation of Clinical Neurophysiology* (IFCN) : 1) Morphologie di- ou triphasique et pointue, 2) fréquence différente du rythme de fond, 3) asymétrie, 4) onde lente subséquente, 5) perturbation du rythme de fond et 6) champ compatible avec une source cérébrale. Tirée de [29], © 2020 Wolters Kluwer Health, Inc., avec permission de Wolters Kluwer Health, Inc.

Les facteurs gouvernants le risque de crise, ou seuil convulsif, sont multifactoriels. Certains sont fixes et incluent un historique d'insulte cérébrale (e.g., accident vasculaire cérébral, encéphalite, trauma crânien), l'historique familial, l'âge, la présence de certaines lésions à l'imagerie, et la nature même des épisodes ayant mené à consulter [45], [46]. D'autres varient dans le temps, comme la dose et l'efficacité du traitement anticrise, les cycles hormonaux et circadiens, le stress et l'efficacité du sommeil [47], [48]. Un biomarqueur quantifiable du seuil convulsif améliorerait grandement la précision diagnostique des cliniciens pour l'épilepsie et pourrait aider à l'optimisation du traitement anticrise [49]. Par exemple, un risque de crise très élevé motiverait un

régime thérapeutique plus agressif, une référence précoce pour évaluation chirurgicale et un resserrement des consignes de sécurité en lien avec les occupations et la conduite automobile [49]. De plus, un tel biomarqueur améliorerait l'efficacité des études cliniques en identifiant précocement les patients à risque d'épilepsie réfractaire ou même ceux ayant une haute chance de rémission [50], [51].

À ce jour, il n'existe pas de biomarqueur prédictif du seuil épileptique validé pour l'usage clinique [2], [52], [53]. Plusieurs efforts se sont concentrés sur des analyses génétiques et métaboliques, la neuroimagerie et l'EEG. En génétique, certains gènes ou scores polygéniques sont fortement associés à l'épilepsie, et permettent donc d'estimer plus précisément le risque de crise future après une crise unique non-provoquée. Certaines mutations de gènes impliqués dans la régulation du potentiel de membranes neuronales peuvent directement abaisser le seuil convulsif [54]. Des études plus récentes sur l'ensemble du génome mettent en lumière certaines altérations génétiques plus subtiles [55], [56]. Dans Heyne et al., les auteurs proposent un score de risque polygénique qui prédit le développement d'une épilepsie généralisée génétique (rapport de risque [RR] = 1.73) ou focale (RR = 1.13) après une première crise [57]. Le principal désavantage des biomarqueurs génétiques est leur aspect statique: ils ne permettent pas d'évaluer la variation du seuil dans le temps. Des biomarqueurs sériques comme les microARN circulants [58], [59] ou métabolites tels que le glutamate, lactate et citrate [60] pourraient bien saisir l'aspect dynamique du seuil épileptique, mais leur développement est miné par la faible spécificité, la susceptibilité au traitement anticrise et l'utilisation de devis cas-contrôle [60]. La neuroimagerie structurelle et métabolique est une autre source prometteuse de biomarqueur d'épilepsie, mais les efforts sont concentrés sur les patients avec épilepsie réfractaire pour l'identification de lésions subtiles et la prédiction de la réponse à la chirurgie [61], [62], [63].

L'EEG demeure un candidat idéal pour la découverte de biomarqueurs prognostiques [53]. Premièrement, il est effectué chez virtuellement tout patient avec suspicion ou diagnostic d'épilepsie, et donc n'est pas sujet à des biais de sélection en fonction de la probabilité pré-test. Cela permet d'effectuer des études rétrospectives à grande échelle. Deuxièmement, l'EEG échantillonne directement le processus considéré pathologique chez les patients avec épilepsie: l'activité électrique cérébrale. L'EEG peut être appliqué sur de longue périodes (minutes à heure et même à jours) et possède une résolution temporelle assez élevée comparée aux autres techniques de neuroimagerie [64], [65]. Finalement, le test est standardisé et disponible partout dans le monde,

avec une barrière technologique basse [31], [66]. Une de ses faiblesses majeures réside dans son interprétation visuelle et la dépendance sur les pointes épileptiformes. Au-delà de sa subjectivité et de la nécessité d'une formation surspécialisée, l'interprétation visuelle fait fi de plusieurs caractéristiques cachées dans le signal, incluant les dynamiques temporelles complexes, des tendances sur longues échelles temporales (>10–20s, la taille de fenêtre typiquement utilisée pour analyser l'EEG de routine) et les interactions de haut-niveau entre les régions cérébrales. Dans une ère où les algorithmes peuvent converser avec les humains, rédiger des textes complexes et même analyser avec précision des imageries médicales, est-ce qu'il est possible d'exploiter davantage la richesse du signal capté par l'EEG et en tirer une information utile cliniquement?

2.2 Des biomarqueurs computationnels d'épilepsie à l'EEG: revue systématique

L'analyse automatisée de l'EEG à la recherche de biomarqueurs d'épilepsie intéresse les chercheurs en neurosciences, informatique et génie depuis plusieurs décennies. Plusieurs caractéristiques du signal EEG sont différentes dans le cerveau d'un patient épileptique comparé à des sujets sans épilepsie. Ces marqueurs incluent des caractéristiques du réseau de connectivité fonctionnelle, la complexité et prédictibilité du signal, la puissance spectrale et la chaoticité. L'analyse automatisée permettrait d'extraire cette information qui est invisible à l'œil nu, de manière quantitative et rapide, améliorant l'utilité diagnostique et prognostique de l'EEG. Malgré ces promesses, aucun de ces marqueurs n'est utilisé en clinique. Nous avons effectué une revue systématique pour évaluer la performance diagnostique de ces marqueurs en épilepsie, les populations étudiées et la qualité méthodologique des études, afin de comprendre pourquoi leur impact clinique est faible et émettre des recommandations pour les projets futurs. Je résume ici les résultats de la revue systématique, disponible en annexe (ainsi que son protocole publié dans un journal avec revue par les pairs) [67], [68], suivi d'une mise à jour sur les travaux publiés depuis.

Notre revue systématique inclus toute étude rétrospective ou prospective comparant au moins un biomarqueur computationnel pour le diagnostic d'épilepsie sur l'EEG de routine (<24h), sans dépendre uniquement de la détection de pointes ou de crises. La population d'intérêt incluait des individus qui ont un EEG de routine dans le cadre de la clinique ou de la recherche, sans restreindre pour l'âge, les comorbidités, ou la médication. Le standard de référence était le diagnostic

d’épilepsie évalué cliniquement, tel que défini par les critères de la Ligue Internationale contre l’Épilepsie [2].

Notre stratégie de recherche, mise sur pied par deux bibliothécaires médicales, a identifié plus de 10 000 articles publiés entre 1961 et 2022. Deux évaluateurs indépendants ont effectué le triage puis la sélection des études. Deux autres évaluateurs indépendants ont procédé à l’extraction des données et à l’évaluation méthodologique basée sur l’outil QUADAS-2 et adapté à la question de recherche de la revue [68]. Ces derniers évaluateurs ont aussi évalué la reproductibilité des études.

Après le triage et la sélection, 37 études ont été incluses. Ces études ont exploré différentes approches d'analyse du signal EEG: des méthodes linéaires (43% des études), non-linéaires (27%), de connectivité (38%) et d'apprentissage profond (10%). La taille des échantillons variait considérablement, avec une moyenne de 54 participants et seulement six études incluant plus de 100 sujets (Figure 2.3). La majorité des études incluaient à la fois des enfants et des adultes, et environ deux tiers des études incluaient tout type d'épilepsie.

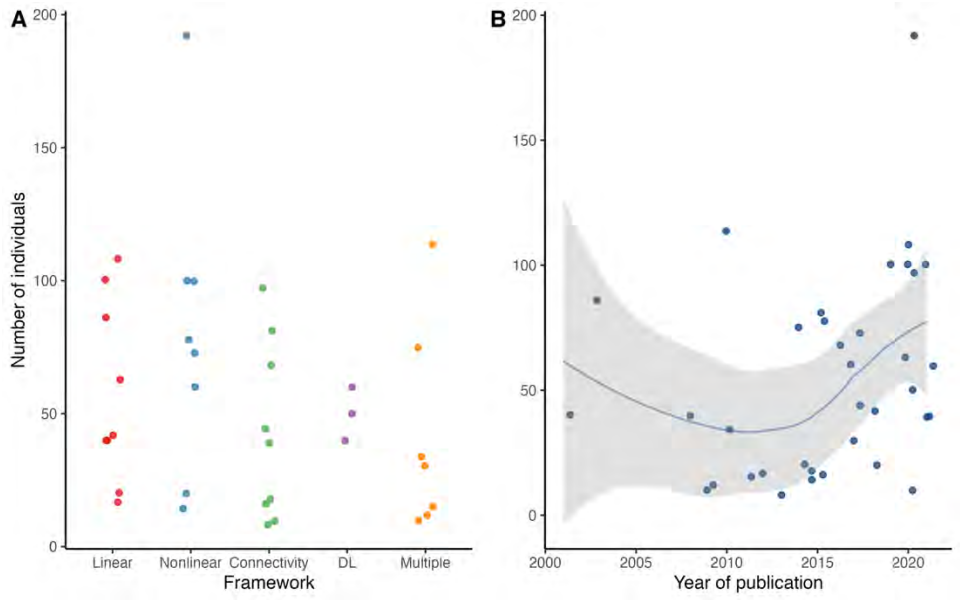


Figure 2.3 : Tailles d’échantillon des études incluent dans la revue systématique en fonction du type de marqueur (A) et de l’année de publication (B). Image tirée de [67] sous licence CC BY 4.0.

Les méthodes linéaires se sont concentrées sur l'analyse spectrale, notamment la puissance relative dans les bandes de fréquence. Des différences entre patients avec épilepsie et sans épilepsie ont été détectées dans toutes les bandes de fréquences: delta (≤4 Hz), thêta (4–8 Hz), alpha (8–13 Hz), beta

(13–40 Hz) et gamma (≥ 40 Hz), ainsi que les sous-bandes alpha [69], [70], [71], [72], [73], [74], [75]. D'autres approches ont exploré la stationnarité du signal et ses moments statistiques d'ordre supérieur (« High-order spectrum »)[71], [75]. Les paramètres de Hjorth, qui quantifient les moments statistiques du signal dans les domaines temporel et fréquentiel [76], semblaient discriminants dans deux études. Une approche originale consistait à détecter des événements lents paroxystiques (« Paroxysmal Slow Wave Events »), définis comme des segments de 2 secondes avec une fréquence médiane inférieure à 6 Hz, montrant une aire sous la courbe ROC (AUROC) de 0.72 pour prédire le risque de récidive à 18 mois après une première crise [77].

Les méthodes non-linéaires ont cherché à caractériser la complexité et la prévisibilité du signal EEG. L'entropie, sous ses différentes formes (Shannon, spectrale, approximative, de permutation), a été étudiée dans sept études [78], [79], [80], [81], [82], [83], [84]. Ces mesures tentent de quantifier le degré d'organisation ou de chaos dans l'activité cérébrale, supposément altéré chez les patients avec épilepsie. Certaines études ont calculé l'entropie après avoir filtré différentes bandes de fréquences, permettant d'évaluer l'entropie sur différentes échelles temporelles. D'autres caractéristiques non-linéaires incluaient les dimensions fractales [75], [78], l'exposant de Hurst [79], l'analyse des intervalles de passage à zéro (« zero-crossings interval analysis ») [85] et l'analyse quantitative de récurrence [83].

L'analyse de la connectivité fonctionnelle a été utilisée par 14 études [69], [78], [81], [86], [87], [88], [89], [90], [91], [92], [93], [94], [95], [96]. Ces méthodes évaluent les interactions entre différentes régions cérébrales, soit par des mesures de synchronisation de phase, soit par des analyses de causalité. La plupart des études ont utilisé une analyse basée sur les capteurs plutôt que sur les sources, avec diverses mesures de connectivité comme l'information mutuelle [86], la cohérence [86], la valeur de verrouillage de phase [86], [88] et la causalité de Granger [87]. Une approche rapportée par deux études consistaient en un modèle des interactions entre les régions cérébrales inspiré de l'oscillateur de Kuramoto afin d'isoler les paramètres qui dictent la propension du réseau à générer des crises [88], [93]. Toutes les études ont analysé la connectivité sur plusieurs bandes de fréquence. Après avoir estimé la force de connectivité entre chaque électrode, la majorité des études ont extraits des caractéristiques qui décrivent la topologie du réseau. Trois études ont directement utilisé les matrices de connectivité comme entrée dans un algorithme de classification [87], [94], [95]. Les caractéristiques de réseau des patients avec

épilepsie étaient peu reproductibles d'une étude à l'autre. La seule caractéristique qui était associée à l'épilepsie dans toutes les études l'ayant étudiée était l'efficacité du réseau [89], [96], [97].

Quatre études ont utilisé des réseaux de neurones à convolution (CNN), avec des architectures variant d'une simple couche de convolution à trois blocs de deux couches convolutives [73], [74], [87], [98]. Certaines études ont prétraité les données en matrices de connectivité [87] ou en représentations temps-fréquence [73], tandis que d'autres ont directement utilisé le signal EEG brut en segments de 2 à 10 secondes [74], [98]. Les réseaux étaient de taille modeste, variant d'environ 3 000 à 92 000 paramètres. L'augmentation du chevauchement des segments lors de l'analyse a semblé améliorer les performances, possiblement en augmentant artificiellement la taille de l'ensemble d'entraînement [74].

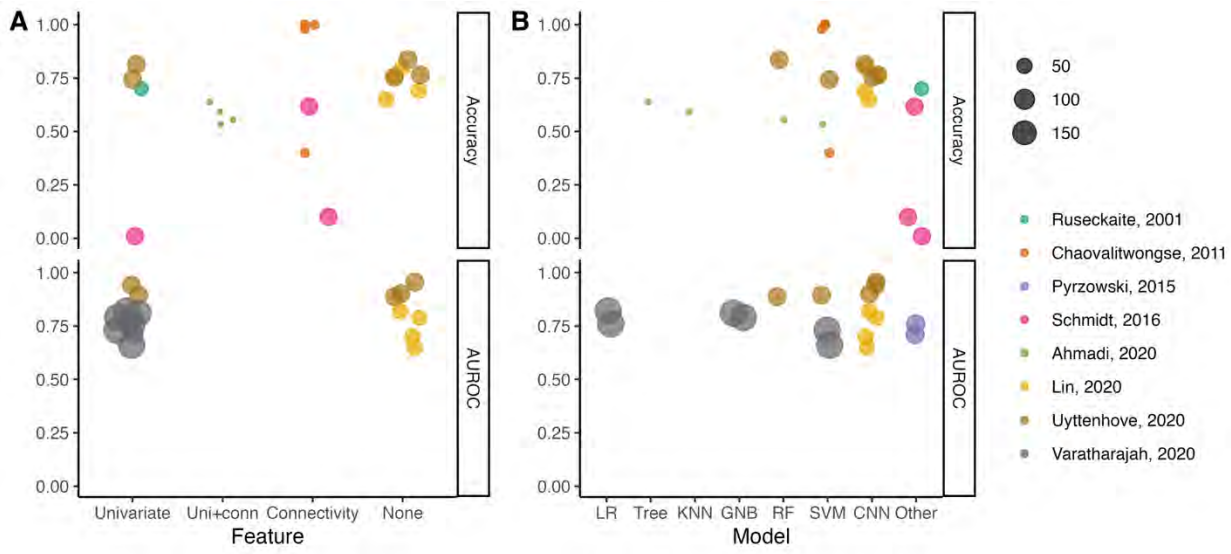


Figure 2.4 : Performances diagnostiques des études incluent dans la revue systématique et exemptent de fuite de données entre les ensembles d'entraînement et de test. A : Performances en fonction du type de caractéristiques. B : Performance en fonction du modèle d'apprentissage. Image tirée de [67] sous licence CC BY 4.0.

Cependant, malgré des performances diagnostiques rapportées allant de 64% à 100%, l'évaluation rigoureuse de ces biomarqueurs est sévèrement limitée par d'importantes faiblesses méthodologiques. Aucune étude n'a démontré un faible risque de biais selon les critères standardisés d'évaluation QUADAS-2 [68], [99]. Les problèmes méthodologiques majeurs incluaient la sélection non-représentative des patients (utilisation fréquente d'un plan d'étude cas-témoins plutôt qu'une cohorte consécutive), les fuites de données dans la validation des

performances (partage d'information entre les ensembles d'entraînement et de test) et le manque de reproductibilité des analyses. La sélection manuelle des segments EEG, effectuée dans 54% des études, introduisait une source additionnelle de subjectivité dont l'impact sur les performances n'a pas été quantifié. Seules six études (16%) ont été jugées reproductibles.

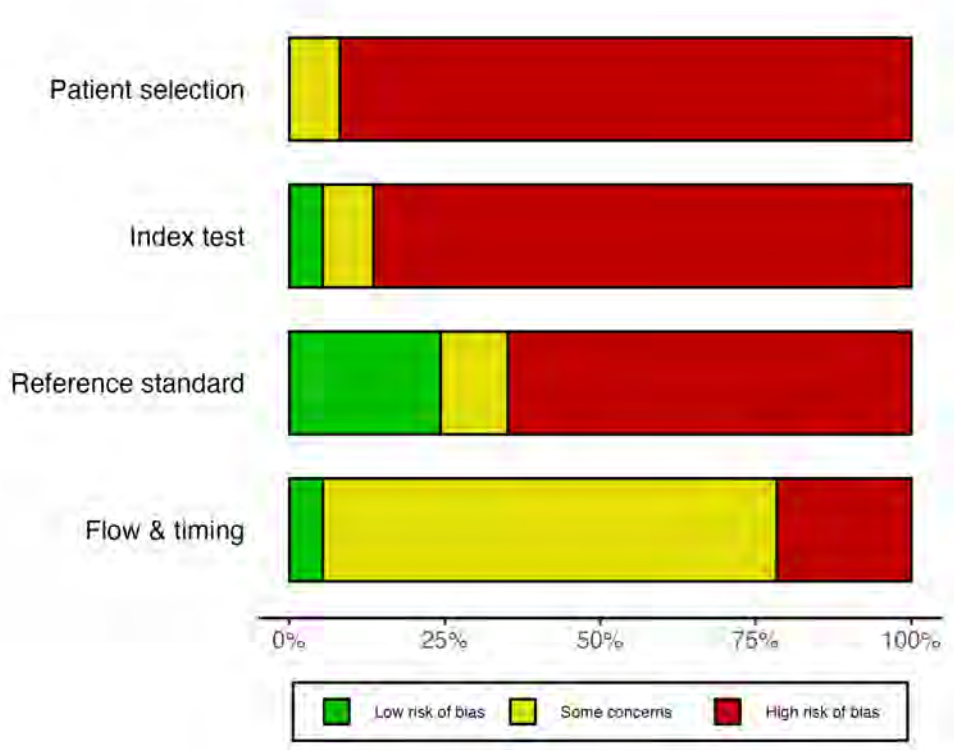


Figure 2.5 : Résumé du risque de biais des études de la revue systématique par domaine PRISMA. Image tirée de [67] sous licence CC BY 4.0.

Ces limitations empêchent actuellement de conclure sur l'utilité clinique réelle de ces biomarqueurs computationnels. Des études futures devront adopter une méthodologie plus rigoureuse, incluant des populations cliniquement pertinentes, une validation externe robuste et une documentation détaillée des méthodes d'analyse. Les approches basées sur l'apprentissage profond, bien que prometteuses, nécessiteront des bases de données considérablement plus grandes pour atteindre leur plein potentiel. Les futures études devront également prioriser l'automatisation complète du traitement des signaux EEG, incluant la détection et le rejet des artéfacts, pour faciliter l'application clinique de ces méthodes. L'avènement de grandes bases de données d'EEG standardisées et l'amélioration continue des techniques d'apprentissage automatique laissent espérer des avancées

significatives dans ce domaine, qui pourraient ultimement compléter l'analyse visuelle traditionnelle.

2.2.1 Mise à jour sur le diagnostic automatisé de l'épilepsie à l'EEG

Depuis la recherche effectuée dans le cadre de la revue systématique présentée à la section précédente, quelques articles additionnels ont tenté d'utiliser l'analyse automatisée de l'EEG pour améliorer le diagnostic de l'épilepsie.

Myers et al. ont développé "EpiScalp", un modèle de régression logistique basé sur des caractéristiques spectrales et de connectivité, pour différencier l'épilepsie des conditions mimiques comme les crises psychogènes non-épileptiques [100]. Sur une cohorte de 218 patients admis à l'unité de monitoring d'épilepsie dans différents centres hospitaliers américains avec un EEG de routine préalable normal, les auteurs ont sélectionné les patients ayant un épisode habituel lors de l'admission permettant de confirmer ou infirmer le diagnostic d'épilepsie. Par la suite, ils ont identifié des marqueurs de connectivité et de puissance spectrale permettant de distinguer les patients avec et sans épilepsie avec une AUROC de 0.94 en validation croisée et une précision de 80% sur un ensemble test. Les critères d'inclusion stricts (confirmation diagnostique par enregistrement d'un épisode habituel en vidéo-EEG prolongé) diminuent cependant la généralisabilité de leurs résultats à une population plus large. Cette performance doit aussi être interprétée avec prudence étant donné la petite taille de l'ensemble de test (20 patients) et un possible biais de sélection des caractéristiques, celle-ci ayant été effectuée sur l'ensemble des données.

Une autre étude multicentrique par Tait et al. a validé un ensemble de biomarqueurs computationnels sur 281 EEG consécutifs normaux collectés dans huit centres au Royaume-Uni [101]. Le diagnostic d'épilepsie était basé sur les notes médicales, avec un suivi d'au moins un an. Leur approche combinait huit biomarqueurs: deux mesures spectrales, quatre de réseau et deux basées sur des modèles dynamiques. Leur modèle atteignait une précision balancée de 68% sur un ensemble test (sensibilité 61%, spécificité 75%). Les auteurs ont aussi confirmé l'absence d'impact de variables confondantes comme l'âge, le genre, le statut de traitement et les comorbidités. Cette étude présente une méthodologie robuste comparativement aux études précédemment décrites. Elle est tout de même limitée par l'absence d'un ensemble test et le manque de généralisabilité aux EEG anormaux, pour lesquels les performances étaient non-significatives.

Faiman et al. se sont concentrés spécifiquement sur la différenciation entre épilepsie et crises psychogènes non-épileptiques chez 148 patients naïfs au traitement [102]. Des segments visuellement normaux de 20s étaient sélectionnés pour l'analyse. Leur approche était divisée en deux: une première étude testait des biomarqueurs précédemment rapportés (puissance thêta et fréquence alpha pic), alors qu'une seconde étude explorait de nouvelles caractéristiques avec un algorithme de sélection de caractéristiques (pour un total de 7 729 caractéristiques). Ni l'approche guidée par hypothèse (précision moyenne 48%) ni l'approche exploratoire (précision 45-60%) n'ont permis d'identifier des biomarqueurs robustes. Cette étude négative souligne la difficulté de la tâche en question. Les faiblesses méthodologiques incluent une taille d'échantillon trop faible pour l'apprentissage, l'utilisation d'un segment unique de 20s pour chaque EEG et une possible sous-optimisation des hyperparamètres.

En parallèle, deux études majeures ont démontré le potentiel de l'apprentissage profond pour l'interprétation automatisée de l'EEG, particulièrement pour la détection de DÉIs. Bien que cette tâche soit différente du diagnostic, ces deux études présentent des parallèles intéressants avec les objectifs de cette thèse comme l'utilisation de modèles profonds, une méthodologie robuste et des données à grande échelle. Tveit et al. ont développé SCORE-AI, un réseau de neurones à convolution entraîné sur plus de 30 000 EEG pour classifier les enregistrements en catégories cliniquement pertinentes (normal, épileptiforme focal/généralisé, non-épileptiforme focal/diffus) [103]. Le modèle a atteint des performances similaires aux experts humains avec une aire sous la courbe ROC entre 0.89 et 0.96 selon la catégorie. La validation externe sur trois jeux de données indépendants, incluant un ensemble multicentrique de 100 EEG évalués par 11 experts et un ensemble monocentrique de 9 785 EEG, démontre la robustesse du modèle. Jing et al. ont créé SpikeNet, un réseau de neurones profond pour la détection automatisée de DÉI [104]. Entraîné sur plus de 9 500 EEG annotés par huit neurophysiologistes certifiés, SpikeNet a surpassé les performances humaines et les solutions commerciales existantes, avec une erreur de calibration plus faible (0.041 vs 0.183 pour les experts) et une meilleure discrimination (AUROC 0.98 vs 0.88 pour le standard commercial). Bien que ces deux études ne visent pas directement le diagnostic d'épilepsie, elles démontrent la faisabilité d'analyser de grandes bases de données d'EEG avec l'apprentissage profond.

Ces études récentes illustrent l'intérêt croissant pour la recherche de biomarqueurs alternatifs à l'EEG à l'aide de méthodes computationnelles. Les approches basées sur l'extraction de

caractéristiques prédéfinies montrent des résultats encourageants mais variables, possiblement limités par la taille des cohortes et des biais méthodologiques. Notamment, la difficulté à reproduire certains biomarqueurs précédemment rapportés souligne l'importance d'une validation externe rigoureuse et met en garde contre la surinterprétation d'études pilotes sur de petits échantillons sur-sélectionnés. L'apprentissage profond émerge comme une approche prometteuse, particulièrement lorsqu'entraîné sur de grandes bases de données. Cependant, son application au diagnostic d'épilepsie reste à démontrer.

2.3 L'apprentissage profond et l'EEG

Au-delà de l'épilepsie, l'apprentissage profond promet de révolutionner l'analyse de l'EEG, avec des applications allant de l'interface cerveau-machine à la détection des stades de sommeil [105], [106]. En épilepsie, les avancées significatives se concentrent principalement à la détection de DÉI qui bénéficie d'ensembles de données comptant des dizaines de milliers d'EEG [103], [104]. Le choix de modèle optimal pour la classification de signaux EEG est un sujet de recherche très actif. Actuellement, deux familles de modèles dominent la littérature récente sur l'apprentissage profond et l'EEG: les réseaux de neurones à convolution et les Transformeurs.

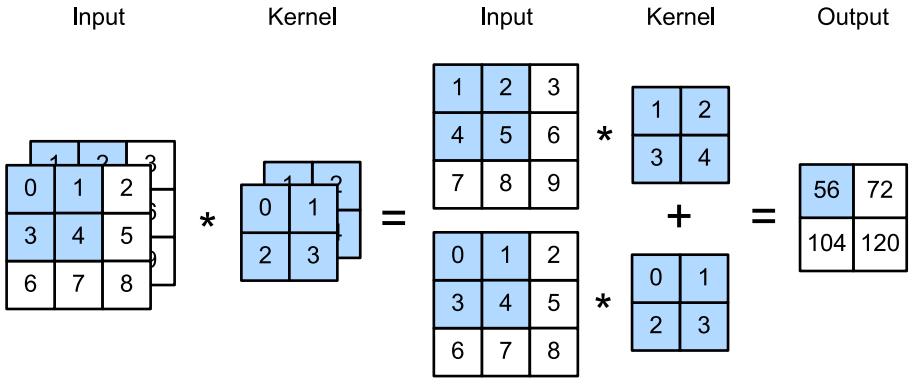


Figure 2.6 : Exemple de l'opération de « convolution » (qui est en réalité une corrélation croisée) avec plusieurs canaux. Image tirée de [107] sous licence CC BY-SA 4.0.

Le CNN est l'approche la plus courante pour la modélisation du signal EEG [67], [105], [108], [109]. Introduit en 1998 avec LeNet-5 pour la reconnaissance de caractères manuscrits [110], le CNN combine trois concepts clés: les couches de convolutions, les fonctions d'activation non-linéaires et le *pooling* (Figure 2.6 et Figure 2.7). Les couches de convolutions consistent à faire « glisser » un filtre sur l'entrée. Les paramètres du filtre sont optimisés au cours de l'entraînement.

Cela limite les connexions possibles entre les paramètres d'entrée (appelé champ réceptif ou *receptive field*) et confère une structure hiérarchique aux CNN, où les premières couches captent des caractéristiques locales alors que les couches plus profondes, des caractéristiques plus générales [111]. Les convolutions permettent aussi le partage de poids, ce qui réduit drastiquement le nombre de paramètres et force le réseau à apprendre des filtres invariants à la translation. Les fonctions d'activation non-linéaires, comme la fonction sigmoïde ou le *Rectified Linear Unit* (ReLU), introduisent la non-linéarité nécessaire pour modéliser des relations complexes et facilite l'entraînement du réseau. Le *pooling* agrège l'information spatiale locale avec une fonction max ou moyenne, réduisant la dimensionnalité et augmentant la robustesse aux petites transformations. À travers le CNN, les couches de *pooling* réduisent progressivement la dimensionnalité de la représentation. Ces propriétés définissent le biais inductif qui confère aux CNN une capacité naturelle à identifier les relations à travers diverses échelles temporelles et spatiales et leur confère une robustesse aux jeux de données modestes [112].

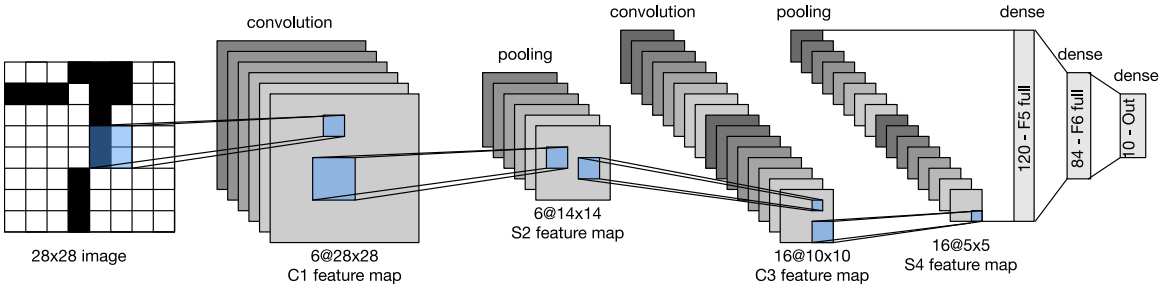


Figure 2.7 : Le modèle LeNet, reconnu comme un pionnier des réseaux de neurones à convolution. Image tirée de [110] sous licence CC BY 4.0.

Depuis son invention, plusieurs changements importants ont été apportés au CNN, lui permettant d'être entraîné efficacement et avec stabilité sur des jeux de données massifs. En 2012, AlexNet a démontré la puissance des réseaux profonds en remportant la compétition ImageNet avec une marge considérable. La principale contribution était la profondeur du modèle, qui comportait 60 millions de paramètres (contre 60 000 pour LeNet de 1998) et rendue possible grâce aux avancées en optimisation sur GPU [15]. Par la suite, VGGNet (2014) a standardisé l'architecture en utilisant des blocs répétitifs de petites convolutions (3x3) et de *pooling*, démontrant qu'une succession de

blocs de convolutions avec un petit champ récepteur est plus performante qu'une séquence plus courte de grandes convolutions [113]. Finalement, ResNet (2015) a révolutionné l'entraînement des réseaux très profonds en introduisant les connexions résiduelles (Figure 2.8), permettant au gradient de circuler plus efficacement à travers le réseau [114]. Ces avancées architecturales ont établi les CNN comme l'état de l'art dans pratiquement tous les domaines de vision par ordinateur [115], [116], [117] et ont inspiré leur application à d'autres types de données comme l'EEG.

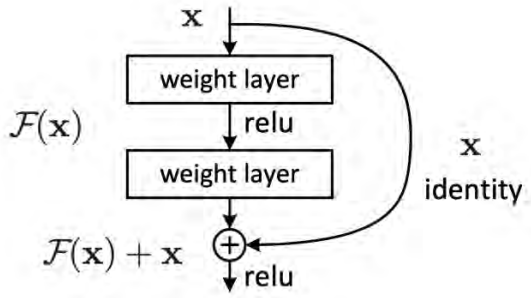


Figure 2.8 : Connection résiduelle au sein d'un réseau de neurones. Image tirée de [114]. © 2016, IEEE

Pour l'EEG, les deux CNN phares sont l'EEGNet [118] et le ShallowConvNet [119], [120]. Ces deux modèles ont en commun l'utilisation séquentielle de *point-wise* convolution dans le domaine temporel puis spatial (Figure 2.9). Cette approche est inspirée de l'algorithme *Filter Bank Common Spatial Patterns* (FBCSP) utilisé pour décoder le signal EEG dans les interfaces cerveaux-machine [121] et permet de réduire considérablement le nombre de paramètres dans leur modèle. Dans un article récent [122], les auteurs du ShallowConvNet ont comparé les performances de ces algorithmes pour l'EEG de routine sur un jeu de données publiquement disponible, le *Temple University Hospital EEG corpus* [123], [124]. Il s'agit d'un ensemble de 10 700 EEG provenant de 8 710 patients avec le rapport correspondant. Pour certains sous-ensembles, les EEG ont été étiquetés à partir du rapport d'EEG (e.g., EEG globalement normal vs. anormal, patient avec épilepsie vs. sans épilepsie). Il est important de noter que les critères d'inclusion de ces EEG ne sont pas spécifiés, et les étiquettes sont à haut risque de biais puisqu'elles sont attribuées sur la base du rapport EEG et non d'une revue plus compréhensive du dossier longitudinal des patients. Dans leur analyse, les auteurs se sont intéressés à la performance des différents CNN en fonction de la taille de l'échantillon d'entraînement. Ils ont démontré que les performances de classification suivent une loi de puissance avec saturation, en fonction de la taille du modèle et de la taille du jeu

de données. Selon leurs conclusions, les modèles présentés saturent après entraînement sur 2 000 exemples, avec une augmentation de la précision de seulement 1.5% lorsqu'entraînés sur 8 000 autres exemples. Les limitations importantes de leur étude sont le biais dans l'étiquetage des données et la taille maximale de leur modèle (1.7M de paramètres).

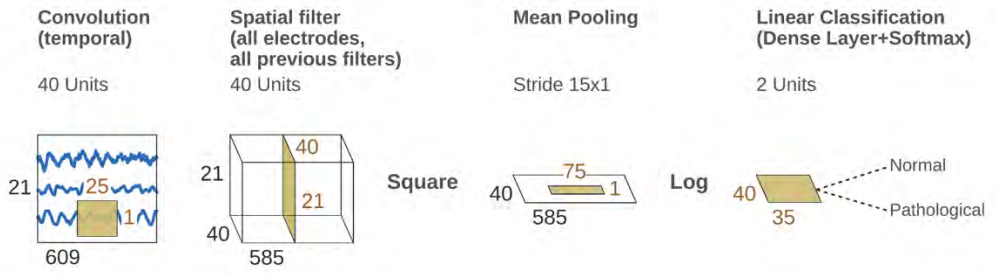


Figure 2.9 : Le modèle ShallowConvNet pour la classification d'EEG. Image tirée de [125]. © 2017, IEEE

Une autre architecture qui gagne en popularité pour l'EEG est le Transformeur. Le Transformeur a initialement été conçu pour modéliser les séries comme les séries temporelles ou le texte. Auparavant, les réseaux de neurones privilégiés pour les données en séries étaient les Réseaux de Neurones Récurrents (RNN) comme les réseaux Long Short-Term Memory (LSTM). Ces réseaux traitent les données séquentiellement en maintenant un état caché (*hidden state*) qui garde en mémoire l'information des états précédents [112]. Les RNN souffrent de trois limitations majeures: premièrement, le traitement séquentiel limite la parallélisation. Deuxièmement, ils sont sujets au problème d'évanescence des gradients où le signal d'apprentissage s'atténue exponentiellement lors de la rétropropagation à travers le temps ; et troisièmement, malgré les mécanismes de portes des LSTM, ces réseaux peinent à identifier les dépendances à long terme car l'information doit traverser de nombreuses étapes de traitement [112]. Les Transformeurs adressent ces problèmes grâce à leur mécanisme d'attention [17]. L'attention permet à chaque élément de la séquence d'interagir directement avec tous les autres éléments, facilitant ainsi la modélisation des dépendances temporelles sans les contraintes des architectures récurrentes (Figure 2.10). De plus, cette architecture permet un traitement parallèle, accélérant significativement l'entraînement. Les Transformeurs sont aussi reconnus pour leur extensibilité: des jeux de données plus volumineux et des modèles plus complexes résultent en de meilleures performances, et ce avec une utilisation plus efficace des ressources computationnelles [126]. Ces avantages en ont fait l'architecture dominante

pour les tâches complexes de traitement du langage naturel, et leur succès s'étend maintenant à d'autres domaines comme l'analyse de signaux EEG [127], [128], [129].

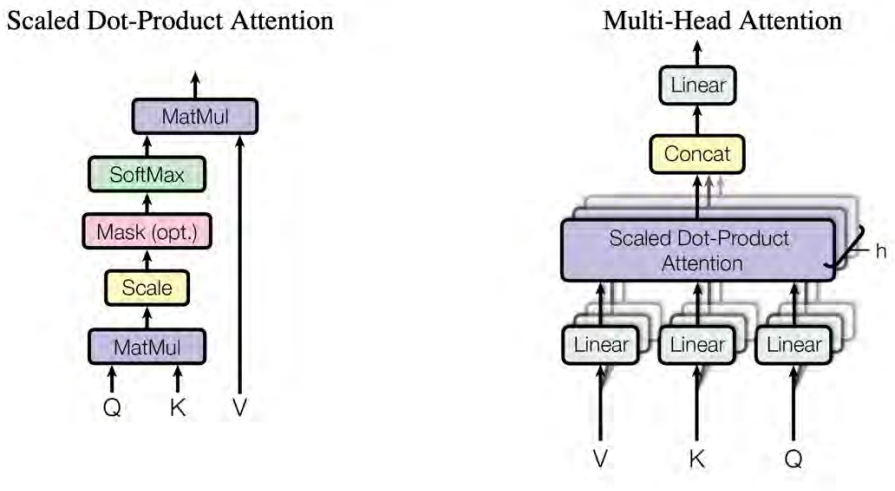


Figure 2.10 : À gauche : Attention avec produit scalaire. À droite : Attention *multi-head*, où les multiples « têtes » sont des modules d'attention avec produit scalaire parallèles. Image tirée de [17] sous licence CC BY-SA 4.0.

Le *Vision Transformer* (ViT) est une adaptation du Transformeur à l'imagerie [130]. Son principe est de découper l'image en morceaux (*patches*) qui sont ensuite organisés en séquence et traités comme des lexèmes (*tokens*), similaires aux mots dans le traitement du langage (Figure 2.11). Cette approche permet d'appliquer directement les mécanismes d'attention aux données visuelles, en considérant chaque « patch » comme une unité distincte pouvant interagir avec toutes les autres. Depuis la publication du ViT, plusieurs améliorations ont été proposées pour faciliter leur entraînement, particulièrement aux jeux de données plus modestes. Le *Compact Vision Transformer* (CvT) combine une tokenisation par convolution avec un Transformeur compact, et il peut atteindre des performances état-de-l'art avec aussi peu que 0.28M paramètres [131], [132]. Un autre élément clé pour l'entraînement efficace du Transformeur est la régularisation.[133] Celle-ci peut être obtenue avec des techniques d'optimisation comme le *weight decay* ou le *dropout*, mais surtout via des techniques d'augmentation de données comme le masquage aléatoire et la rotation [133], [134]. Cette régularisation est essentielle pour adapter les Transformeurs aux jeux de données limités [135]. Cette caractéristique est particulièrement pertinente pour l'analyse de l'EEG, où les données étiquetées sont souvent rares et où l'augmentation de données doit être appliquée avec précaution pour préserver la signification physiologique du signal.

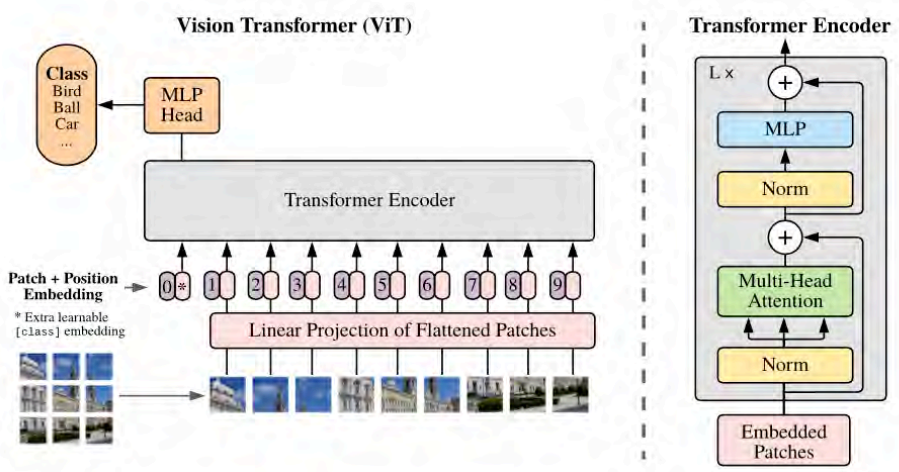


Figure 2.11 : Aperçu du *Vision Transformer*. Image tirée de [130] sous licence CC BY 4.0.

En EEG, le Transformeur a été utilisé pour le suivi des mouvements extra-oculaires [128], la prédiction de crises en temps réel [136], [137], la reconnaissance d'émotions [127] et le décodage des patrons moteurs [129]. Toutes ces études utilisent le ViT, mais varient dans leur méthode pour découper le segment d'EEG d'entrée. Deux études, ViT2EEG et EEGFormer [127], [128], utilisent les convolutions *point-wise* pour encoder le signal avant l'entrée dans le Transformeur, inspiré par les ShallowConvNet et EEGNet présentés plus tôt [118], [119]. Les deux autres études utilisent comme tokenizer soit une projection linéaire comme le ViT original [130], ou bien une couche de convolution avec activation ReLU. Trois des études utilisent le signal EEG sous la forme C x T [127], [128], [137], alors que la quatrième transforme le signal en spectrogramme de la forme C x F x T (dans leur cas, C = 1) [136]. Bien que chacun des articles démontrent une bonne performance du ViT pour l'EEG, les tailles de jeux de données se limitent à 15 000–80 000 segments provenant de 23–70 patients et les segments d'EEG sont courts (jusqu'à 50 000 points ou 10 secondes). Une autre initiative notable est le *Brain Foundation Model* par Bayazi et al. [138]. Ce Transformeur applique la complexité d'un large modèle de langage à l'analyse de l'EEG et de l'IRM fonctionnelle. Cependant, le modèle traite un canal à la fois, et donc ne peut pas modéliser les relations spatiales inhérentes à l'EEG. Ainsi, la capacité réelle du Transformeur à modéliser un large ensemble de données EEG pour une tâche clinique reste incertaine.

Malgré ces avancées prometteuses, plusieurs limitations importantes persistent dans l'application de l'apprentissage profond à l'EEG. Premièrement, le plein potentiel de ces algorithmes pour le diagnostic d'épilepsie reste largement inexploré: les plus grands modèles actuels, qui ne comptent

que quelques dizaines de milliers de paramètres,[73], [74], [87], [98] sont modestes comparés aux modèles état de l'art dans d'autres domaines. Deuxièmement, de nombreuses questions méthodologiques fondamentales restent sans réponse claire: le choix optimal entre Transformeurs et CNN, la méthode de tokenisation la plus appropriée, l'utilisation de spectrogrammes versus signaux bruts, ou encore la longueur optimale des segments à analyser. Enfin, ces modèles profonds restent largement des "boîtes noires": nous avons une compréhension limitée des caractéristiques qu'ils extraient des signaux EEG et de la façon dont ils prennent leurs décisions. Cette opacité constitue un obstacle majeur à leur adoption clinique et souligne le besoin de développer des méthodes d'interprétabilité robustes.

CHAPITRE 3 OBJECTIFS ET ORGANISATION DE LA THÈSE

Dans le chapitre précédent, nous avons démontré qu'il y a un besoin criant de biomarqueurs du risque de crise en épilepsie. L'EEG pourrait contenir une information pertinente invisible à l'œil nu mais détectable via des méthodes quantitatives. Cependant, les études actuelles sont limitées par des biais méthodologiques reliés à la sélection des patients et la validation. L'apprentissage profond est une approche intéressante pour modéliser l'EEG, mais sa capacité à détecter des caractéristiques liées au diagnostic d'épilepsie et au risque de crise est incertaine.

Cette thèse tente de répondre à la question suivante: Comment extraire de l'EEG des biomarqueurs quantitatifs et robustes du risque de crise en épilepsie, indépendamment de la présence de DÉI? Plus spécifiquement:

1) Quelles caractéristiques décrites dans la littérature permettent d'identifier les patients à risque de crise de façon indépendante des DÉI, et quelle est leur précision diagnostique?
2) Est-ce que l'apprentissage profond peut améliorer les performances diagnostiques et prognostiques comparés à l'extraction de caractéristiques et à la présence de DÉI?

Pour répondre à ces questions, les objectifs suivants ont été établis:

1) Développer une base de données d'EEG de routine avec données cliniques détaillées provenant de patients consécutifs, permettant la découverte et la validation de biomarqueurs;
2) Valider les performances des biomarqueurs neurophysiologiques précédemment décrits et explorer de nouvelles caractéristiques du signal EEG associées à l'épilepsie;
3) Concevoir et optimiser un modèle d'apprentissage profond interprétable pour la détection de l'épilepsie et la prédiction du risque de crise à partir de l'EEG de routine.

3.1 Organisation de la thèse

La thèse est divisée en deux parties principales. La première partie (Chapitres 1 et 2) présentait une revue critique de la littérature, démontrant le besoin de biomarqueurs en épilepsie et le potentiel de l'apprentissage profond appliqué à l'EEG, tout en exposant les limitations méthodologiques actuelles.

La deuxième partie présente trois études complémentaires qui explorent progressivement la modélisation du risque de crise, des approches traditionnelles aux méthodes d'apprentissage profond. Le Chapitre 4 établit les fondations en évaluant un modèle basé sur des caractéristiques computationnelles classiques pour le diagnostic d'épilepsie et la prédiction de récurrence de crise à un an. Le Chapitre 5 introduit une nouvelle approche en développant une architecture d'apprentissage profond capable de modéliser directement le signal EEG brut pour prédire le diagnostic d'épilepsie. Le Chapitre 6 étend cette méthodologie en proposant un modèle qui prédit le risque de crise à travers le temps tout en améliorant l'interprétabilité clinique.

CHAPITRE 4 ARTICLE 1: MACHINE-LEARNING FOR THE PREDICTION OF ONE-YEAR SEIZURE RECURRENCE BASED ON ROUTINE ELECTROENCEPHALOGRAPHY

Emile Lemoine, MD, MSc[1,2,3], Denahin Toffa, MD, PhD[1,3], Geneviève Pelletier-Mc Duff, BSc[1,3], An Qi Xu[3], Mezen Jemel, BSc[1,3], Jean-Daniel Tessier, BSc[1,3], Frédéric Lesage, PhD[2,4], Dang K. Nguyen, MD, PhD[1,3], Elie Bou Assi, MEng, PhD[1,3*]

[1]Department of Neurosciences, Université de Montréal, Montréal, Qc, Canada

[2]Institute of biomedical engineering, École Polytechnique de Montréal, Montréal, Qc, Canada

[3]Centre de Recherche du CHUM (CRCHUM), Montréal, Qc, Canada

[4]Centre de Recherche de l'institut de Cardiologie de Montréal, Montréal, Qc, Canada

*Corresponding author

Cet article aborde les premier et deuxième objectifs visant à mettre sur pied une base de données d'EEG de routine avec information clinique et de mesurer la performance diagnostique pour l'épilepsie des marqueurs computationnels existants. L'article a été publié le 4 août 2023 dans la revue *Scientific Reports* (https://www.nature.com/articles/s41598-023-39799-8) et compte maintenant 18 citations (*Google Scholar*). Il a remporté le prix Mary Ann Lee de la Ligue Canadienne contre l'Épilepsie pour meilleure publication par un résident en neurologie au Canada (2024) et le prix Pavel Hamet pour meilleure publication au CRCHUM (2024). Ce travail a fait l'objet de multiples présentations orales, notamment à l'*American Academy of Neurology Annual Meeting* (Boston, 2023; Prix *Futures in Neurological Research*) et au *Canadian Neurological Sciences Federation Congress* (Montréal, 2022; Prix Herbert Jasper pour meilleur abrégé en neurophysiologie clinique), ainsi que des présentations par affiches, incluant à *ICTALS* (Berne, 2022; Bourse de congrès), l'*American Epilepsy Society Annual Meeting* (Virtual [Chicago], 2021) et l'*International Epilepsy Conference* (Virtual, 2021).

Ma contribution à cet article comprend l'identification de la problématique, la collecte de données, le développement de la méthode d'apprentissage automatique, le prétraitement des données EEG, la réalisation des expériences computationnelles, l'interprétation et l'analyse des résultats, la conception des visualisations, ainsi que la rédaction du manuscrit.

4.1 Abstract

Predicting seizure recurrence risk is critical to the diagnosis and management of epilepsy. Routine electroencephalography (EEG) is a cornerstone of the estimation of seizure recurrence risk. However, EEG interpretation relies on the visual identification of interictal epileptiform discharges (IEDs) by neurologists, with limited sensitivity. Automated processing of EEG could increase its diagnostic yield and accessibility. The main objective was to develop a prediction model based on automated EEG processing to predict one-year seizure recurrence in patients undergoing routine EEG. We retrospectively selected a consecutive cohort of 517 patients undergoing routine EEG at our institution (training set) and a separate, temporally shifted cohort of 261 patients (testing set). We developed an automated processing pipeline to extract linear and non-linear features from the EEGs. We trained machine learning algorithms on multichannel EEG segments to predict one-year seizure recurrence. We evaluated the impact of IEDs and clinical confounders on performances and validated the performances on the testing set. The receiver operating characteristic area-under-the-curve for seizure recurrence after EEG in the testing set was 0.63 (95%CI: 0.55–0.71). Predictions were still significantly above chance in EEGs with no IEDs. Our findings suggest that there are changes other than IEDs in the EEG signal embodying seizure propensity.

4.2 Introduction

Epilepsy is a chronic neurological condition defined as an enduring, pathological propensity to recurring seizures [2]. Predicting the risk of seizure recurrence is at the heart of the diagnosis and management of people with epilepsy (PWE). The electroencephalogram (EEG), a 20- to 60-minute recording of the electrical activity of the cerebral cortex via scalp electrodes, is a cornerstone of the estimation of seizure recurrence risk. The hallmark of epilepsy on the EEG is the interictal epileptiform discharge (IED): a brief, sharp discharge emanating from the background rhythm between seizures. In several clinical scenarios, such as after a first unprovoked seizure, before withdrawing antiseizure medication (ASM), and after surgical resection of an epileptic focus, visual identification of IEDs on routine EEG grossly doubles the risk that a patient will have seizure recurrence in the next years [31]. This impacts ASM management and prescription of ancillary tests.

Unfortunately, spikes are elusive: in PWE, a 20-minute EEG captures spike in only 29–55% of cases [10], [28]. As a result, the sensitivity of EEG for predicting seizure recurrence is limited [30].

In addition, IEDs are subject to overinterpretation with an inter-rater agreement that is only moderate, even among fellowship-trained neurophysiologists [37]. The overidentification of sharply contoured waveforms and normal variants as epileptiform can lead to the misdiagnosis of epilepsy, particularly in the event of a poor clinical history [139], [140]. Finally, once the diagnosis of epilepsy is established, IEDs on routine EEG do not correlate well with disease activity, limiting their usefulness to monitor ASM therapy [141], [142], [143], [144]. A biomarker of seizure propensity that is automated, objective, and independent of IEDs would heavily impact clinical practice by reducing diagnostic error, accelerating treatment in patients at high risk of seizures, avoiding the consequences of overdiagnosis in the others, and monitor disease activity.

Several studies have suggested that the routine EEG can capture non-visible anomalies in cortical activity in patients with both focal and generalized epilepsies [145], [146], [147], [148], [149], [150]. These differences include subtle power shifts in specific frequency bands [151], [152], [153], [154], changes in regularity of the signal [80], [155], or presence of power scaling laws [79]. However, key questions were not addressed in previous studies, such as the reproducibility on external data and the impact of confounders like age and antiseizure therapy. In addition, previous studies are underpowered (with samples smaller than 100 patients) [156]. Thus, the potential predictive performances of computational EEG biomarkers in the clinical setting remain unknown [157]. There is a need for high-powered cohort studies to assess the diagnostic accuracy of these biomarkers and initiate their clinical translation.

In this paper, we develop and validate predictive models for the prediction of seizure recurrence at one year based on the computational extraction of biomarkers from the routine EEG signal. We train the model on a large retrospective cohort of consecutive patients undergoing routine EEG and validate the predictive accuracy on a temporally shifted cohort of patients. We investigate whether predictive accuracy is independent of IEDs and other clinical confounders.

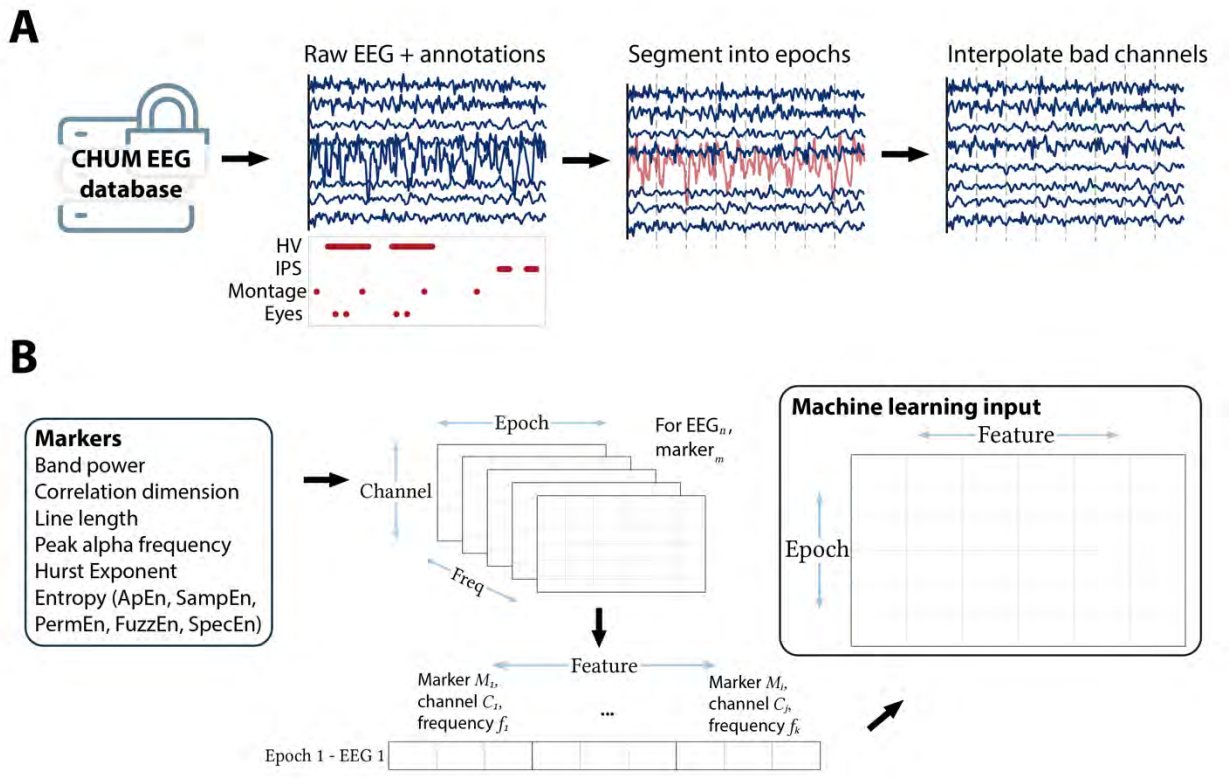


Figure 4.1 : EEG processing and marker extraction methods. A: Processing of a single EEG: extraction from the database in which the EEG is stored with annotations, segmentation into 10s epochs according to pre-defined timepoints, identification of artefactual channels (in red) and interpolation. B: Marker extraction: for each marker, one value is computed at each channel, epoch, and frequency bands. The values for a given epoch are used as input for the machine learning algorithm.

4.3 Methods

4.3.1 Patient population and clinical file review

We retrospectively recruited all consecutive patients who underwent a routine EEG at the University of Montreal Hospital Center (CHUM) between January 2018 and June 2019. Routine EEGs (both awake and sleep recordings) recorded between January and December 2018 constituted the training set, while EEGs recorded between January and June 2019 constituted the held out testing set. We excluded EEGs with no follow-up visit available after the EEG, uncertain diagnosis at the end of the available follow-up, or with excessive artifacts or seizures (as per the EEG report). For the testing set, we additionally excluded patients for whom an EEG was already included in the training set. We reviewed the patients' entire medical chart for clinical information:

demographics (age, sex), comorbidities at the time of the EEG, epileptogenic factors, reason for EEG, presence of focal brain lesion on neuroimaging (when available), and number of ASMs. For PWE, we collected type of epilepsy, age of epilepsy onset, and date of the first seizure after the EEG. If the date of first seizure after the EEG was not available, we estimated it by linear interpolation based on the seizure frequency reported in the visit after the EEG. From the EEG report, we extracted the type of recording (awake or sleep deprived), deepest sleep stage achieved, presence of IED, and presence and degree of abnormal slowing. All clinical information was stored on a REDCap database hosted on the CHUM research center's servers.

4.3.2 Outcomes

The primary outcome is the patient-reported seizure recurrence during the first year of follow-up after the EEG, as provided in the medical notes. We considered only unprovoked seizures and auras, which include seizures that occurred in the setting of sleep deprivation and medication non-compliance [2]. The secondary outcome was the diagnosis of epilepsy, based on information available in medical notes by the appointed neurologist. The starting date of the diagnosis would be the date of the first seizure experienced by the patient. We only considered the diagnosis valid if it was concordant with International League Against Epilepsy criteria (Fisher et al., 2017): having had at least one seizure and either 1) a second seizure >24 hours apart or 2) an estimated risk of seizure recurrence ≥ 60% over the next 10 years [2]. If the diagnostic was not concordant, the patient would be excluded. The last outcome was active epilepsy: this required a diagnosis of epilepsy, at least one seizure in the year preceding the EEG, and seizure recurrence at any point after the EEG.

4.3.3 EEG recording and processing

EEGs were recorded using a Nihon Kohden EEG system. The recording protocol is standardized based on national recommendations [158]. Awake EEGs were 20–30 minutes in duration and are recorded at 200 Hz through 19 electrodes arranged with the standard 10-20 distribution. They included two 90s periods of hyperventilation and photic stimulation from 4 Hz to 22 Hz. Hyperventilation was not performed in patients >80 y.o., in patients unable to cooperate with technologists, nor in patients with medical contraindications. Moreover, the patients were instructed to open or close their eyes at several times during the recording. Sleep deprived recordings were 60 minutes in duration, with the same activation procedures. An EEG technologist

annotates the EEG in real-time. The protocol includes changes of montage at regular intervals during the recording, rotating through a total of seven montages. For this study, we converted the digital EEG to an average referential montage. EEGs were converted into EDF format and stored on the CHUM research center's server for analysis.

The processing pipeline is illustrated in Figure 4.1A. EEG recordings were high pass filtered at 0.75 Hz and notch filtered at 60 Hz with a fast-impulse response (FIR) filter (hamming window). Ten-second epochs were extracted at pre-specified time points: every change of montage, every 15s during hyperventilation, every 15s for 2 minutes post-hyperventilation, every photic stimulation frequency, and every eye closure or opening. Artifact detection and interpolation were done using the *autoreject* algorithm [159]. For a given EEG recording, an optimal peak-to-peak amplitude threshold was found for each individual epoch/channel combination using 5-fold cross-validation (CV). Rejected channels were interpolated using spherical splines. The preprocessing pipeline was written in Python (version 3.8) and is based on the *MNE* library.

4.3.4 Extraction of computational markers

Ten univariate markers were extracted from the EEGs. The markers were selected based on previous literature, with the aim of capturing distinct linear and non-linear properties of the EEG signal across the time- and spectral-domain at each channel. The markers' algorithms, mathematical details, and references are supplied in the **Supplementary method 1**. Linear features were: band power (BP) in ten frequency bands (low [1–2 Hz] and high [2–4 Hz] delta, low [4–6 Hz] and high [6–8 Hz] theta, low [8–10 Hz] and high [10–13 Hz] alpha, low [13–20 Hz] and high [20–40 Hz] beta, low [40–75 Hz] and high [75–100 Hz] gamma), peak alpha frequency (PAF), and Hurst exponent (HE). For BP, the EEG were band pass filtered using a FIR (hamming window). Power spectrum density was calculated using a multitaper method, and the integral in each frequency band was estimated using Simpson's method. For the PAF, the peak frequency of the alpha band (8–13 Hz, band pass filtered using a FIR window) was extracted for each EEG, epoch, and channel. HE was calculated for each EEG, epoch, channel, and wavelet decomposition level (see next paragraph), with a minimum window size of 10 points.

Non-linear features were: line length (LL), correlation dimension (CD), and five different entropy estimates: Approximate (ApEn), Sample (SampEn), Fuzzy (FuzzEn), Permutation (PermEn), and Spectral entropy (SpecEn). For non-linear features and for HE, one value was calculated for each

feature, EEG, epoch, channel, and wavelet decomposition level (Figure 4.1B**)**. The *Sym5* wavelet was used with six decomposition levels (with frequency range: 100–50 Hz, 50–25 Hz, 25–12.5 Hz, 12.5–6.25 Hz, 6.25–3.125 Hz, and 3.125–1.56 Hz) [160]. For entropies, optimal parameters were selected to maximize the inter-EEG vs. intra-EEG variance on five EEGs that were excluded from the study ($m = 3$, $r = 0.25$, $\tau = 5$, $n = 2$, and $k = 3$). Missing values were imputed using multivariate iterative imputation. The calculation of the markers was independent on the outcomes.

4.3.5 Machine learning model development and validation

The ML model's task is to map the vector of linear and non-linear features' values for a single EEG to a clinical outcome. The training was done epoch-wise—each EEG epoch fed as an independent learning observation. To prevent data leakage, epochs from the same patient were grouped together in the same CV split. The predictions for epochs of a single EEG were aggregated using the median to yield a single prediction per EEG. We also tested other percentiles (0.1–0.9 in 0.1 steps) for the aggregation of predictions (**Supplementary Table S1**). The clinical outcomes are described in the **Outcomes** section. The performance metric is the receiver operating characteristic area-under-the-curve (ROC AUC), selected for its robustness to class imbalance. Improvement over chance (IoC) was defined as an AUC significantly over 0.50.

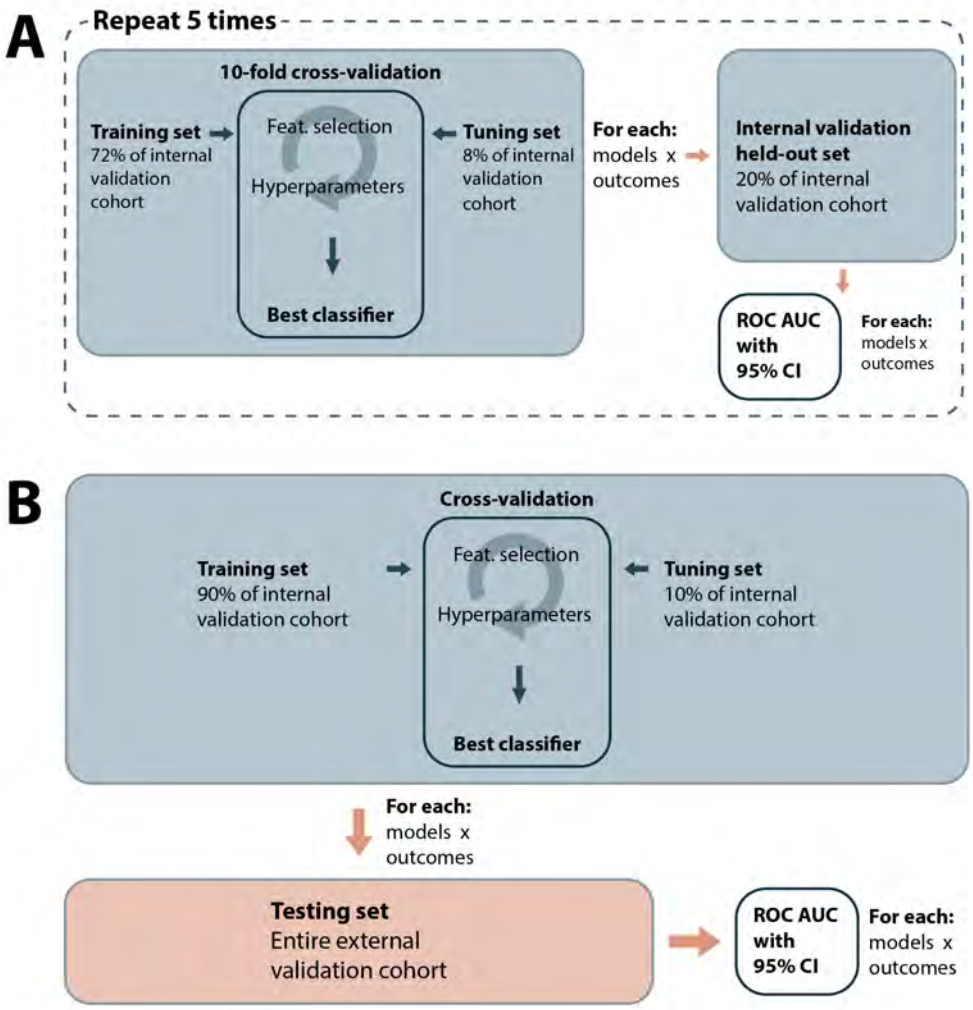


Figure 4.2 : Machine learning evaluation methods. A: Nested cross-validation with 10-fold inner loop and 5-fold outer loop. B: Evaluation on the temporally shifted validation cohort (testing set). In blue: internal validation cohort (EEGs recorded in 2018). In red: testing cohort (EEGs recorded in 2019).

Four ML algorithms were evaluated: Generalized linear model (GLM; L_1- and L_2-regularized logistic regression) [161], support vector machine with radial basis function (SVM), random forest (RF), and gradient boosted trees (LightGBM) [162]. Supervised feature selection was performed with a linear L_1-regularized SVM.

A nested CV was used first to evaluate the models, features, and clinical confounders on the training set (Figure 4.2A). In nested CV, an inner-loop is used to optimize hyperparameters of the feature selection and learning model, and an outer-loop is used to validate the performances on separate data. It allows to estimate confidence intervals and is more robust than CV [163]. We used

a 10-fold CV for the inner-loop and a 5-fold CV for the outer-loop, with patient-wise grouping of the EEG epochs at each level. ROC AUC values across outer-loop CV splits were averaged and 95% confidence intervals were estimated using LeDell's curve based approach [164]. ROC AUC values were compared against the random classifier (ROC AUC = 0.50). Statistics were calculated at the EEG level (after aggregating predictions for all epochs in a single EEG). ML models were trained and validated using Python 3.8 (with classifiers from *Scikit-learn* and *LightGBM* libraries).

We tested the interacting effect of age, presence of IEDs, and presence of a focal lesion on neuroimaging to increase the performances of the algorithm. For age, we added interaction terms between scaled age and features to the set of features. For IEDs and focal lesions, we used a two-step classifier: first, if the factor is positive (e.g.: presence of IEDs), the model automatically outputs a positive prediction. If the factor is negative, the ML model's predictions are used.

Validation of predictive performances on testing set

We validated the performances of the best performing ML models on the testing cohort (Figure 4.2B). First, we removed features that did not show IoC on the training cohort. Then, we performed a 10-fold CV on the training data (EEGs from 2018) to select the best features and best hyperparameters for each of the four previously described models and three outcomes. The best feature-set/hyperparameters were used to train the models on the training data. The trained models were then applied to the testing set (EEGs from 2019) to emit probabilistic predictions. We computed the ROC AUC values from the probabilistic predictions, with 95% confidence intervals estimated by DeLong's approach (single prediction by patient) [165]. For the primary outcome, we calculated the performance using only patients who had at least a one-year follow-up after the EEG. We also tested the outcomes on all testing patients (including those with follow-up shorter than one year). For comparison, we tested the classification performance of IED alone (presence vs. absence) and focal lesion alone (presence vs. absence) on the risk of seizure recurrence.

4.3.6 Post-hoc analyses

Post-hoc analyses were performed on predictions from the LightGBM classifier. For each outcome, we evaluated the risk of bias of the classification for different subgroups by recomputing average AUC and 95% CI. The subgroups were age group (18–40, 40–60, and >60 years old), sex, presence of focal lesion, presence of IED (absence, presence, and uncertain), presence of slowing, number of ASM (0, 1, ≥2), and epilepsy type (focal, generalized, and unknown). We investigated the

performances in two specific subgroups: patient not yet diagnosed with epilepsy (undergoing evaluation for suspected seizures), and patients undergoing EEG pre-ASM withdrawal.

We also investigated the time-dependence of the predictions for seizure recurrence as well as the impact of clinical confounders using a multivariate survival model. We used the model's predictions to separate the patients into a low-predicted risk and high-predicted risk (above vs. below average). We then fit a cox proportional hazard model to estimate the hazard ratio of seizure recurrence dependent on the model's predictions, controlled for the following characteristics: age, sex, and number of ASMs (selected based on a directed acyclic graph presented in **Supplementary Figure S1**). We checked the robustness of the choice of covariates with a sensitivity analysis (see **Supplementary method 2**).

Comparison of individual markers

We compared the predictions for seizure recurrence of individual markers between each other and with their combination. We repeated the nested CV independently for each marker, using only the values from this marker as input to the classification pipeline (keeping CV splits identical between markers).

4.3.7 Sample-size estimation

Power analysis is described in **Supplementary method 3**. With a significance level of 0.05, accounting for 12 multiple comparisons (3 outcomes x 4 models), we estimated that routine EEGs from a single year would provide us with a power > 0.9 for the expected effect size.

4.3.8 Ethics

Ethics approval was provided by the CHUM Research Centre's Research Ethics Board (Montreal, Canada, project number: 19.334). A waiver of informed consent was provided by the CHUM Research Centre's Research Ethics Board due to the absence of diagnostic/therapeutic intervention and the absence of risk for the subjects involved. All methods were carried out in accordance with Canada's Tri-Council Policy statement on Ethical Conduct for Research Involving Humans.

4.3.9 Reporting standards

The reporting of the study conforms with the TRIPOD statement (Transparent reporting of multivariate prediction model for individual prognosis or diagnosis) when applicable.[166]

4.4 Results

4.4.1 Patient characteristics

Patients' characteristics for the training cohort are described in **Table 4.1**. We screened 816 records for eligibility; 517 patients were included (549 EEG recordings). In this cohort, 132 EEGs were from patients who had seizure recurrence after the EEG (24%). There were 346 EEGs (63%) from PWE. The median age was 50 y.o. (IQR: 33–62 y.o.). Median follow-up after the EEG was 100 weeks (IQR: 42–135). In PWE, 248 EEGs (72%) did not show IEDs. The EEG was part of the initial evaluation of suspected seizure(s) in 286 cases (Supplementary Table S1).

Table 4.1 : Description of the training (EEG recordings between January and December 2018) and testing cohorts (EEG recordings between January and June 2019)

	Training cohort (EEGs from 2018)		**Testing cohort (EEGs from 2019)**	
	Seizure freedom at one year	**Seizure recurrence at one year**	**Seizure freedom at one year**	**Seizure recurrence at one year**
Number of EEGs	417	132	217	84
Epilepsy diagnosis (%)	214 (51.3)	132 (100.0)	98 (45.2)	84 (100.0)
Age (median [IQR])	52.00 [35.00, 64.00]	37.00 [26.75, 55.00]	54.00 [36.00, 66.00]	37.00 [30.00, 57.25]
Sex = woman (%)	223 (53.5)	56 (42.4)	115 (53.0)	51 (60.7)
Total follow-up after eeg in weeks (median [IQR])	94.00 [35.00, 133.00]	136.00 [97.50, 163.25]	78.00 [32.00, 122.00]	123.50 [94.25, 141.00]
Epilepsy type (%)				
Focal	148 (35.5)	88 (66.7)	71 (32.7)	66 (78.6)
Generalized	56 (13.4)	41 (31.1)	21 (9.7)	13 (15.5)
No epilepsy	203 (48.7)	0 (0.0)	119 (54.8)	0 (0.0)
Unknown	10 (2.4)	3 (2.3)	6 (2.8)	5 (6.0)
Age of epilepsy onset (median [IQR])	24.00 [15.00, 43.00]	18.00 [13.00, 35.00]	28.00 [15.00, 53.00]	28.00 [14.00, 48.00]
Number of days since last seizure (median [IQR])	801.00 [225.00, 3525.00]	43.00 [9.50, 127.00]	613.00 [165.00, 1489.50]	32.50 [4.00, 90.00]
Number of ASM (%)				
0	197 (47.2)	11 (8.3)	108 (49.8)	19 (22.6)
1	149 (35.7)	58 (43.9)	80 (36.9)	31 (36.9)
2	48 (11.5)	45 (34.1)	21 (9.7)	15 (17.9)
3	19 (4.6)	13 (9.8)	6 (2.8)	14 (16.7)
4	4 (1.0)	5 (3.8)	2 (0.9)	3 (3.6)
5	0 (0.0)	0 (0.0)	0 (0.0)	2 (2.4)
Focal lesion on brain imaging (%)	142 (34.1)	53 (40.2)	76 (35.0)	44 (52.4)
Sleep deprived EEG (%)	51 (12.2)	13 (9.8)	33 (15.2)	18 (21.4)
IED (%)				
Absence	333 (79.9)	77 (58.3)	174 (80.2)	47 (56.0)
Presence	52 (12.5)	46 (34.8)	28 (12.9)	30 (35.7)
Uncertain	32 (7.7)	9 (6.8)	15 (6.9)	7 (8.3)
Abnormal slowing on EEG (%)	107 (25.7)	37 (28.0)	72 (33.2)	39 (46.4)

For the testing set, 429 records were screened for eligibility. After applying exclusion criteria, we included 301 EEGs from 261 patients (Table 4.1). The prevalence of seizure recurrence after EEG in this cohort was 32%. Other variables have a similar distribution to the training and validation cohort.

4.4.2 Predictive performances on the internal validation cohort

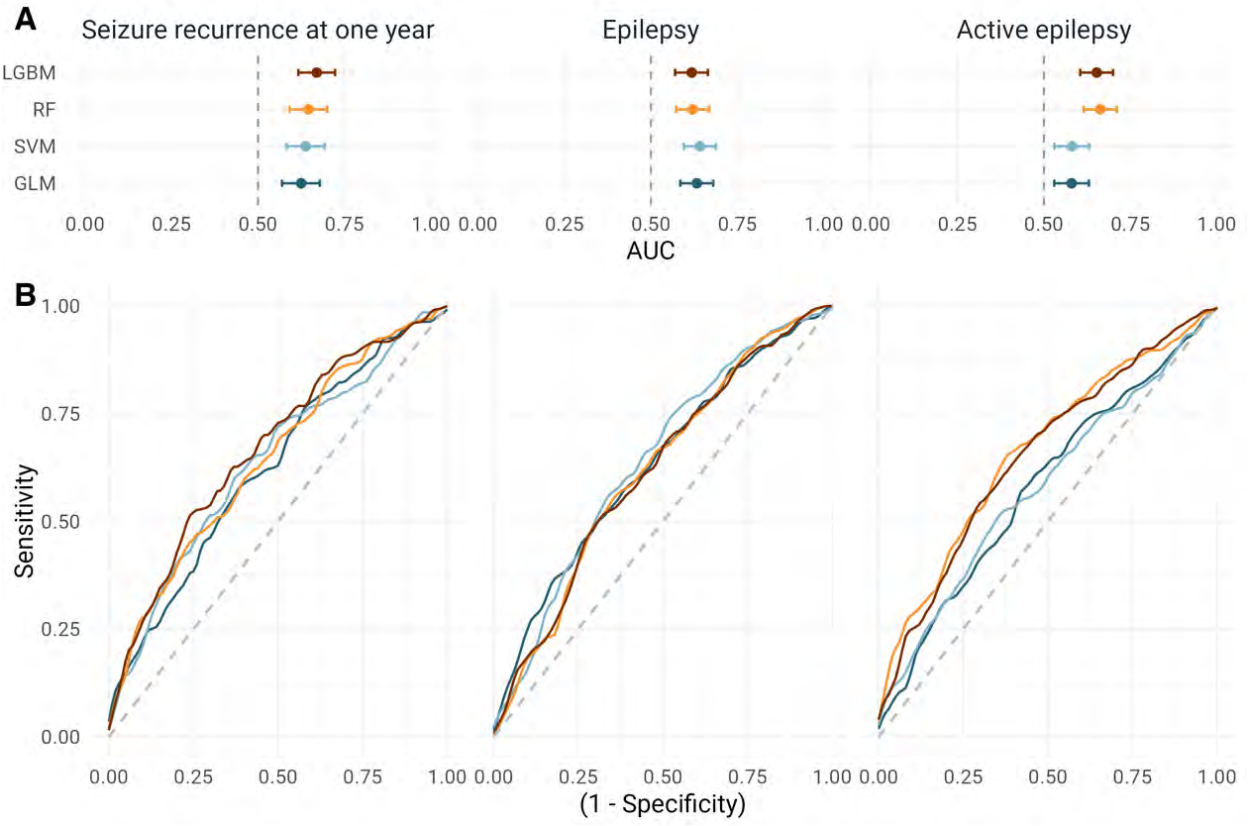


Figure 4.3 : Classification performances for each classification algorithm and each clinical outcome. A: AUC ROC with 95% interval estimated using nested five-fold cross-validation. B: ROC curves for each algorithm

For all outcomes, all four algorithms had statistically significant IoC (Figure 4.3 and Table 4.2). For the prediction of seizure recurrence at one year, the best model was LightGBM with a ROC AUC of 0.67 (0.62–0.72). For the diagnosis of epilepsy, the best model was SVM with a ROC AUC of 0.64 (0.60–0.69). For active epilepsy, the best model was RandomForest with a ROC AUC of 0.66 (0.62–0.71). There was no statistical difference in performances across models for each outcome. The quantile used for aggregating predictions of the epochs from a single EEG had no significant impact (**Supplementary Table S2**).

Table 4.2 : Classification performances on internal validation cohort estimated from nested CV

Outcome	Classifier	AUC
Seizure recurrence at one year	GLM	0.62 (0.57–0.68)
	SVM	0.64 (0.58–0.69)
	RandomForest	0.65 (0.59–0.70)
	LightGBM	0.67 (0.62–0.72)
Epilepsy diagnosis	GLM	0.63 (0.58–0.68)
	SVM	0.64 (0.60–0.69)
	RandomForest	0.62 (0.57–0.67)
	LightGBM	0.62 (0.57–0.66)
Active epilepsy	GLM	0.58 (0.53–0.63)
	SVM	0.58 (0.53–0.63)
	RandomForest	0.66 (0.62–0.71)
	LightGBM	0.65 (0.60–0.70)

Adding clinical information to the feature set did not have a statistically significant effect. By integrating age, AUC was 0.67 (0.62–0.72). For the two-step model with IEDs, AUC was 0.66 (0.60–0.71) and for the two-step model with focal lesion, AUC was 0.58 (0.53–0.64).

4.4.3 Survival analysis

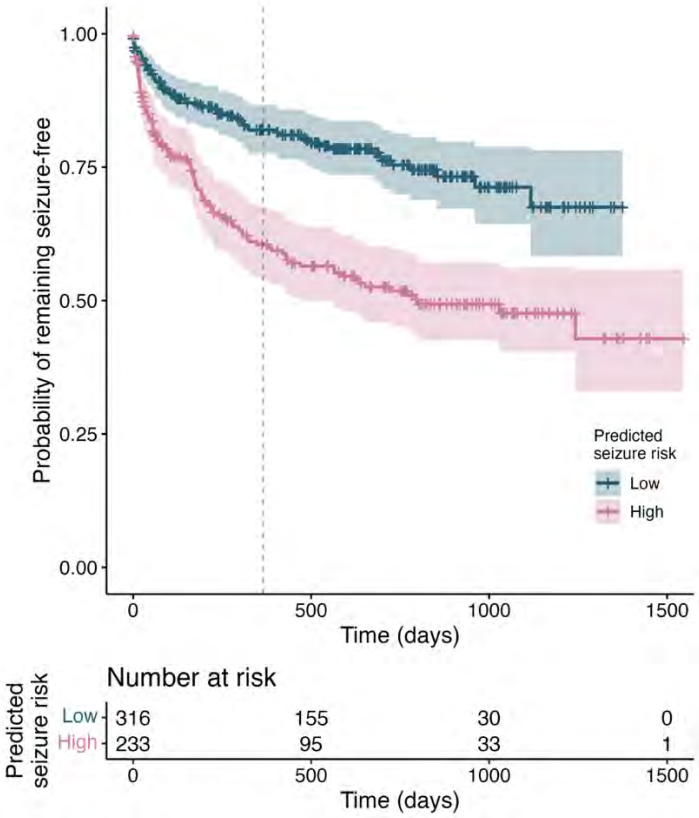


Figure 4.4 : Survival curve for seizure freedom after the EEG, dependent on the risk predicted by the LightGBM model. Dashed line indicates one-year follow-up.

The overall probability of remaining seizure free at one-year was 0.73 (95%CI: 0.69–0.77). When stratifying by the LightGBM model predictions, the one-year seizure free survival was 0.82 (0.78–0.87) in the high-predicted risk, and 0.61 (0.54–0.68) in the low predicted risk. In contrast, the one-year seizure free survival as a function of IEDs was 0.49 (0.40–0.61) in the presence of IEDs, and 0.78 (0.74–0.82) in the absence of IEDs.

In the multivariate survival analysis, the adjusted hazard ratio of seizure recurrence for the model's predictions was 1.22 (95%CI 1.07–1.40, $p = 0.0029$). The Kaplan-Meier curve (Figure 4.4) shows separation between groups up to one year after the EEG. The risk of seizure recurrence was strongly associated with age (aHR: 0.68, 0.56–0.82, $p < 0.001$) and number of ASM (1.66, 1.43–1.92, $p<0.001$). The sensitivity analysis showed robustness to different sets of covariates (**Supplementary method 2**).

4.4.4 Subgroup analyses

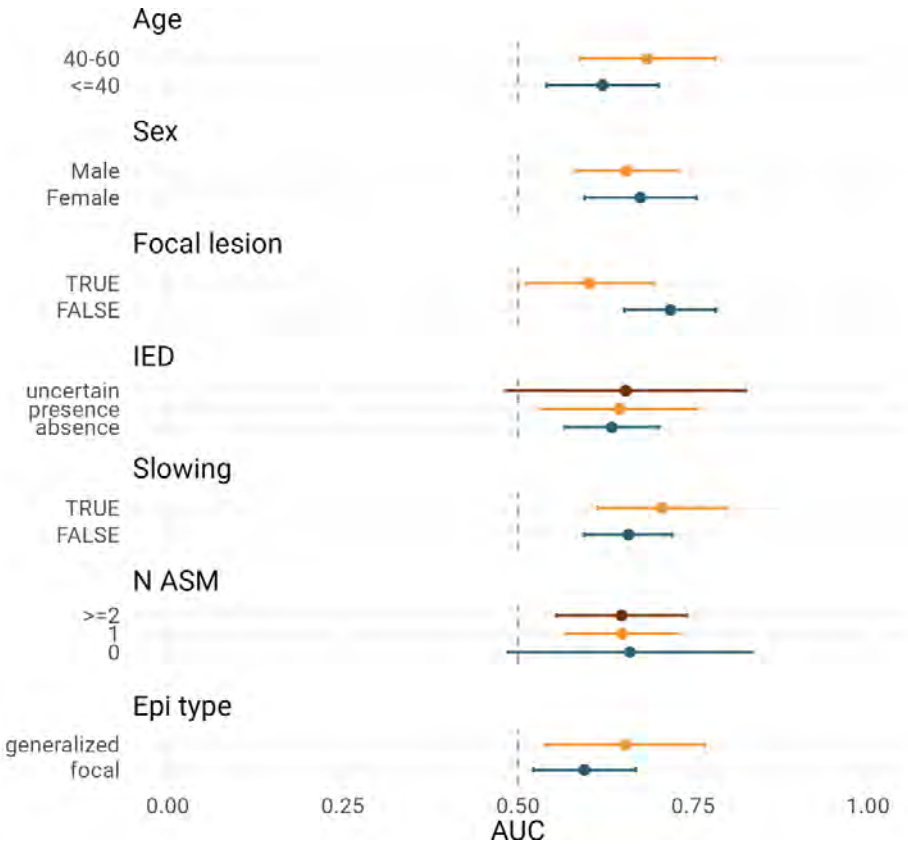


Figure 4.5 : Predictive performances (ROC AUC) for the LightGBM model stratified by subgroups for each of the three outcomes, with 95% confidence intervals. The dotted line indicates AUC of 0.50. AUC ROC: Area-under-the-receiver operating characteristic curve.

In the subgroup analysis, there was no statistical differences between any strata for almost all outcomes (Figure 4.5 and **Supplementary Figure S2**). For seizure recurrence at one year almost all subgroups had performances that were significantly above chance. Only the absence vs. presence of focal lesion showed a trend towards increased AUC. For the epilepsy outcome, performances were not above chance for patients between 40 and 60 y.o. For some subgroups, sample size was small, and estimation were either not reliable ("uncertain" IEDs [all outcomes], no ASM [outcome "seizure recurrence"]) or impossible (patients > 60 y.o. [outcome "seizure recurrence"], ≥2 ASM [outcome "Epilepsy"]).

In the subgroup of patients undergoing initial evaluation for seizure(s) (N = 227), the ROC AUC was also statistically significantly above chance (0.65 [0.57–0.74]). For the subgroup of patients

undergoing EEG before ASM withdrawal, AUC could not be estimated because of the small sample size (N=32, < 1 per outcome in each nested CV fold).

4.4.5 Comparison between markers

The comparison between markers is shown in Figure 4.6. The best markers were BP (AUC ROC 0.65 [95%CI: 0.59–0.70]), LL (0.63 [0.58–0.68]), and FuzzEn (0.61 [0.56–0.67]). PermEn, PAF, and SpecEn did not show IoC. The combination of all features had the greatest predictive performances (0.65 [0.60–0.70]).

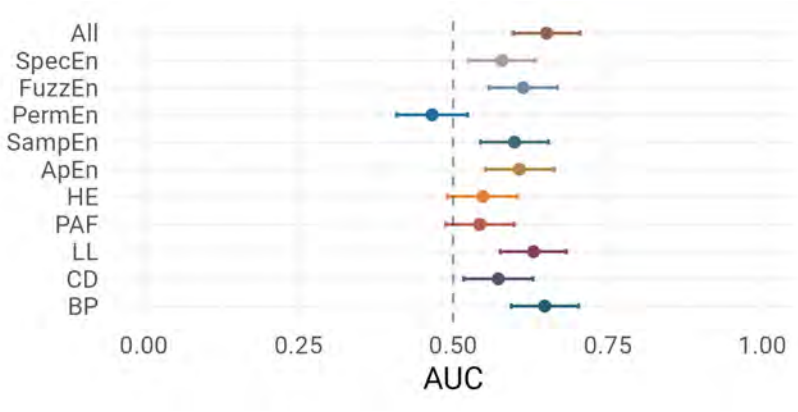


Figure 4.6 : Comparison of predictive performances for all markers using a LightGBM model. ApEn: Approximate entropy; BP: Band power; CD: Correlation dimension; FuzzEn: Fuzzy entropy; HE: Hurst exponent; LL: Line length; PAF: Peak alpha frequency; PermEn: Permutation entropy; SampEn: Sample entropy; SpecEn: Spectral entropy.

We repeated the subgroup analysis for each individual feature (**Supplementary Figure S3**). For age, relative performances of CD and SpecEn were decreased in patients 40–60 y.o., while for BP, they were decreased in patients ≤ 40 y.o. Presence of focal lesion decreased performances in all markers. Presence of IEDs decreased performance of all markers except BP and LL. The presence of abnormal slowing particularly reduced the performances of CD. In patients with one ASM, LL and BP had higher performances, while in patients with ≥2 ASM, only BP had IoC. The combination of markers reduced the impact of stratification, and was the only feature set that showed IoC in all strata.

4.4.6 Validation on a temporally shifted cohort

In the testing cohort, the LightGBM model had IoC for predicting seizure recurrence at one year, with an AUC of 0.63 (95%CI: 0.55–0.71). The performances on the entire testing cohort (including

those with follow-up shorter than a year) were similar (0.64 [0.58–0.71]). The binary predictions had a negative predictive value (NPV) of 78%, a positive predictive value (PPV) of 37%, a sensitivity of 64%, and a specificity of 55%. In the absence of IEDs, the LightGBM predictions were still significantly above chance (seizure recurrence: 0.63 [0.55-0.71]). For comparison, in this cohort, IEDs (presence vs. absence/uncertain) had an AUC of 0.61 (0.56–0.69) for seizure recurrence at one year, while the presence of a focal lesion had an AUC of 0.59 (0.53–0.65). For the outcomes "epilepsy diagnosis" and "active epilepsy", AUC for the LightGBM model was 0.64 (0.57–0.70) and 0.57 (0.50–0.63), respectively.

We tested the two-step classifier with IEDs on the temporally shifted cohort for the prediction of seizure recurrence at one year, achieving an AUC was 0.70 (0.63–0.76). For the binary predictions, NPV was 80%, PPV was 51%, sensitivity was 83%, and specificity was 45%.

4.5 Discussion

This study demonstrates that machine learning models trained on computational features automatically extracted from 20-minute scalp EEG can predict seizure recurrence at one year with above-chance performances in a cohort of 778 consecutive patients undergoing routine EEG. The predictive performances for estimating seizure recurrence risk after routine EEG were validated in a temporally shifted cohort of 261 patients, where ROC AUC was 0.63, significantly above chance. In comparison, the presence of IEDs—the only validated marker of seizure risk in the clinical setting—was 0.61. A two-step model that uses first IEDs and then computational features on IED-negative EEGs achieved a testing AUC of 0.70. These performances were still significant with EEGs that did not capture any visible IEDs. The best performances were achieved in patients without focal lesion.

The most important finding of this study is the robust evidence for non-visible changes in the EEG signal associated with the propensity to have seizures. For decades, researchers have been hinting at non-visible differences in the EEG signal of PWE compared to healthy controls [145], [167]. Two important frameworks to model the EEG are linear and non-linear models. Linear models assume that the signal arises from a linear combination of independent oscillators. In general, alpha frequency is found to be slower in patients with focal [154], [168] and generalized epilepsy [151]. Non-linear models represent the signal as a non-linear dynamical system that can be characterized by entropy and dimensionality, among others. Entropy and correlation dimension, both measures

of signal complexity, tend to be reduced in PWE [80], [155], [169]. It must be emphasized that this body of literature is built upon small case-control studies [145], a study design that overestimates diagnostic performances [170].

In contrast to case-control designs, cohort or nested case-control studies reduce the risk of selection bias when evaluating diagnostic accuracy [171]. Two studies used this more robust approach to predict seizure recurrence from automated analysis of routine EEG [69], [77]. In the first, the authors evaluated Paroxysmal slow wave events (PSWE, 2-second EEG windows with a median peak frequency of <6 Hz) on a cohort of 70 patients undergoing EEG after a first seizure [77]. They found that the rate of PSWE could predict seizure recurrence at 18-month with an AUC of 0.72. In the second, on a cohort of 114 patients undergoing EEG after a first suspected seizure, the connectivity in the theta band could predict a future diagnosis of epilepsy with a specificity of 70% and sensitivity of 53% [69]. Importantly, neither study validated their findings on a separate set of patients. In our study, we adopted a cohort design: subjects were drawn consecutively from a population of patients undergoing routine EEG, i.e., the target population in a real-world setting. We also used a temporally shifted testing cohort, which allows to explore the out-of-sample generalizability of the models in a manner that mimics their real-life deployment. These factors reinforce the robustness of the performance estimation, which should be consistent when deployed in the clinical setting [172]. However, the testing cohort was from the same institution as the training set. The capacity of this method to generalize to other institutions would need to be evaluated in a future study.

While the correlation between seizure frequency and presence of IEDs on routine EEG is not well established, the ability to predict long-term seizure recurrence from routine EEG would ~~greatly~~ impact both the management of patients presenting with suspected seizure(s) and patients diagnosed and treated for epilepsy. Currently, the prognostic value of EEG at diagnosis is mostly focused on the evaluation of patients with an unequivocal, single unprovoked seizure. In these patients, identification of IEDs on a single routine EEG confers a two-fold increase in the risk of subsequent seizures if untreated, generally warranting ASM therapy [173]. In addition, the prognostic value of EEG before ASM withdrawal is demonstrated in patients with at least two-year seizure freedom (especially in children) [174]. Beyond these clinical scenarios, there is still little evidence to support the use of EEG to adjust ASM therapy and prognosticate the disease since a highly active EEG does not necessarily correlate with seizure frequency; this restricts the

usefulness of EEG as a monitoring tool [31], [141], [142], [175]. In this study, we included all consecutive patients undergoing routine EEG at our institution; we are interested in the potential of the routine EEG to quantify at one point in time the propensity to have future seizures. When combined with IEDs in a two-step model, the algorithm showed a testing ROC AUC of 0.70. For context, in our cohort, IEDs alone had an AUC of 0.61. These results demonstrate a certain complementarity between EEG features and IEDs, and bring hope that the routine EEG could potentially be used as a tool to assist clinicians in recommending for or against ASM treatments based on future seizure risk. However, the usefulness and real-life impact of this algorithm would need to be established in a prospective clinical setting. Moreover, while ML-based analysis of EEG holds important promises, it will only ever serve as additional data to physicians, allowing them and their patients to make more informed decisions.

The best models for each outcome had a ROC AUC on the internal validation cohort of 0.67, 0.64, and 0.66 for seizure recurrence at one year, epilepsy diagnosis, and active epilepsy, respectively. While these performances are statistically significant, their clinical usefulness could be questioned, especially with regards to the limited PPV and specificity at a given threshold. Two major restrictions impede predictive performances: the capacity of the model and the reliability of the labels used for training. First, the small variance seen across different models and outcomes might indicate that the amount of data used saturated the capacity of the features and models (i.e.: underfitting). Machine learning studies consistently show that dataset size is correlated with predictive performances given sufficient capacity of the model [176], [177]. The next step to improve performances would therefore consist in gathering more training data to increase the complexity of the EEG features and ML architecture. The second hurdle is the confidence in the labels. Epilepsy diagnosis is probabilistic by definition: a patient could be at “high risk” of seizure and therefore have epilepsy, but never go on to have seizure recurrence [2]. Nevertheless, the performances for this outcome were still above chance on the temporally shifted cohort. The models trained to predict the outcome “Active epilepsy” did not generalize well to the testing cohort. We used this outcome to test the hypothesis that having had recent seizures would strongly affect the EEG signal. The results indicate that the features that we extracted from the EEG signal are more affected by seizure propensity (risk of having seizure recurrence) than past seizures. Ultimately, all these outcomes depend on the imperfect reporting of seizures by patients [49]; there

is promise that in the future, more objective outcomes could directly benefit predictive models in epilepsy [51].

EEG signals are altered by age, sex, comorbidities, antiseizure medications, and possibly several other factors; however, the degree to which these variables can confound predictions made from the EEG signal is unknown [178], [179], [180], [181]. In our study, the models trained to predict seizure recurrence at one year were robust to clinical variables. Patients with a focal finding on neuroimaging had reduced predictive performance, but we did not observe worse performance in patients with focal slowing on EEG. There was a poor correlation of focal findings on neuroimaging to focal slowing on EEG: only 37% (72) of EEGs with focal findings on neuroimaging had focal slowing, and 38% (54/144) with focal slowing on EEG did not have a focal finding on neuroimaging. This poor correlation highlights that neuroimaging and EEG query different aspects of the nervous system. Further work is needed to improve performance of our algorithm in patients with underlying non-epileptic focal abnormalities by using spatially aware features (e.g. left-right or anterior-posterior gradients, topographical voltage maps) [78], [182] or connectivity features [93], [183]. The post-hoc analyses were mostly limited by the number of patients, especially under the robust nested CV framework.

The survival analysis showed that ASMs count and age were important predictors of seizure recurrence risk. Higher ASMs count is a proxy for refractory cases and higher seizure frequency. ASMs are, however, infrequently accounted for in previous studies [73], [97], [184]. In our case, grouping by number of ASMs did not affect performances for prediction of seizure recurrence, suggesting that extracted features were independent of ASM. Performances for patients with no ASM could not be reliably estimated because of the rare seizure recurrence in this subgroup (11 EEGs). A small sample size also prevented the subgroup analysis for predicting the success of ASM withdrawal (32 EEGs). These two clinical applications would need to be explored in future studies. Regarding age, older patients carry a higher disease burden and are at higher risk of syncope, transient neurologic episodes, and confusion—nonepileptic conditions that could lead a patient to undergo an EEG exam. This partly explains why the yield of routine EEG for epilepsy in older patients is much lower [180]. Here, seizure recurrence was rare in the older patient group (>60 y.o.), for which performance could not be estimated. As predicted, the addition of age as an input feature slightly improved predictions. The potential benefit of using age as a predictor may

be limited by the resulting increase in dimensionality that could overthrow the increase in information, especially in a data-scarce setting that is subject to overfitting.

The model had above-chance performances in the absence of IEDs for all outcomes. Previously, only a few studies tested the impact of IEDs on their model; most had a case-control design, and none validated their results on a separate validation set [69], [78], [83], [184]. This finding suggests that automated analysis of EEG would increase the yield of EEG even in the absence of IEDs (74.7% of all EEGs in our cohort). The low sensitivity of EEG for IEDs leads to delays in diagnosis and need for prolonged or repeated exam, so the "negative" subgroup of EEGs (without IEDs) would most directly benefit from an alternative and independent marker. The two-step classifier showed an increase in performance compared to IEDs or computational features alone. This could orient the clinical applications of such an algorithm, complementing the interpretation of the EEG reader when an EEG does not reveal IEDs. Recent studies have demonstrated that machine learning models can detect IEDs with expert-level performances [104], [185]. Our approach could complement these algorithms by predicting seizure risk in IED-negative EEG, to further improve the clinical value and objectivity of EEG.

BP had the greatest performances when used alone to predict seizure recurrence, followed by FuzzEn and LL. Studies have suggested that slight shifts can be observed in the frequency spectrum of patients with epilepsy [151], [152], [153], [186], [187]. This could be secondary to the pathologically increased interictal synchronization, but could also be explained by ASM, age, or other confounders. In our study, compared with other features, BP had greater decrease in performances in younger patients and in the absence of abnormal slowing. Interestingly, it was the most performant feature in the presence of multiple ASMs (in opposition to FuzzEn and SampEn), suggesting that the changes in the frequency spectrum are not related to these confounders. Regarding entropy, it is generally found to be lower in patients with focal and generalized epilepsy [80], [155]. Increased predictability and reduced complexity could result from the constraints imposed by epileptogenic processes [188]. Several algorithms have been proposed to estimate the entropy of physiological time-series, without clear evidence that this measure can embody seizure propensity [189]. Compared with other features, FuzzEn was more affected by the presence of IEDs, focal lesion, and in patients with focal epilepsy; other entropy markers followed a similar trend. Entropies also had poor performances in patients with ≥ 2 ASMs, in line with previous

studies on non-linear analysis of EEG [85]. This complementarity between BP and entropy could be leveraged with clinical priors in the modeling process.

Despite the methodological strengths of this study, some limitations must be highlighted. First, the data comes from a single center, preventing us to generalize the results to other institutions. Second, the data collection was retrospective. For some patients, follow-up might have been too short to detect seizure recurrence, and these patients would have been inappropriately flagged as "no seizure recurrence". This limitation decreases the potential effect size (our capacity to discriminate between groups), and is, as such, conservative in nature. Third, most patients were on ASM at the time of EEG. ASM are known to affect the EEG signal and several of the features used in this study, such as BP and entropy [151], [152], [189]. A larger sample size is required to estimate the performances of such markers in patients with no ASM or in those undergoing ASM withdrawal. Finally, while statistically significant, the clinical impact of the performances reported in this study is modest. Applied to the clinical setting, the models would affect risk estimation by only a few percent, as demonstrated by the modest PPV, NPV, specificity, and sensitivity. This could potentially be addressed by using more powerful (albeit data-hungry) models to represent the EEG data, such as deep neural networks. With these limitations in mind, the findings in this study still robustly suggest that there exist changes in the EEG signal other than IEDs that can inform us about long-term seizure propensity; this opens the door to the possibility of using automated markers of epilepsy in the clinical setting, and strongly motivates future research in this direction.

In conclusion, we demonstrate that there are changes other than IEDs in the EEG signal embodying seizure propensity. These changes have a predictive horizon of one year after the EEG and their significance is independent of IEDs, age, and number of antiseizure medications. While significant, the potential impact on decision making in the clinical setting is modest. Future work will focus on improving the representation of the EEG to increase the performances of this approach and evaluate its real-life impact on clinical decision making.

Acknowledgements: DKN reports unrestricted educational grants from UCB and Eisai, and research grants for investigator-initiated studies from UCB and Eisai. EL is supported by a scholarship from the Canadian Institute of Health Research. DKN and FL are supported by the Canada Research Chairs Program, the Canadian Institutes of Health Research, and Natural Sciences and Engineering Research Council of Canada. EBA is supported by the Institute for Data

Valorization (IVADO, 51628), the CHUM research center (51616), and the Brain Canada Foundation (76097). We would like to acknowledge the work of the CHUM EEG technologists for their contribution to the recording of the EEGs. We would also like to thank Manon Robert and Véronique Cloutier for their help regarding the access to the EEG data and the submission to the ethics review board.

Competing interests: None of the authors declare any conflict of interest. The funding sources were not involved in study design, data collection, analysis, redaction, nor decision to submit this paper for publication.

Data availability: All code written for this study will be publicly available upon publication at the following address: https://gitlab.com/chum-epilepsy/epilepsy_markers_reeg.git. Anonymized data in the form of extracted features with clinical outcomes will be provided upon reasonable request to the corresponding author and conditional to the approval by our ethics research board. The TRIPOD reporting checklist is provided as supplementary material.

Contributors: EL, DT, GPMD, DKN, FL, and EBA conceived and planned the experiments. EL, AQX, MJ, and JDT collected the data. EL, AQX, MJ, JDT, and EBA had direct access and verified the underlying data. EL performed the experiments. EL, DT, DKN, FL, and EBA contributed to the interpretation of the results. EL wrote the first draft of the manuscript. All authors provided critical feedback and reviewed the manuscript.

Supplementary Information: The online version contains supplementary material available at https://www.nature.com/articles/s41598-023-39799-8.

CHAPITRE 5 ARTICLE 2: IMPROVING DIAGNOSTIC ACCURACY OF ROUTINE EEG FOR EPILEPSY USING DEEP LEARNING

Emile Lemoine[1,2,3], Denahin Toffa[1,3], An Qi Xu[3], Jean-Daniel Tessier[1,3], Mezen Jemel[1,3], Frédéric Lesage[2], Dang K. Nguyen[1,3†], Elie Bou Assi[1,3†*]

[1]Department of Neuroscience, Université de Montréal, Canada

[2]Institute of biomedical engineering, Polytechnique Montréal, Canada

[3]Centre de Recherche du Centre hospitalier de l'Université de Montréal (CRCHUM), Canada

[†]These authors should be considered co-senior authors

[*]Corresponding author

Dans cet article, nous nous attaquons au troisième objectif de la thèse, qui est d'améliorer la performance diagnostique de l'analyse computationnelle de l'EEG à l'aide de modèles profonds. Ce travail a été accepté pour publication dans la revue *Brain Communications* en date du 22 août 2025. Il a fait l'objet de présentations orales locales au service de neurologie du CHUM et au séminaire du programme Clinicien-chercheur de l'Université de Montréal (2024; Prix Jacques-Lacroix pour meilleure présentation orale), ainsi que des présentations par affiches à l'*American Epilepsy Society Annual Meeting* (Los Angeles, 2024) et à l'Institut Neurologique de Montréal (Montréal, 2024).

Ma contribution à cet article comprend l'identification de la problématique, la collecte de données, le développement de la méthode d'apprentissage profond, le prétraitement des données EEG de routine, la maintenance des ressources computationnelles, la réalisation des expériences, l'interprétation et l'analyse des résultats, la conception des visualisations, ainsi que la rédaction du manuscrit.

5.1 Abstract

The yield of routine EEG to diagnose epilepsy is limited by low sensitivity and the potential for misinterpretation of interictal epileptiform discharges (IEDs). Our objective is to develop, train, and validate a deep learning model that can identify epilepsy from routine EEG recordings, complementing traditional IED-based interpretation. This is a retrospective cohort study of diagnostic accuracy. All consecutive patients undergoing routine EEG at our tertiary care center between January 2018 and September 2019 were included. EEGs recorded between July 2019 and September 2019 constituted a temporally shifted testing cohort. The diagnosis of epilepsy was

established by the treating neurologist at the end of the available follow-up period, based on clinical file review. Original EEG reports were reviewed for IEDs. We developed seven novel deep learning models based on Vision Transformers (ViT) and Convolutional Neural Networks (CNN), training them to classify raw EEG recordings. We compared their performance to IED-based interpretation and two previously proposed machine learning methods. The study included 948 EEGs from 846 patients (820 EEGs/728 patients in training/validation, 128 EEGs/118 patients in testing). Median follow-up was 2.2 years and 1.7 years in each cohort, respectively. Our flagship ViT model, DeepEpilepsy, achieved an area under the receiver operating characteristic curve (AUROC) of 0.76 (95% CI: 0.69–0.83), outperforming IED-based interpretation (0.69; 0.64–0.73) and previous methods. Combining DeepEpilepsy with IEDs increased the AUROC to 0.83 (0.77–0.89). DeepEpilepsy can identify epilepsy on routine EEG independently of IEDs, suggesting that deep learning can detect novel EEG patterns relevant to epilepsy diagnosis. Further research is needed to understand the exact nature of these patterns and evaluate the clinical impact of this increased diagnostic yield in specific settings.

5.2 Introduction

The diagnosis of epilepsy is notoriously challenging. It relies on the occurrence of either two seizures more than 24h apart, one seizure and a high risk of another, or the presence of an epilepsy syndrome.[2] Despite this clear definition, the rate of misdiagnosis remains high [190], [191], being highly dependent on the ability to collect a clear clinical history and accurately interpret the electroencephalogram (EEG).

The EEG can capture ictal and interictal activity, namely interictal epileptiform discharges (IEDs), which are highly specific for epilepsy (98%) [192]. A scalp EEG is cost-effective and technically straightforward, with standard acquisition protocols that have been put in place by the International League Against Epilepsy [193], [194]. However, the sensitivity of a single routine EEG for IEDs is 20–50%, and only 17% in adults after a first unprovoked seizure [10], [28], [30]. Furthermore, the interrater reliability for IEDs is fair to moderate even among experts, with a kappa of 35–50% [37], [195], [196]. Consequently, the EEG has limitations as a diagnostic tool in patients with suspected seizures, with EEG misinterpretation contributing to diagnostic errors in epilepsy [39]. The identification of additional biomarkers beyond IEDs could help overcome these limitations and improve diagnostic accuracy [49], [50].

In recent decades, efforts have focused on overcoming the limitations of traditional EEG interpretation by identifying alternative epilepsy biomarkers through computational methods [67], [145], [150], [153], [197]. While these approaches have shown promise, their translation to clinical practice has been limited by several factors: modest performance [102], [197], [198], small [100] or lack of dedicated [77], [78], [88], [199], [200] testing set, exclusion of patients with neurological comorbidities or abnormal EEGs [100], [199], [200], and reliance on IED-detection [199], [200]. As a result, the expected diagnostic accuracy of these approaches in a real-world population is uncertain.

Deep learning (DL) has emerged as a powerful tool for the analysis of complex signals. DL models can autonomously extract features from time-series or images by optimizing millions of parameters on large datasets. DL has been applied to EEG to decode brain signals for brain-computer interface [119], predict delirium [201], and automatically detect IEDs [103], [104]. Given DL's capacity to capture the complex brain dynamics, we hypothesized that it could enhance the detection of epilepsy-specific patterns on routine EEG recordings.

The present study seeks to address these questions: can modern DL models detect epilepsy on interictal EEG, even in the absence of IEDs? What are the potential diagnostic performances of a DL-assisted EEG interpretation for epilepsy? And what sample size is required to train such models?

5.3 Materials and Methods

5.3.1 Study design

This is a retrospective study on a consecutive cohort of patients undergoing routine EEG in a single tertiary care center in Montreal, Canada.

5.3.2 Participants

We included all patients who underwent a routine EEG (20- to 60-minute, with or without sleep deprivation) between January 2018 and September 2019 at the Centre Hospitalier de l'Université de Montréal (CHUM). Exclusion criteria were the absence of follow-up after the EEG, an uncertain diagnosis of epilepsy at the end of the available follow-up period, or an EEG performed in a hospitalized patient. Under a prespecified protocol, one neurology resident (EL) and three students (AQ, MJ, JDT) collected data from the electronic health record for each visit, including baseline

characteristics (age, sex), comorbidities, number of antiseizure medications, and presence of a focal lesion on neuroimaging. They also reviewed the EEG report for the presence of IED(s) and abnormal background slowing. All clinical information was stored on a REDCap database hosted on the CHUM research center's servers.

We separated the cohort into two independent subsets according to the date of the EEG. Recordings before July 15, 2019, comprised the training and validation set, while recordings after July 15, 2019, comprised the testing set. We excluded from the testing set any recording from a patient already included in the training and validation set. The training and validation set was further separated into a training set and a validation set in a random fashion (80%/20% split).

5.3.3 Test Methods

Reference Standard

The reference standard is the diagnosis of epilepsy according to the treating physician at the end of the available follow-up period. This diagnosis is based on the ILAE definition of epilepsy, i.e. having had two unprovoked seizures more than 24h apart or one unprovoked seizure and be considered at high (>60%) risk of seizure recurrence, or being diagnosed with an epilepsy syndrome [2]. The final diagnosis at the end of the follow up period was used, as opposed to the speculated diagnosis at the time of the EEG, because the follow up period provides additional information such as imaging, additional EEG recordings, video-monitoring admissions, and seizure recurrences.

EEG recording

EEGs were recorded using a standardized protocol on a Nihon Kohden EEG system, following national recommendations [158]. Awake EEGs, 20–30 minutes long, were recorded at 200 Hz with 21 electrodes arranged with the 10-20 system. They included two 90-second periods of hyperventilation (except in patients >80 years old, uncooperative, or with medical contraindications) and photic stimulation from 4 Hz to 22 Hz. Patients were also instructed to open or close their eyes at several times. Sleep deprived recordings lasted 60 minutes, with the same activation procedures. Technologists annotated the EEG in real-time. For this study, EEGs were converted to an average referential montage (A1-A2), saved to EDF format, and stored on the CHUM research center's server for analysis.

Automated processing of EEG and classification

The index test is the classification of the EEG recordings using machine learning. We developed DeepEpilepsy, a Vision Transformer (ViT) model that takes raw EEG segments as input and outputs a probability of the diagnosis of epilepsy (Figure 5.1). EEGs in average referential montage (19 channels) were segmented into overlapping 10- or 30-second windows (95% overlap) and directly used as input into the DL models. The input dimensions were $19 \times 200t$, where t is the window size in seconds. We initially explored two window sizes (10 and 30 seconds) based on common practice in EEG interpretation and computational constraints [67]. A 95% overlap between segments was used to provide sufficient data augmentation while maintaining computational feasibility. We also investigated the impact of other window sizes on performances (**eTable 6**). The model configurations for the ViT models, including DeepEpilepsy, are presented in **eTable 1**. To enhance model generalization, we applied a random data augmentation algorithm during training.[134] For each segment, an augmentation was drawn randomly from a set of transformations, which included filtering (band-pass, low-pass, high-pass), masking (channel, time), and adding noise (**eFigure 1**). These were applied with a 50% probability and randomized intensity. We performed a Bayesian hyperparameter search on the training and validation set to choose DeepEpilepsy's final configuration. We also investigated different learning rates, weight decay, and batch size values. The final models were trained on the entire training and validation set. The optimization hyperparameters and model specifications are described in **eTable 4**.

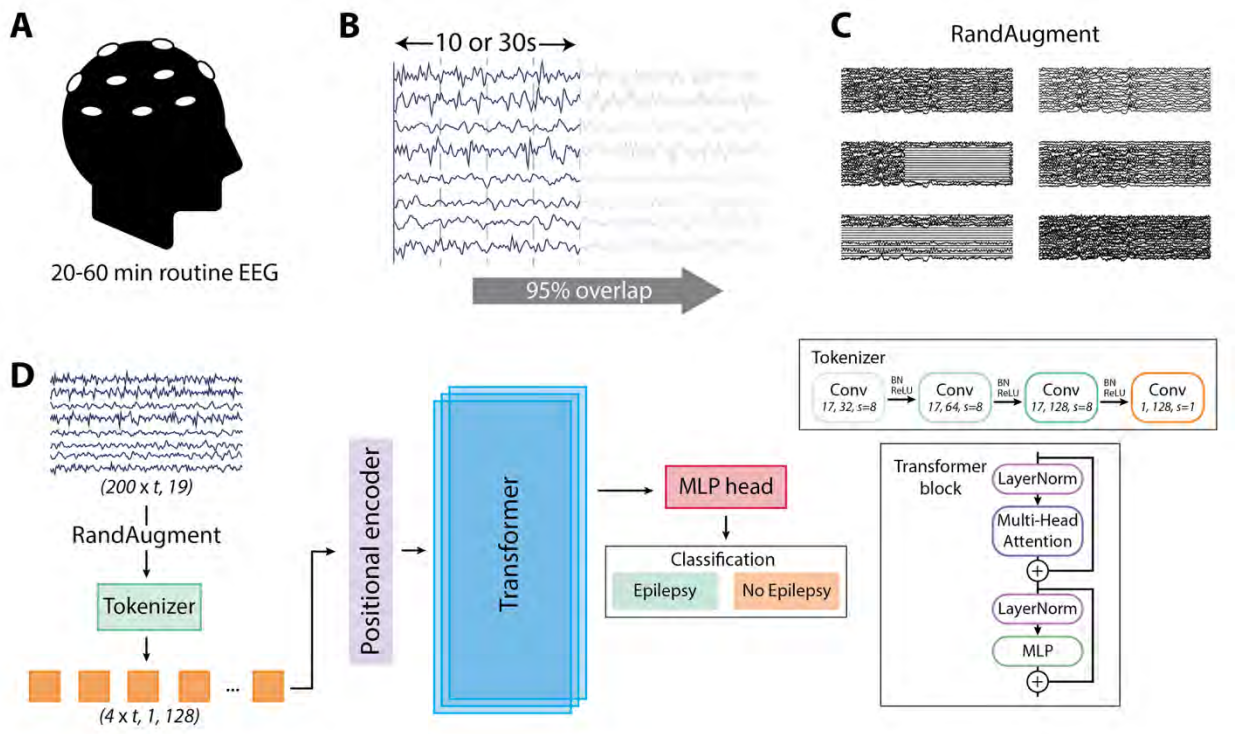


Figure 5.1: Details of the DeepEpilepsy Transformer model. The EEG is first processed through the RandAugm algorithm with 50% probability. A tokenizer is used (upper right: convolutional tokenizer) before positional encoding. The tokens are then input into a Transformer model. A MLP head classifies the embeddings from the Transformer according to the diagnosis of epilepsy. BN: Batch normalization; Conv: Convolutional layer; MLP: Multilayer perceptron; RandAugment: Random Augmentation; ReLU: Rectified linear unit.

In addition, we implemented other Deep Learning models (ViT and ConvNeXt; **eTable 1 and 2**), as well as two previously described methods: the ShallowConvNet inspired by the *Filter Bank Common Spatial Patterns* algorithm (**eTable 3**) [125], and a feature-extraction framework relying on the extraction of linear and nonlinear EEG markers that are used as input into a classifier (LightGBM) [198]. These methods are described in detail in **eMethods 1**.

To obtain the diagnostic performances, the final models/procedures were applied to the testing set. This resulted in a single predicted probability for each EEG segments. To obtain one prediction per EEG recording, we aggregated the predicted probabilities at the EEG-level using the median of the predicted values. In cases where patients had multiple EEGs, each recording was treated as an independent observation. A sensitivity analysis excluding repeated EEGs was performed to assess potential bias.

We further evaluated DeepEpilepsy in a specific subgroup of patients which were not yet diagnosed with epilepsy at the time of the index EEG (i.e., undergoing evaluation for suspected seizures). We also measured the performance bias across different subgroups: age groups (18–40, 40–60, and >60 years old), sex, presence of focal lesion, presence of IED (absence, presence, and uncertain), presence of slowing, sleep deprivation before EEG, and number of ASM (0, 1, ≥2).

5.3.4 Analysis

We calculated the AUROC using the probabilistic predictions for each model, with 95% confidence intervals estimated using DeLong's method (single prediction by patient) [165]. We also computed the Area Under the Precision-Recall Curve (AUPRC) with 95% confidence intervals estimated using bootstrap resampling (1000 iterations). For comparison, we tested the classification performance of IEDs alone (presence vs. absence). We also tested a two-step classification using IEDs first (traditional EEG interpretation), followed by DeepEpilepsy if IEDs were absent (DL interpretation).

The optimal classification threshold was obtained using the validation cohort, minimizing the distance between the curve and the upper left corner of the ROC graph. This threshold was then applied to compute sensitivity, specificity, negative predictive value, and positive predictive value on the testing set.

We performed additional analyses to better quantify the effect of window duration and random augmentation on DeepEpilepsy. For segment duration, we re-trained the model using window durations of 5s, 10s, 15s, 30s, 45s, and 60s (with fixed 1.5s overlap to maintain consistent training sample size). For random augmentation, we re-trained DeepEpilepsy eight times with and without RandAugment (20 epochs each). Performances between augmented and non-augmented models were compared with AUROC on the testing set as the performance metric.

We performed an exploratory analysis of the embeddings learned by DeepEpilepsy and ShallowConvNet to better understand the patterns captured by both models (**eMethods 2**). Embeddings are the internal representations that deep learning models create while processing raw EEG data - they represent how the model "sees" the EEG after transforming it through multiple layers, without any pre-specified features. These learned representations differ from traditional EEG features and can provide insights into what patterns the model considers important for classification.

5.3.5 Sample size

Using Obuchowski's method [202], with a 60% epilepsy prevalence, a power of 0.9, and a significance level of 0.0071 (adjusted from 0.05 divided by 7 DL models), a minimum of 126 EEGs is required to detect an AUROC of 0.70.

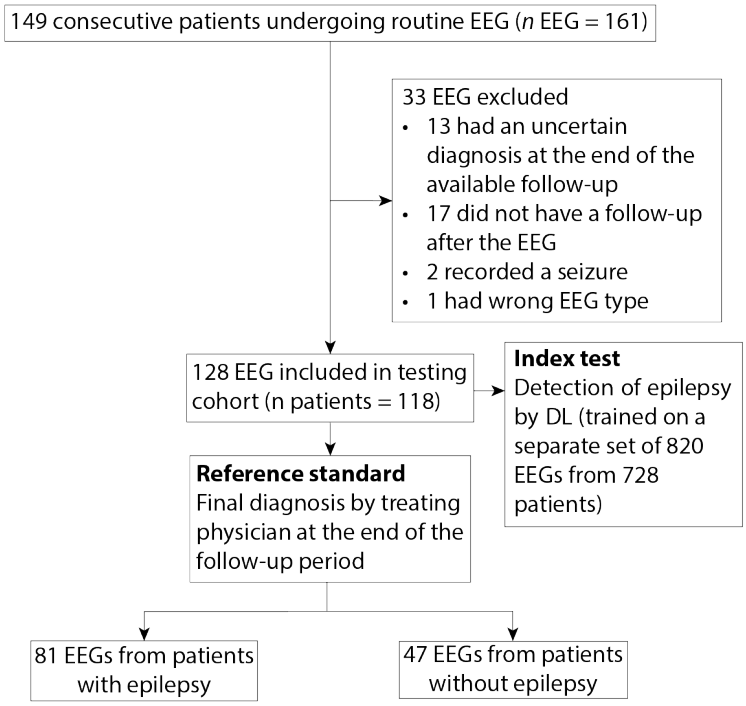


Figure 5.2 : Flowchart of patients included in the testing cohort.

5.3.6 Standard Protocol Approvals, Registrations, and Patient Consents

Ethics approval was granted by the CHUM Research Centre's Research Ethics Board (REB) (Montreal, Canada, project number: 19.334). The REB waived informed consent due to the lack of diagnostic/therapeutic intervention and minimal risk to participants. All methods followed Canada's Tri-Council Policy statement on Ethical Conduct for Research Involving Humans.

5.4 Results

5.4.1 Participants

After exclusion, 948 EEGs from 846 patients were included: 820 EEGs in the training/validation set (728 patients) and 128 EEGs in the testing set (118 patients), with no patient overlap (Table

5.1). Before exclusion, 1 185 EEGs from 1 067 patients and 161 EEGs from 149 patients met the inclusion criteria for the training and testing cohorts, respectively. Reasons for exclusion were absence of follow-up after the EEG, uncertain diagnosis at the end of available follow-up, seizure during the EEG, and wrong EEG type (i.e., performed in a hospitalized patient) (Figure 5.2). Median age was 49 and 51.5 (IQR: 32–62 and 30–62.5) and the proportion of women were 51% and 62.5% in the training and testing cohorts, respectively. Median follow-up was 2.2 years (IQR: 1.0–2.9) and 1.7 years (IQR: 0.9–2.3). Epilepsy prevalence was 63% in both sets.

Table 5.1 : Description of the training (EEG recordings between January 2018 and July 2019) and testing cohorts (EEG recordings between July and September 2019)

	Training/validation cohort (n = 820)		**Testing cohort (n = 128)**	
	Epilepsy	**No Epilepsy**	**Epilepsy**	**No Epilepsy**
Number of EEGs	517	303	81	47
Sex = woman (%)	259 (50.1)	159 (52.5)	54 (66.7)	26 (55.3)
Age (median [IQR])	42.00 [29.00, 58.00]	57.00 [41.00, 67.00]	37.00 [25.00, 57.00]	60.00 [50.50, 71.00]
Total follow-up after EEG in weeks (median [IQR])	133.50 [95.75, 173.00]	59.00 [17.00, 116.00]	99.50 [70.25, 125.00]	62.00 [17.00, 102.00]
Epilepsy type (%)				
Focal	370 (71.6)	–	49 (60.5)	–
Generalized	119 (23.0)	–	26 (32.1)	–
Unknown	28 (5.4)	–	6 (7.4)	–
Age of epilepsy onset (median [IQR])	22.00 [13.00, 40.00]	–	23.00 [14.00, 48.00]	–
Seizure recurrence after EEG (%)	269 (52.0)	0 (0.0)	44 (54.3)	0 (0.0)
Number of days since last seizure (median [IQR])	237 [56, 1134]	–	118 [44, 467]	–
Number of epilepsy risk factors (median [IQR])	3 [1, 4]	2 [1, 4]	2 [1, 3]	1 [0, 3]
History of epilepsy surgery (%)	60 (11.6)	0 (0.0)	4 (4.9)	0 (0.0)
Number of ASM (%)				
0	55 (10.6)	253 (83.5)	17 (21.0)	42 (89.4)
1	280 (54.2)	36 (11.9)	34 (42.0)	5 (10.6)
2	123 (23.8)	12 (4.0)	19 (23.5)	0 (0.0)
3	47 (9.1)	2 (0.7)	6 (7.4)	0 (0.0)
4	10 (1.9)	0 (0.0)	5 (6.2)	0 (0.0)
5	2 (0.4)	0 (0.0)	0 (0.0)	0 (0.0)
Focal lesion on brain imaging (%)	223 (43.1)	84 (27.7)	31 (38.3)	10 (21.3)
Sleep deprived EEG (%)	62 (12.0)	50 (16.5)	22 (27.2)	8 (17.0)
IED (%)				
Absence	333 (64.4)	282 (93.1)	42 (51.9)	46 (97.9)
Presence	139 (26.9)	2 (0.7)	30 (37.0)	0 (0.0)
Uncertain	45 (8.7)	19 (6.3)	9 (11.1)	1 (2.1)
Abnormal slowing on EEG (%)	199 (38.5)	46 (15.2)	32 (39.5)	10 (21.3)

In the training cohort, 141 EEGs (17%) showed definite IEDs and 64 (8%) showed uncertain IEDs. Two definite IEDs were found in patients without epilepsy. In the testing cohort, 30 EEGs (23%)

showed definite IEDs and 10 (8%) showed uncertain IEDs, with all definite IEDs in patients with epilepsy.

5.4.2 Test Results

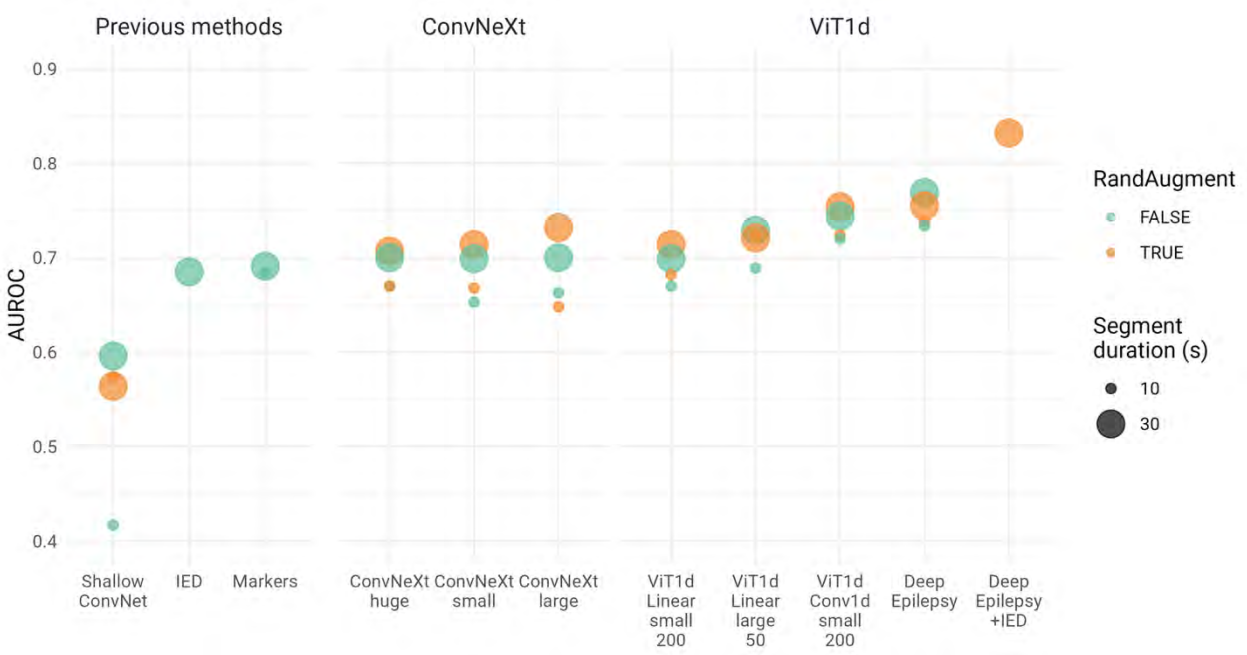


Figure 5.3 : Diagnostic performances of automated EEG analysis for the diagnosis of epilepsy on the testing set (n = 128). Our flagship model, DeepEpilepsy, is shown alone and combined with traditional EEG interpretation based on the identification of IED. The other novel approaches shown are ViTs and ConvNeXt using different configurations (size: small, large, huge; tokenizers: convolutional or linear; window size: 50 pt or 200 pt) as well as presence of RandAugm and the duration of EEG segments used as input. Previous methods are the ShallowConvNet,[23] extraction of computational markers,[21] and the presence of IEDs on EEG. AUROC: Area under the receiver operating characteristic curve; IED: interictal epileptiform discharges; ViT: Vision Transformers.

The AUROC for the diagnosis of epilepsy in the testing cohort for every approach is pictured inFigure 5.3. For DeepEpilepsy, the AUROC was 0.76 (95%CI: 0.69–0.83) and AUPRC of 0.88 (0.83–0.94) (Figure 5.4). Using the threshold computed on the validation cohort (0.86), there were 75 true positives, 38 true negatives, 13 false positive, and 41 false negatives, equating to a sensitivity of 64.7%, a specificity of 74.5%, a positive predicted value (PPV) of 85.2%, and a negative predictive value (NPV) of 48.1%. For comparison, when using the presence of IEDs on EEG (as per the EEG report) as the index test, the sensitivity is 37.0%, specificity is 100.0%, PPV is 100.0%, and NPV is 41.1%, with an AUROC of 0.69 (95% CI: 0.64–0.73) and AUPRC of 0.86

(0.82–0.91) (Figure 5.4). The AUROC of DeepEpilepsy was higher than any other method, although this was only statistically significant when compared to the ShallowConvNet models (AUROC: 0.60, 95%CI: 0.50–0.69). The diagnostic performances of all methods are shown in Table 5.2.

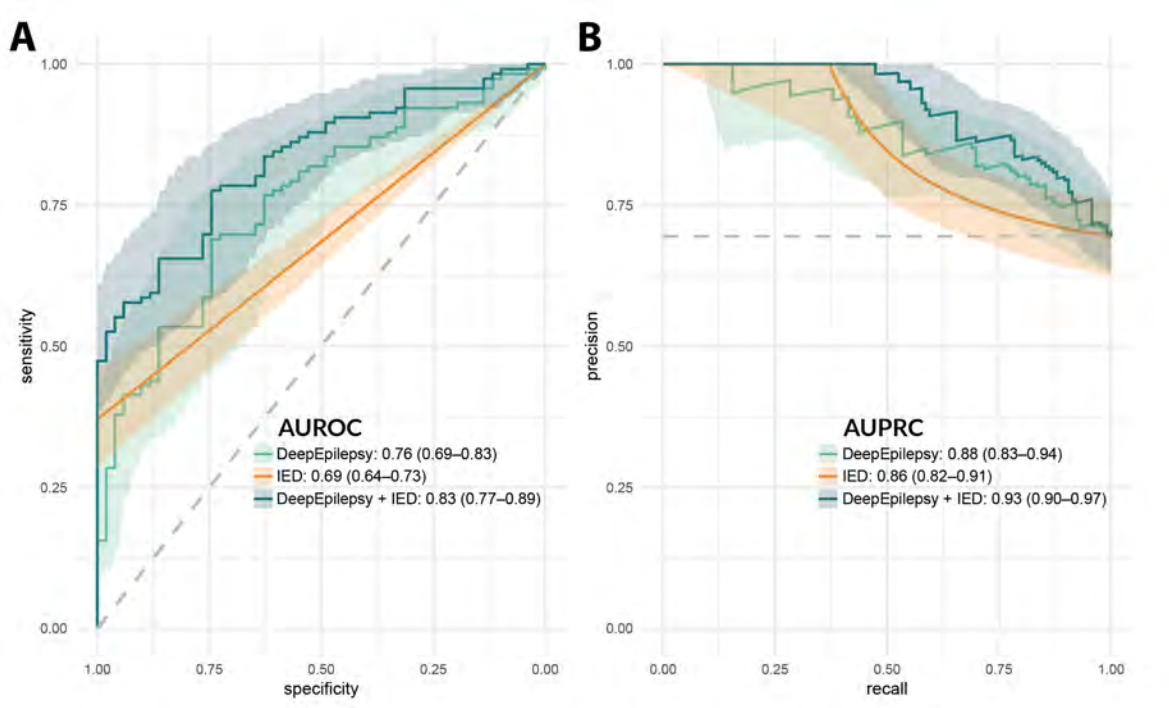


Figure 5.4 : Diagnostic performances on the testing set (*n* = 128). A: ROC curves for DeepEpilepsy, IEDs only, and DeepEpilepsy combined with IEDs in the testing cohort. B: Precision-recall curves for the three approaches. AUPRC: Area under the Precision-Recall curve; AUROC: Area under the receiver operating characteristic curve; IED: interictal epileptiform discharges.

When using the two-step model as the index test (1: presence of IED classified as epilepsy, 2: if no IED: DeepEpilepsy models prediction), the AUROC was 0.83 (95%CI: 0.77–0.89) and AUPRC was 0.93 (0.90–0.96) (Figure 5.4). The sensitivity, specificity, PPV, and NPV were 73.2%, 74.5%, 86.7%, and 55.1%.

A sensitivity analysis excluding repeated EEGs from the testing set showed similar performance, with DeepEpilepsy achieving an AUROC of 0.74 (*n* = 118; 95% CI: 0.65–0.81).

Table 5.2 : Classification performances on the testing set for all machine learning methods

	Segment duration (s)	RandAugment	AUC

DeepEpilepsy	30	False	0.77 (0.69--0.84)
DeepEpilepsy	30	True	0.76 (0.68--0.83)
ViT1d, Conv tokenizer, small	30	True	0.75 (0.68--0.83)
ViT1d, Conv tokenizer, small	30	False	0.74 (0.66--0.82)
DeepEpilepsy	10	True	0.74 (0.66--0.81)
DeepEpilepsy	10	False	0.73 (0.64--0.81)
ConvNeXt, large	30	True	0.73 (0.65--0.81)
ViT1d, Linear tokenizer, large	30	False	0.73 (0.65--0.80)
ViT1d, Conv tokenizer, small	10	True	0.72 (0.64--0.80)
ViT1d, Linear tokenizer, large	10	True	0.72 (0.64--0.80)
ViT1d, Linear tokenizer, large	30	True	0.72 (0.64--0.80)
ViT1d, Conv tokenizer, small	10	False	0.72 (0.64--0.80)
ConvNeXt, small	30	True	0.71 (0.63--0.80)
ViT1d, Linear tokenizer, small	30	True	0.71 (0.63--0.79)
ConvNeXt, huge	30	True	0.71 (0.62--0.79)
ConvNeXt, huge	30	False	0.70 (0.61--0.78)
ConvNeXt, large	30	False	0.70 (0.62--0.78)
ViT1d, linear tokenizer, small	30	False	0.70 (0.61--0.78)
ConvNeXt, small	30	False	0.70 (0.61--0.78)
Feature extraction with LightGBM	30	---	0.69 (0.60--0.78)
ViT1d, Linear tokenizer, large	10	False	0.69 (0.60--0.76)
Feature extraction with LightGBM	10	---	0.68 (0.59--0.77)
ViT1d, linear tokenizer, small	10	True	0.68 (0.59--0.76)
ConvNeXt, huge	10	False	0.67 (0.58--0.76)
ConvNeXt, huge	10	True	0.67 (0.58--0.75)
ViT1d, linear tokenizer, small	10	False	0.67 (0.58--0.75)
ConvNeXt, small	10	True	0.67 (0.58--0.76)
ConvNeXt, large	10	False	0.66 (0.58--0.75)
ConvNeXt, small	10	False	0.65 (0.57--0.74)
ConvNeXt, large	10	True	0.65 (0.56--0.73)
ShallowConvNet	30	False	0.60 (0.49--0.69)
ShallowConvNet	10	True	0.57 (0.47--0.67)
ShallowConvNet	30	True	0.56 (0.46--0.66)
ShallowConvNet	10	False	0.42 (0.32--0.51)

5.4.3 Subgroup analyses

In the testing cohort, 75 patients (64%) had an uncertain diagnosis at the time of the EEG, 28 of which were eventually diagnosed with epilepsy. In the 47 others, the most common final diagnoses were syncope/faintness (11), dementia-related fluctuations (6), and non-specific sensitive symptoms (5). Within this subgroup of uncertain diagnoses, 10 patients who were diagnosed with

epilepsy showed IEDs and 6 had uncertain sharp transients (vs. 1 in patients without epilepsy). The complete characteristics of the subgroup are detailed in **eTable 5**.

In the subgroup of 75 patients not diagnosed with epilepsy at the time of the EEG, DeepEpilepsy still had above-chance performances (AUROC: 0.69, 95%CI 0.56–0.80), and the two-step model had the following performances: sensitivity of 65.6%, specificity of 76%, PPV of 63.6% and NPV of 77.6%, with an AUROC of 0.77 (0.65–0.87). The ROC curves for IEDs only, DeepEpilepsy, and DeepEpilepsy combined with IEDs for this subgroup are shown in Figure 5.4.

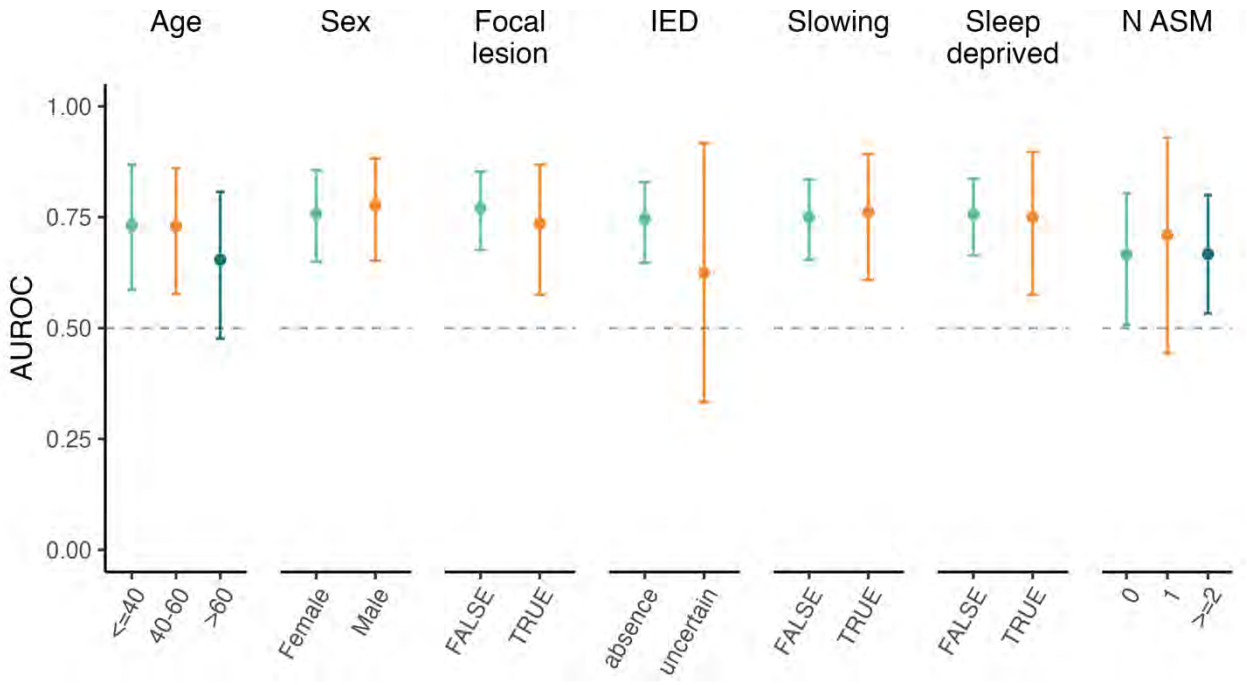


Figure 5.5 : Performance of DeepEpilepsy for classification of epilepsy diagnosis from routine EEG in different subgroups of the testing set. The subgroups have the following sample sizes: 1) Age < 40: *n* = 40, >40–≤60: *n* = 44, >60: *n* = 44; 2) male: *n* = 48, female: *n* = 80; 3) focal lesion: *n* = 41, no focal lesion: *n* = 87; 4) uncertain IED: *n* = 10, absence of IED: *n* = 88. 5) focal slowing: *n* = 42, no focal slowing: *n* = 86; 6) Sleep deprived EEG: *n* = 30, awake EEG: *n* = 98; 6) no ASM: n = 59, one ASM: *n* = 39, ≥2 ASMs: n = 30. ASM: Antiseizure medication; AUROC: Area under the receiver operating characteristic curve; IED: interictal epileptiform discharges.

The results for other subgroups are presented in Figure 5.5. Across all subgroups, performances were above chance except for patients > 60 years old and patients with a single antiseizure medication. Notably, in absence of IEDs (*n* = 98), AUROC was 0.74 (0.65–0.83), with NPV of 0.55%, PPV of 76%, sensitivity of 57%, and specificity of 75%. By comparison, in patients where DeepEpilepsy predicted low epilepsy risk (*n* = 79), IEDs had a AUROC of 0.62 (0.56–0.68), with

NPV of 55%, PPV of 100%, sensitivity of 24%, and specificity of 100%. Also, DeepEpilepsy performed similarly in sleep deprived EEG and awake EEGs (AUROC = 0.76 [0.67–0.84] and 0.76 [0.58–0.90], respectively).

5.4.4 Sample size, segment duration, and RandAugment analysis

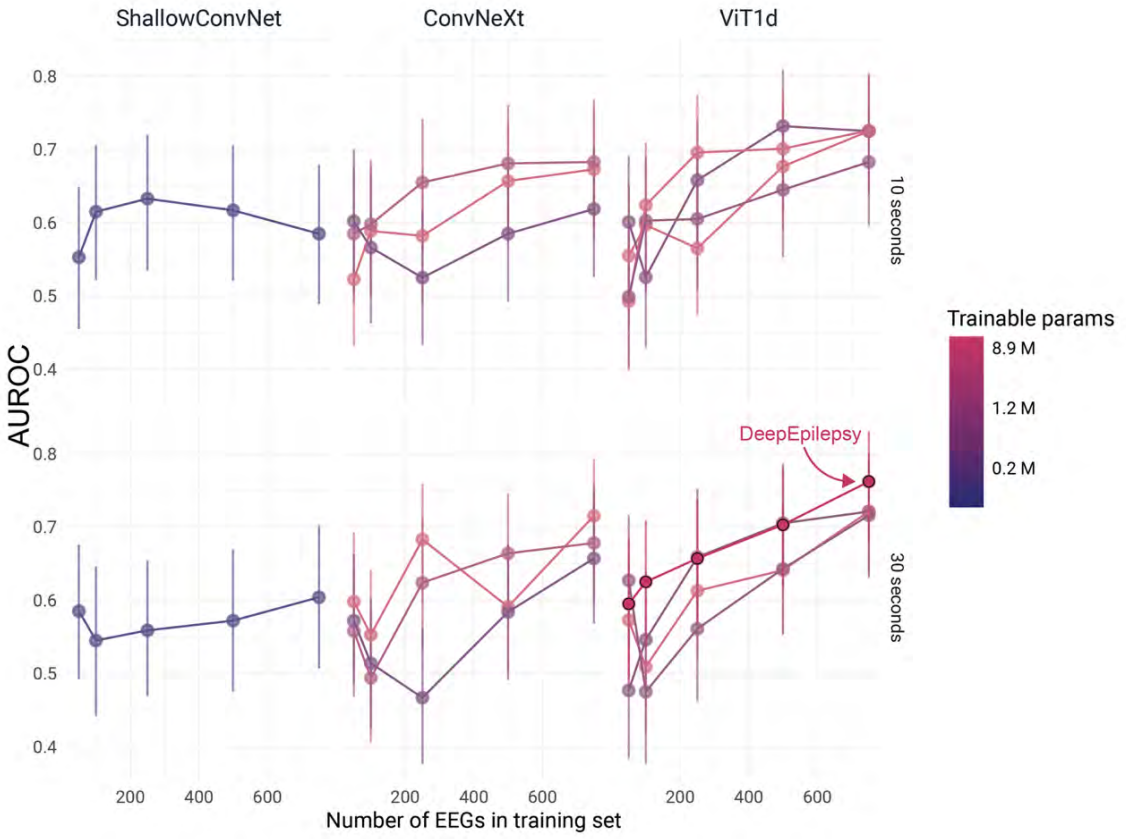


Figure 5.6 : Impact of training sample size on the performance of four deep learning models (ShallowConvNet, ConvNeXt, DeepEpilepsy, and other Vision Transformers) for detecting epilepsy from EEG segments. Performance is measured by the AUROC score on the testing set (n = 128), with models trained on varying numbers of EEGs (50, 100, 250, 500, and 750). The models were trained on 10s (top row) and 30s (bottom row) overlapping EEG segments. AUROC: Area under the receiver operating characteristic curve; IED: interictal epileptiform discharges; ViT1d: Vision Transformer with one-dimension tokenizer.

We trained the different neural network models on subsets of the data (50, 100, 250, 500, and 750 EEGs) to assess the impact of the size of the training sample on performance (Figure 5.6). With 10-second segments, the ShallowConvNet had highest performances when trained on 250 EEG recordings. The other models tended towards increased performances, with a ceiling at 500 EEGs. Using 30-second segments, the ShallowConvNet showed a slight increase in performances with

increased training size, with a maximal AUROC of 0.6 at 750 EEGs. In contrast, the performance of the ConvNeXt and ViT models increased almost linearly with sample size, achieving the highest performances with 750 EEGs. In almost all cases, 500 EEGs was the minimal training size required to achieve above-chance performances. For reference, using our segmentation strategy, 500 EEGs resulted in 765,000 10-second overlapping segments or 500,000 30-second overlapping segments.

A systematic evaluation of segment durations confirmed that 30-second windows achieved optimal performance (**eTable 6**). Models trained with RandAugment showed higher maximal performance, but increased variability compared to models without data augmentation (max AUROC 0.73 vs 0.72, mean AUROC 0.71 vs. 0.71, standard deviation of AUROC: 0.011 vs. 0.017).

5.4.5 Relationship between learned representations and traditional EEG features

Deep learning models such as DeepEpilepsy transform raw EEG signals into hidden representations (embeddings) that are optimal for distinguishing patients with and without epilepsy. To understand what patterns these models capture, we analyzed how these embeddings relate with traditional EEG features (namely band power and entropy) using clustering analysis. For band power, DeepEpilepsy's embeddings showed higher variance in the high-frequency range (> 13 Hz), particularly in the 20–40 Hz, 40–75 Hz, and 75–100 Hz bands. In contrast, ShallowConvNet's embeddings exhibited relatively higher variance in the low-frequency range (< 10 Hz) (**eFigure 4**). Although DeepEpilepsy showed significant heterogeneity across all frequency bands, ShallowConvNet had non-significant analysis of variance in the 20–40 Hz range ($p = 0.24$) Regarding entropy, both models showed significant heterogeneity across all frequencies, but ShallowConvNet displayed higher inter-cluster variance, especially for bands above 1.6 Hz, suggesting that this was a key feature learned by this model (**eFigure 5**).

5.5 Discussion

This study assessed the diagnostic performance of DL-based analysis of routine EEG for epilepsy. We developed and trained the DL models on 948 consecutive EEGs from 846 patients, testing them on a temporally shifted cohort of 128 EEGs from 118 patients. Our flagship model, DeepEpilepsy, had a testing AUROC of 0.76 (95%CI: 0.69–0.83), outperforming other methods including conventional IED-based interpretation and previously proposed computational methods.

Combining the presence of IEDs with DL analysis increased the AUROC to 0.83 (95%CI: 0.77–0.89), demonstrating a potential for clinical translation.

Epilepsy diagnosis is primarily clinical, guided by individualized seizure recurrence risk assessment, which can be challenging due to limited reliable data [2]. The identification of IEDs on rEEG is commonly used to support the diagnosis of epilepsy, but their low sensitivity and risk of over-interpretation can often lead to both over- or underdiagnosis [39]. In our study, IEDs had an AUROC of 0.69 with a sensitivity as low as 37%. Our DL models provided higher overall diagnostic performances from the EEG than IEDs. Combining both approaches allowed to leverage the model's higher sensitivity and the high specificity of IEDs. Currently, no definitive, quantitative, non-ictal biomarkers have been validated for clinical use [2]. Although several studies have explored changes in the EEG such as shifts in band power [151], [152], [153] or changes in entropy [80], [203], many remain at the "proof-of-concept" stage, limited by case-control designs or inadequate validation [67]. More recent studies on computational analysis of EEG for the diagnosis of epilepsy have shown mixed results [102], [197]. Unlike prior work [67], our validation cohort corresponds to the group of patients in which the algorithm would be used in real-life, reducing bias in performance evaluation. Furthermore, the gold-standard in our study was based on a thorough review of clinical notes with a median follow-up period of over two years, allowing the clinician to build a more complete clinical picture integrating seizure recurrence, imaging, video-EEG evaluations, or new clinical symptoms. This is in contrast with studies that based the diagnosis on the EEG report or a single clinical visit [67]. These methodological strengths reduce bias and represent key steps towards the clinical integration of automated EEG analysis [67].

DeepEpilepsy is based on the Transformer architecture [17], which has greatly advanced our capacity to model sequence data. Transformers have been adapted for EEG-based tasks such as eye-tracking [128], seizure prediction [136], [137], and decoding of motor patterns [129]. A critical component in adapting Transformers to EEG is the tokenization method, which influences feature extraction and the timescales captured by the model. Previous studies have used separable convolutions as the tokenizer [128], [129], a popular approach in EEG models since the ShallowConvNet and EEGNet CNNs [118], [119]. However, in our early experiments, we found this approach underperformed and was inefficient, leading us to discard it. In contrast to the original ViT model, which "patchified" the input signal with a linear, non-overlapping tokenizer [130], we showed that a deep convolutional embedding results in higher performances. This improvement is

likely due to the convolution's inductive bias towards hierarchical dynamics across timescales and spatial scales [204]. The discrepancies between our findings and previous studies on Transformer-based EEG models probably arise, in part, from dataset size and complexity: our training dataset included over 1 million samples from more than 900 patients, while prior studies used significantly smaller training samples (15 000–80 000 segments from 23–70 patients [127], [128], [136], [137]) as well as shorter EEG segments (up to 50 000 points [127], [128], [136], [137], compared to our 114,000 points per segment).

A notable advantage of Transformers over CNNs is their scalability. DeepEpilepsy showed continual improvement as the size of the training sample increased, without hitting a performance ceiling. Recent studies have further demonstrated CNNs' limitations in scaling to large EEG datasets [122]. While data augmentation through RandAugment increased model variability without clear performance benefits in our dataset, it might prove more valuable with larger training samples. The absence of a performance ceiling in DeepEpilepsy suggests potential for further improvements with larger datasets, motivating multicenter collaborations to expand the training sample.

Unlike other approaches to automated EEG interpretation that rely on explicit IED detection [103], [104], [199], [200], DeepEpilepsy was trained without specific emphasis on IEDs. The model's good performance in EEGs without IEDs (AUROC of 0.74) and its complementarity with IED-based classification suggests it captures additional epilepsy-specific patterns. Our embedding analysis suggests these patterns may be linked to changes in the higher frequency spectrum (40–100 Hz), which include the lower range of high-frequency oscillations (HFOs, typically 80–500 Hz). HFOs on intracranial EEG may have a prognostic value in patients with refractory temporal lobe epilepsies [205], [206], and some studies have successfully detected them on scalp EEG with promising correlation with seizure outcomes [207], [208], [209]. However, the role of HFOs on scalp EEG remains limited, largely due to technical challenges such as requirement for a high sampling frequency (most studies using > 500 Hz) and low signal-to-noise ratio [207].

The superiority of DeepEpilepsy over our benchmark model (LightGBM) [198], which used carefully selected traditional EEG features (spectral power, nonlinear measures, peak alpha frequency), suggests that learning from raw EEG data captures relevant patterns that might be missed by conventional analysis. This is particularly evident in two aspects: the model's sensitivity

to high-frequency patterns (40–100 Hz) and its improved performance with longer segments (30s), suggesting it captures both fine-scale spectral features and longer-term dynamics not typically considered in routine EEG interpretation. These findings warrant further investigation to better understand the clinical significance of these patterns.

Integrating DL models like DeepEpilepsy in the clinical workflow could enhance clinical decision-making by the increasing the information available in case of diagnostic uncertainty. However, this must be balanced against the risks of false positive predictions. While IEDs showed perfect specificity in our dataset, DeepEpilepsy's improved sensitivity comes at the cost of lower specificity, which is particularly concerning given the significant impact of an incorrect epilepsy diagnosis (unnecessary medications, driving restrictions, and psychosocial consequences) [12], [140].

Therefore, we envision DeepEpilepsy as a decision support tool rather than a diagnostic test. A positive prediction by the model in a patient with neurological events of uncertain significance and negative workup (no IEDs on EEG, no epileptogenic lesion on MRI) could increase the suspicion of epilepsy, prompting to more frequent follow-ups or repeat EEGs. Conversely, a patient with a low pre-test probability of epilepsy, absence of IEDs and a negative DL prediction could reduce clinical suspicion. Most likely, combined with advances in other domains such as text processing, imaging and genetics [210], [211], [212], the automated EEG analysis will lead to a more comprehensive phenotyping of these patients and potentially lead to quantifying the seizure likelihood. This could also improve clinical trials in epilepsy, which are currently limited by self-reported and unreliable outcome measures [49], [51].

This study has limitations. Our data comes from a single center, and although routine EEG recording is standardized, variability in hardware, software, and technique may affect generalizability. Additionally, at our center, patients with a first unprovoked seizure presenting at the emergency department generally undergo their EEG there and not as outpatient, limiting their inclusion in our cohort. Another limitation is the use of the EEG report as a measure of whether an EEG contains IEDs, which could be biased as EEG readers are not blinded to the diagnosis. However, for patients which were “undiagnosed” at the time of the EEG, the limitation does not apply. Finally, subgroup analyses were limited by the relatively small sample size.

In conclusion, this study demonstrates that DeepEpilepsy, a Transformer model, could identify epilepsy on routine EEG independently of IEDs. The DL algorithm alone had an AUROC of 0.76, surpassing previously proposed methods, which was increased to 0.83 when combined with IEDs. Several questions remain such as the exact nature of brain dynamics captured by DeepEpilepsy, the optimal sample sizes for training the model, and the true clinical impact of this increased diagnostic yield in specific clinical settings.

5.5.1 Code and Data Availability

The code for the study will be available upon publication at the following address: https://gitlab.com/chum-epilepsy/dl_epilepsy_reeg. Anonymized data will be made available to qualified investigators upon reasonable request, conditional to the approval by our REB. The STARD checklist is provided as Supplementary material.

5.5.2 Acknowledgements

We would like to acknowledge the work of the CHUM EEG technologists for their contribution to the recording of the EEGs. We would also like to thank Manon Robert and Véronique Cloutier for their help regarding the access to the EEG data and the submission to the ethics review board.

5.5.3 Funding

EL is supported by a scholarship from the Canadian Institutes of Health Research (CIHR) and the Fonds de Recherche du Québec–Santé (FRQS). DKN and FL are supported by the Canada Research Chairs Program, the CIHR, and Natural Sciences and Engineering Research Council of Canada. DKN reports unrestricted educational grants from UCB and Eisai, and research grants for investigator-initiated studies from UCB and Eisai. EBA is supported by the Institute for Data Valorization (IVADO, 51628), the CHUM research center (51616), the Brain Canada Foundation (76097), and the FRQS.

5.5.4 Competing interests

None of the authors declare any conflict of interest. The funding sources were not involved in study design, data collection, analysis, redaction, nor decision to submit this paper for publication.

5.6 Supplementary material

5.6.1 eMethod 1: Automated processing of EEG and classification

Deep Learning: We implemented two DL approaches. First, we adapted to EEG data the ConvNeXt model, a deep CNN analog to the ResNet, selected for its robust performance in computer vision tasks.[116] Second, we implemented a novel model coined *DeepEpilepsy*, based on the Vision Transformer (ViT) architecture, a Transformer model that takes images as input and that outputs class probabilities.[130] DeepEpilepsy uses a three-layer convolutional tokenizer plus a bottleneck convolution, restricting the complexity of the model and allowing to capture multi-scale features (Error! Reference source not found.), akin to the Compact Convolutional Transformer.[132] We also tested a tokenizer with non-overlapping linear patch embedding proper to the original ViT model.[130]

For all DL models, EEGs were segmented into overlapping 10- or 30-second segments and scaled so that each channel had a mean of zero and standard error of one. These scaled segments were used as input into the DL models. To enhance model generalization, we applied a random data augmentation algorithm during training, similar to the RandAugm algorithm.[134] For each EEG segment, one augmentation was drawn randomly from a set of transformations, which included filtering (band-pass, low-pass, high-pass), masking (channel, time), and adding noise (**eFigure 1**). These augmentations were applied with a 50% probability. The intensity of the augmentation (e.g., filter frequency, noise level, mask length) was also randomized and controlled by a hyperparameter M. Based on initial experiments on the training and validation data, we set M = 8.

We performed a Bayesian hyperparameter search on the training and validation set to select four different configurations for the ViT (one of which was DeepEpilepsy) (**eTable 1**) and three for ConvNeXt (**eTable 2**). We also investigated different learning rates, weight decay, and batch size values. The final models were trained on the entire training and validation set. The optimization hyperparameters and model specifications are described in **eTable 4**.

ShallowConvNet: We reimplemented the ShallowConvNet model following the configuration outlined in Schirrmeister et al.[125] However, after conducting a hyperparameter search on the training and validation set, we identified a more optimal configuration specific to our dataset, which we used for testing (**eTable 3**). The EEG segmentation and standardization were consistent with the other DL models. Similarly, we optimized training hyperparameters (learning rates, weight

decay, and batch size values) through a Bayesian search on the training and validation set (**eTable 4**).

EEG markers: We followed the methodology described in Lemoine et al.,[198] selecting only the best-performing markers and testing both 10- and 30-second segments. EEGs were segmented at pre-specified time points (every change of montage, every 15s during hyperventilation, every 15s for two minutes post-hyperventilation, every photic stimulation frequency, and every eye closure or opening). We applied an automated artifact detection/rejection algorithm (*AutoReject*)[159] and extracted the following markers: fuzzy entropy, line length, correlation dimension, band power, and peak alpha. Band power was calculated using a multitaper method, with integrals estimated using Simpson's method (frequency ranges: 100–75 Hz, 75–40 Hz, 40–20 Hz, 20–13 Hz, 13–10 Hz, 10–8 Hz, 8–6 Hz, 6–4 Hz, 4–2 Hz, and 2–1 Hz). For nonlinear features (fuzzy entropy, line length, and correlation dimension), the *Sym5* wavelet was used with six decomposition levels (with frequency ranges: 100–50 Hz, 50–25 Hz, 25–12.5 Hz, 12.5–6.25 Hz, 6.25–3.125 Hz, and 3.125–1.56 Hz).[160] One value was extracted per marker, EEG, segment, channel, and frequency band. Missing values were imputed using multivariate iterative imputation. Markers were used as input features for an L1-regularized boosted-trees classifier (LightGBM). The optimal hyperparameters for the classifier were selected via Bayesian optimisation using a 5-fold cross-validation on the training and validation set.

5.6.2 eMethod 2: Interpretability

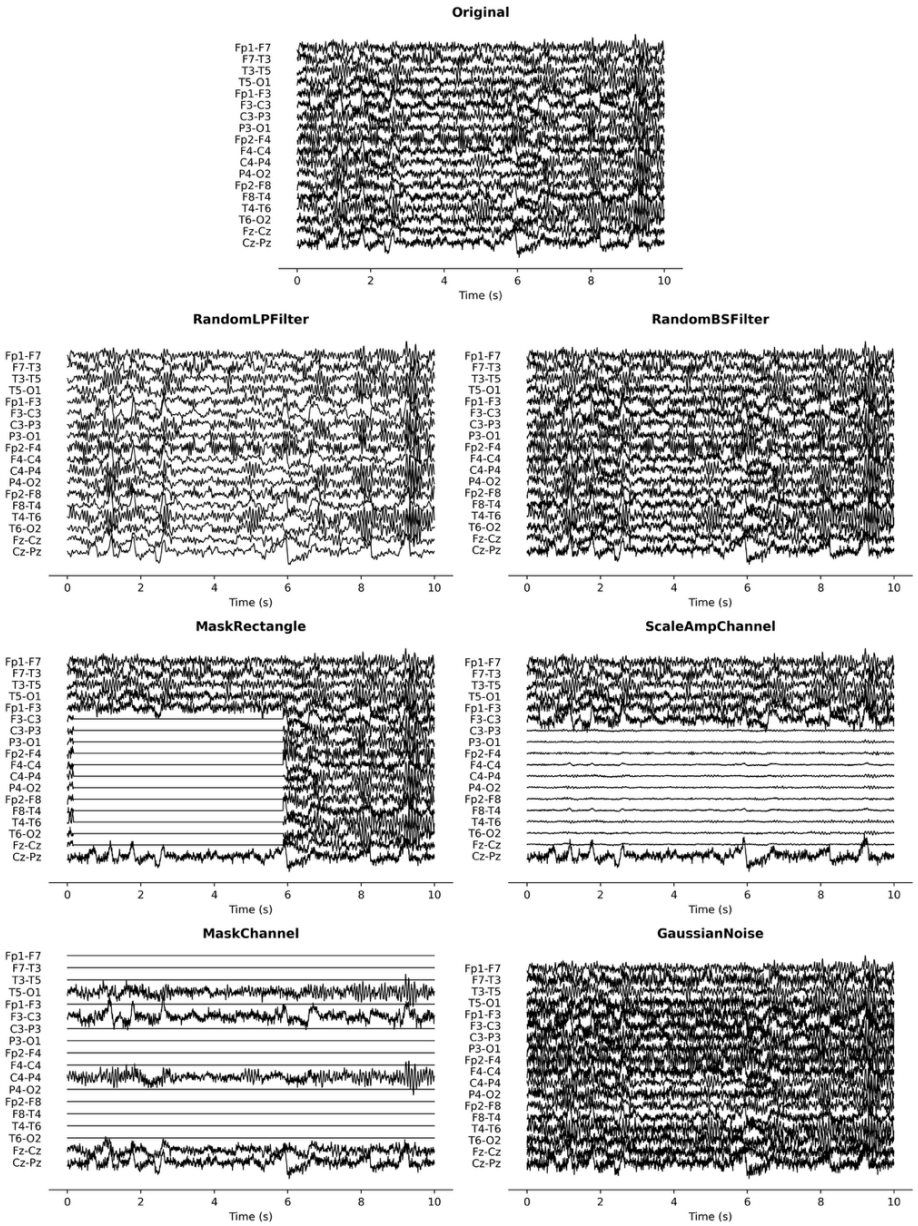
We performed an exploratory analysis of the embeddings learned by DeepEpilepsy and ShallowConvNet to better understand which patterns were captured by these DL models from the raw EEG data. This analysis is distinct from the LightGBM benchmark model, which used pre-specified EEG features for classification.

For the embedding analysis, thirty-second segments from the testing set were processed through each DL model, and their embeddings (internal representation before the classification layer) were extracted. A clustering algorithm was then used to group the embeddings into 12 distinct clusters. To understand what patterns these clusters represented, we computed two traditional EEG features (band power and entropy) from the original EEG segments and analyzed how these features were distributed across the clusters. These features were computed using the same methods and

frequency ranges described in the section **Automated processing of EEG and classification: EEG markers**.

To test for heterogeneity between clusters, we applied an analysis of variance (Krusper-Wallis test) at each frequency band. We then compared the F-score between both models and between frequency bands to identify which frequency ranges showed the greatest variation between clusters, suggesting these were important patterns learned by each model.

5.6.3 eFigure 1: Data augmentations used by the RandAugm algorithm



eFigure 5.1: Data augmentations used by the RandAugm algorithm, alongside the original EEG sample. RandomLPFilter: Low-pass filter, with random cut-off frequency. RandomBSFilter: band stop filter with randomly chosen frequency window. MaskRectangle: masking of data points contiguous in both time and space. ScaleAmpChannel: Random scaling of channels. MaskChannel: masking of all data points in randomly selected channels. GaussianNoise: Addition of gaussian noise with a random intensity. The intensity of the augmentations is scaled according to a hyperparameter M. For example, higher values of M result, on average, in lower values of cutoff frequencies for LPFilter, larger mask area for MaskRectangle, and a larger number of channels affected by ScaleAmpChannels as well as a higher amplitude of scaling.

5.6.4 eTable 1–4: Deep learning hyperparameters for the final model configurations

eTable 1: Model configurations for the Vision Transformer (ViT) models

Model	patch size	tokenizer	tokenizer: layers	hidden dim	layers	heads	MLP size	dropout	attention dropout	params (M)
ViT1d, linear, small	200	Linear	1	128	2	2	128	0.25	0.25	0.7
ViT1d, linear, large	50	Linear	1	512	6	8	512	0.25	0.25	10.0
ViT1d, Conv, small	200	Convolution	3	128	2	2	128	0.25	0.25	0.4
DeepEpilepsy: ViT1d, Conv, large	50	Convolution	3	512	6	8	512	0.25	0.25	10.6

eTable 2: Model configurations for the ConvNext models

Model	blocks	channels	stem: downsampling scale	drop path rate	params (M)
ConvNeXt, small	1, 1, 3, 1	16, 32, 64, 128	4	0.1	0.3
ConvNeXt, large	2, 2, 6, 2	32, 64, 128, 256	2	0.1	2.0
ConvNeXt, huge	3, 3, 9, 3	64, 128, 256, 512	2	0.1	11.9

eTable 3: Model configuration for the ShallowConvNet model

Model	kernel size (stride)	space conv channels	time conv channels	max pool window	dropout	params (M)
ShallowConvNet	16 (1)	64	128	80	0.25	0.040

eTable 4: Optimization parameters for all neural networks

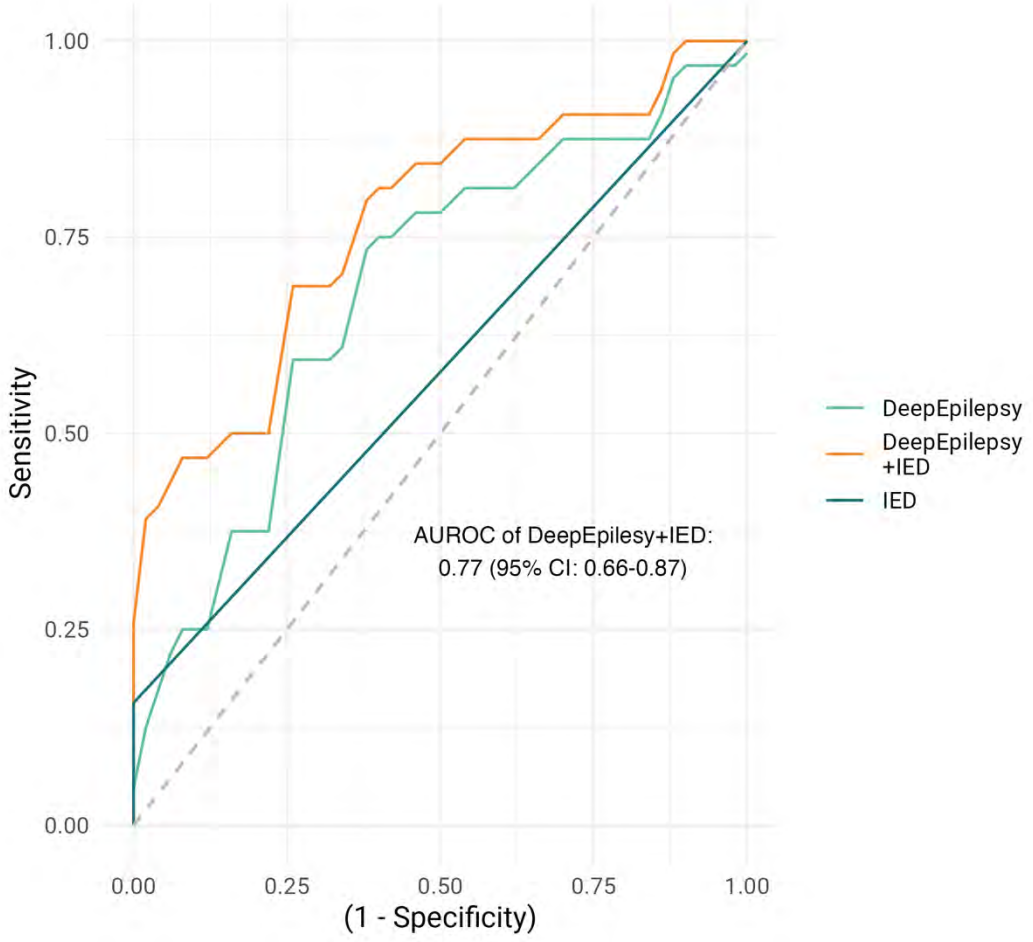
Parameter	value
Optimizer	AdamW
Base learning rate	1.0e-5
Weight decay	0.05
Optimizer momentum	β1, β2=0.9, 0.999
Batch size	512
Training epochs	30
Learning rate schedule	Cosine decay
Warmup iterations	1000

Warmup schedule	Linear
RandAugm M	7
Gradient clipping	1.0 (ViT only)

5.6.5 eTable 5: Clinical characteristics of the "undiagnosed" subgroup of the testing cohort

	Epilepsy	**No Epilepsy**
Number of patients	28	47
Sex = woman (%)	19 (67.9)	26 (55.3)
Age (median [IQR])	41.00 [34.75, 58.25]	60.00 [50.50, 71.00]
Total follow-up after EEG in weeks (median [IQR])	119.00 [95.00, 134.50]	62.00 [17.00, 102.00]
Epilepsy type (%)		
Focal	23 (82.1)	–
Generalized	3 (10.7)	–
Unknown	2 (7.1)	–
Age of epilepsy onset (median [IQR])	37.00 [23.25, 52.00]	–
Seizure recurrence after EEG (%)	17 (60.7)	–
Number of days since last seizure (median [IQR])	87.50 [33.00, 164.00]	–
Number of epilepsy risk factors (median [IQR])	2.00 [1.00, 4.00]	1.00 [0.00, 3.00]
History of epilepsy surgery (%)	0 (0)	–
Number of ASM (%)		
0	9 (32.1)	42 (89.4)
1	12 (42.9)	5 (10.6)
2	5 (17.9)	0 (0.0)
3	2 (7.1)	0 (0.0)
4	0 (0.0)	0 (0.0)
5	0 (0.0)	0 (0.0)
Focal lesion on brain imaging (%)	10 (35.7)	10 (21.3)
Sleep deprived EEG (%)	9 (32.1)	8 (17.0)
IED (%)		
Absence	12 (42.9)	46 (97.9)
Presence	10 (35.7)	0 (0.0)
Uncertain	6 (21.4)	1 (2.1)
Abnormal slowing on EEG (%)	10 (35.7)	10 (21.3)

5.6.6 eFigure 2: Performance of DeepEpilepsy compared to interictal epileptiform discharges on the undiagnosed subgroup

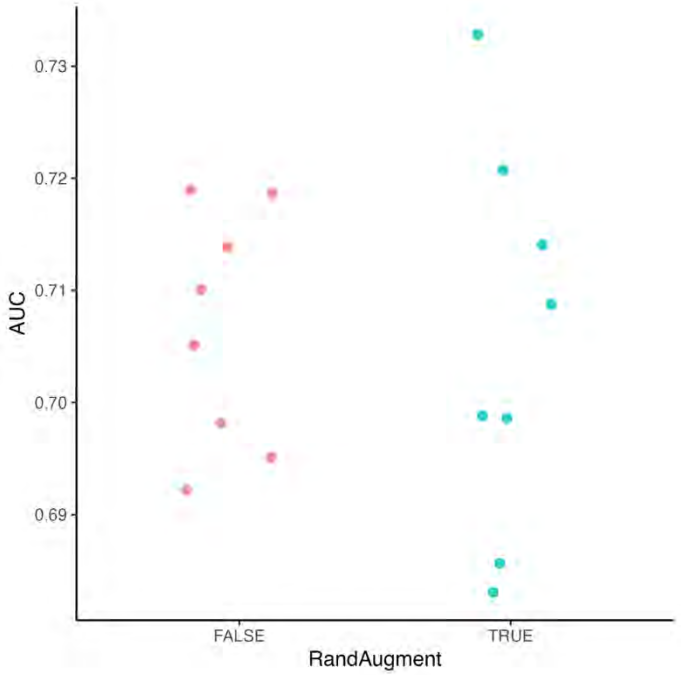


eFigure 2: ROC curves for IEDs only, DeepEpilepsy, and DeepEpilepsy combined with IEDs in the subgroup of patients not diagnosed with epilepsy at the time of the EEG (n = 77). AUROC: Area under the receiver operating characteristic curve; IED: interictal epileptiform discharges.

5.6.7 Segment duration and RandAugment Analysis

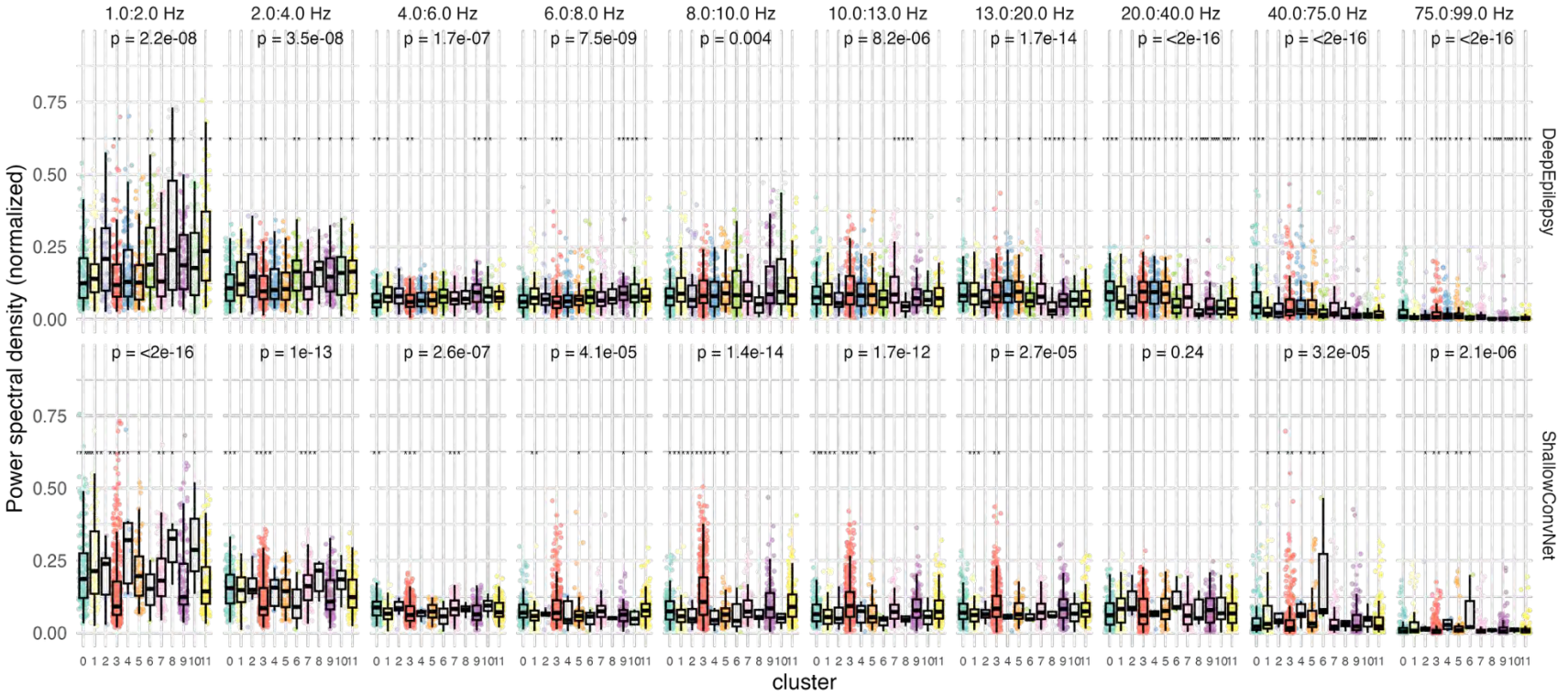
eTable 6: Effect of segment duration on DeepEpilepsy's performances

Segment duration (s)	AUROC (95% CI): RandAugment	AUROC (95% CI): No RandAugment
5	0.721 (0.639–0.804)	0.711 (0.625–0.792)
10	0.713 (0.632–0.789)	0.691 (0.603–0.772)
30	**0.746 (0.663–0.821)**	**0.679 (0.589–0.763)**
45	0.733 (0.644–0.815)	0.717 (0.627–0.798)
60	0.716 (0.634–0.796)	0.705 (0.615–0.787)



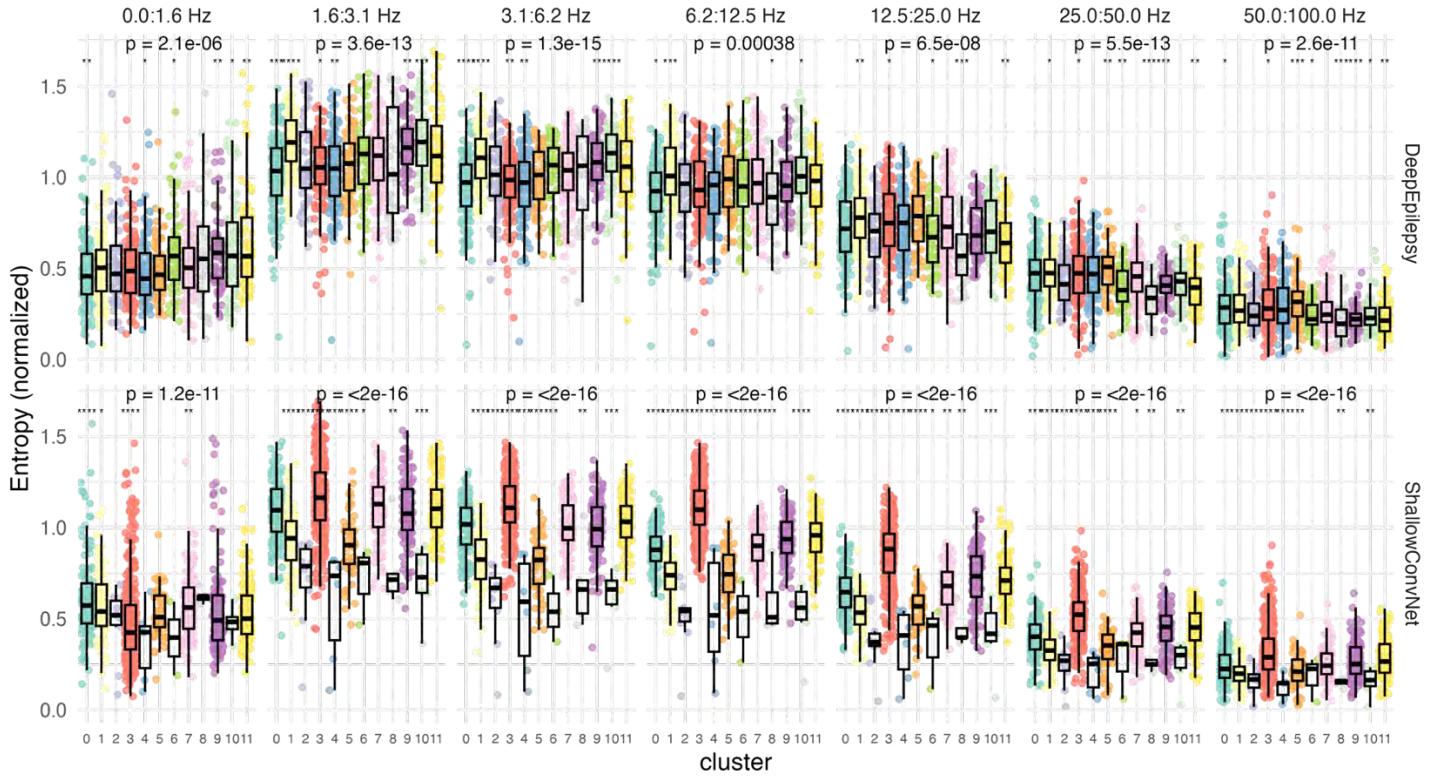
eFigure 3: Effect of RandAugment on DeepEpilepsy's performances. Each point represents the AUROC achieved by DeepEpilepsy on the testing set after independent training runs (20 epochs each) with (blue) or without (orange) RandAugment data augmentation. AUC: area under the curve.

5.6.8 eFigure 3: Power spectrum density of EEG segments clustered according to their latent representations using DeepEpilepsy vs. ShallowConvNet



eFigure 4: Power spectrum density of 30s EEG segments clustered according to their latent representations using DeepEpilepsy vs. ShallowConvNet. Each point represent the normalized power spectrum density for individual EEG segments of 30s at different frequency band for both models. Each EEG segments was processed through either the trained ShallowConvNet (bottom) or the trained DeepEpilepsy (top) to generate a latent vector. The latent vectors were then clustered using K-means clustering (K=12). The power in each band was calculated for the input segment (1 Hz:2 Hz, 2 Hz:4 Hz, etc.) and plotted on the y-axis. A statistical analysis of inter-cluster variance was perform in each frequency band using the Krusper-Wallis test (*p*-values at the top of each facet, ***n*** = 1 024 segment per test). A lower p-value correspond to a larger heterogeneity between clusters in that frequency bands.

5.6.9 eFigure 5: Entropy of EEG segments clustered according to their latent representations using DeepEpilepsy vs. ShallowConvNet



eFigure 5: Entropy of EEG segments clustered according to their latent representations using DeepEpilepsy vs. ShallowConvNet. Each point represent the normalized entropy for individual EEG segments of 30s at different frequency band for both models. Each EEG segments was processed through either the trained ShallowConvNet (bottom) or the trained DeepEpilepsy (top) to generate a latent vector. The latent vectors were then clustered using K-means clustering (K=12). The entropy in each band was calculated for the input segment (1 Hz:2 Hz, 2 Hz:4 Hz, etc.) and plotted on the y-axis. The fuzzy entropy algorithm was used with parameters m=2 and r=0.2. A statistical analysis of inter-cluster variance was perform in each frequency band using the Krusper-Wallis test (p-values at the top of each facet, n = 1 024 segment per test). A lower p-value correspond to a larger heterogeneity between clusters in that frequency bands.

CHAPITRE 6 UN MODÈLE DE SURVIE PROFOND POUR PRÉDIRE LE RISQUE DE CRISE APRÈS L'EEG DE ROUTINE

Le chapitre précédent a démontré la capacité de l'apprentissage profond à détecter des marqueurs d'épilepsie à l'EEG avec une performance supérieure aux DÉI et aux marqueurs computationnels. Cependant, le modèle DeepEpilepsy présente trois limitations en lien le format de ses prédictions. Premièrement, sa prédiction binaire (épilepsie vs. absence d'épilepsie) ne reflète pas la variabilité du risque de crise, notamment chez les patients bien contrôlés sous traitement. Deuxièmement, l'hétérogénéité des durées de suivi n'est pas prise en compte dans l'apprentissage, ce qui peut biaiser le modèle en faveur des patients avec un suivi plus court. Troisièmement, une quantification dynamique du risque de crise serait cliniquement plus pertinente qu'une classification binaire, permettant une meilleure personnalisation de la prise en charge.

Ce chapitre complémentaire présente les travaux d'un troisième article en préparation qui adresse ces limitations en développant à partir de DeepEpilepsy un modèle de survie capable de prédire le risque de crise à travers le temps. De plus, des modifications substantielles à l'architecture et au prétraitement des données améliorent la robustesse et l'interprétabilité du modèle. À terme, le projet inclura une validation multicentrique du modèle, dont je présente le plan à la fin du chapitre. Les résultats de ce travail ont fait l'objet de présentations orales, notamment à l'*Eastern Association of Electroencephalographers* (Boston, 2025; Prix de Congrès *Burnham Fellowship*) et *ICTALS* (Montréal, 2025), ainsi que par affiche à l'*American Academy of Neurology Annual Meeting* (San Diego, 2025).

6.1 Introduction

L'épilepsie est définie cliniquement par un risque accru de crises [2]. L'évaluation de ce risque nécessite une approche multimodale intégrant la sémiologie des épisodes suspects, les facteurs de risque cliniques, la neuroimagerie et l'électroencéphalogramme (EEG). La présence d'anomalies épileptiformes à l'EEG est particulièrement informative, prédisant un risque de récidive 1.5–3 fois plus élevé dans plusieurs situations cliniques [8], [30], [31], [32].

Malheureusement, l'EEG présente plusieurs limitations. Sa sensibilité est faible, avec seulement 29–55% des patients épileptiques qui présenteront des anomalies épileptiformes sur un EEG de 30 à 60 minutes [8], [10], [28]. Son interprétation requiert une expertise surspécialisée, et même entre

experts, l'accord inter-observateur pour les anomalies épileptiformes est au plus modéré [37], [213]. L'analyse computationnelle de l'EEG, particulièrement par apprentissage profond, offre une alternative prometteuse en extrayant automatiquement des biomarqueurs quantitatifs qui représentent les interactions complexes entre fréquences et régions cérébrales à différentes échelles temporelles [74], [103], [104], [214].

Les modèles actuels d'analyse automatisée de l'EEG se concentrent principalement sur la classification du diagnostic [74], [100], [101], [214] ou la prédiction de récidive à des horizons temporels fixes [77], [198]. Bien que la classification puisse théoriquement prendre en compte plusieurs classes de risque, cette approche ne reflète pas pleinement la nature temporelle du risque épileptique. L'épilepsie est fondamentalement définie par un risque de crise qui évolue dans le temps [2], [3].

L'analyse de survie est un outil épidémiologie qui permet d'estimer la fonction de survie $S(t)$ qui décrit l'évolution d'une maladie après un facteur de risque ou un traitement [215]. Les modèles de survie sont particulièrement adaptés à l'épilepsie pour modéliser le risque de crise à travers le temps [45], [216], [217], [218]. Ces modèles statistiques sont généralement limités par la complexité des variables d'entrée, souvent constituées de quelques variables cliniques [215], [216], [217], [218]. Cependant, le couplage entre l'apprentissage profond et les modèles de survie permet de combiner une entrée complexe avec la prédiction du risque dans le temps [219], [220]. Quelques études ont démontré l'utilité de ces modèles de survie profond en oncologie [220], [221], [222], [223] et en soins intensifs [224], mais leur applicabilité en épilepsie est inconnue.

Cette étude propose donc pour la première fois un modèle de survie profond en épilepsie, EEGSurvNet, qui analyse le signal EEG pour prédire le délai jusqu'à la prochaine crise sur un horizon de deux ans. La performance du modèle est comparée à un modèle de survie traditionnel avec les prédicteurs cliniques standards, incluant la présence d'anomalies épileptiformes. Au-delà de la prédiction, nous explorons l'interprétabilité du modèle pour identifier les caractéristiques du signal EEG associées au risque de crise ainsi que les paramètres qui optimisent sa généralisation.

6.2 Méthodes

6.2.1 Design et population

Il s'agit d'une étude rétrospective chez une cohorte de patients consécutifs ayant obtenu un EEG de routine au Centre hospitalier de l'Université de Montréal (CHUM), au Canada. Tous les patients qui ont eu un EEG au service de neurophysiologie entre le 1er janvier 2018 en le 31 décembre 2019 étaient inclus. L'EEG de routine inclut les EEG de 30 à 60 minutes avec et sans déprivation de sommeil effectué chez des patients en centre ambulatoire. Les critères d'exclusion sont l'absence de suivi après l'EEG, un diagnostic d'épilepsie incertain à la fin de la période du suivi, ou la présence de crise à l'EEG. Les dossiers des patients ont été révisés par un résident en neurologie/neurologue (EL) ainsi que trois étudiants en neurosciences selon un protocole pré-spécifié. Les données extraites incluaient l'âge, le sexe, les comorbidités, la présence de facteurs de risque d'épilepsie, la présence d'anomalies à l'imagerie neurologique (IRM ou scan) et le nombre de médicaments anticrises. Pour chaque visite, le nombre de crises depuis la visite précédente est extraite. Le diagnostic d'épilepsie est déterminé selon les critères de la Ligue Internationale contre l'Épilepsie [2] à partir de la dernière note disponible du neurologue traitant. Les rapports d'EEG ont été révisés à la recherche d'anomalies épileptiformes ou de ralentissement anormal. Les données cliniques sont stockées dans une base de données REDCap située sur les serveurs sécurisés du CRCHUM.

Les EEG enregistrés avant septembre 2019 constituent l'échantillon d'entraînement et validation, et ceux après septembre 2019, l'échantillon test (Figure 6.1). Les patients ayant eu un EEG à la fois avant et après septembre 2019 sont exclus de l'échantillon test.

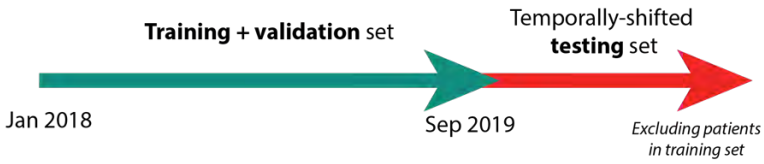


Figure 6.1 : Illustration de la séparation temporelle des échantillons d'entraînement/validation et de test.

6.2.2 Issue primaire

L'issue primaire est le délai (en jours) jusqu'à la prochaine crise épileptique après l'EEG, extraite des notes médicales de suivi. L'issue englobe tout type de crises épileptiques, incluant les crises focales sans altération de l'état de conscience et les myoclonies, mais pas les crises non-épileptiques. Pour environ 25% des patients, la date exacte n'était pas rapportée; dans ce cas, nous effectuons une interpolation linéaire en fonction de la fréquence rapportée de crises, assumant une distribution uniforme des crises pour la période. Les patients n'ayant pas présenté de crises pendant leur suivi sont considérés comme censurés à la date de la dernière visite documentée.

6.2.3 Prédicteurs

EEG

Les EEG sont enregistrés sur un appareil Nihon-Kohden selon un protocole standard en accord avec les lignes directrices canadiennes [225]. Les EEG d'éveil, d'une durée de 20 à 30 minutes, sont enregistrés avec une fréquence d'échantillonnage de 200 Hz via 19 électrodes respectant l'arrangement 10–20 [226]. Ils incluent deux périodes de 90s d'hyperventilation (excepté chez les patients de plus de 80 ans, non-coopératif, ou avec une contre-indication médicale) et stimulation photique de 4 à 22 Hz. Les EEG de sommeil durent 60 minutes et comportent les mêmes procédures d'activation. Les EEG sont enregistrés avec une référence A1-A2, convertis en format EDF, puis stockés sur les serveurs sécurisés du CRCHUM en suivant le standard BIDS [227].

Modèle de survie profond

Pour cette étude, nous avons développé un modèle de survie profond pour l'EEG appelé EEGSurvNet. EEGSurvNet est bâti sur DeepEpilepsy, un *Vision Transformer* (ViT) qui prend en entrée des segments d'EEG multi-canaux de 10 ou 30s [214]. Plusieurs améliorations ont été apportées à l'architecture originale dans le but d'augmenter ses performances, sa robustesse aux

artéfacts et données externes et son interprétabilité. Ces améliorations ont été testées itérativement sur l'ensemble d'entraînement/validation ainsi que sur un jeu de données externes, le *Temple University Abnormal EEG corpus* (v3.0.1) [123].

La première amélioration consiste à allonger la durée des segments analysés de 30 et de 60 secondes. Au-delà de 60 secondes, nous notons un plafonnement des performances au prix de ressources computationnelles très élevées. La deuxième amélioration consiste à appliquer un filtre passe-bande à 60 Hz pour éliminer la contribution du bruit de courant alternatif et de potentiellement rendre le modèle robuste aux EEG enregistrés dans différents pays. La troisième amélioration concerne le format d'entrée des données: les EEG sont convertis en spectrogrammes en utilisant la transformée avec ondelettes de Morlet (nombre de cycles: 7, nombre de bandes de fréquences : 24). Cette amélioration permet surtout d'améliorer l'interprétabilité. La résolution temporelle est diminuée par un facteur de 4 pour limiter la taille des données d'entrées. Le tokeniseur du modèle a également été modifié pour traiter ces spectrogrammes, utilisant maintenant des convolutions en 2D. La version finale comporte 27M de paramètres (Table 6.1).

Tableau 6.1 : Architecture d'EEGSurvNet

Hyperparamètre	**Valeur**
Taille de patch	0.2s
Tokenizeur : type	Convolution
Tokenizeur : nombre de couches	3
Tokenizeur : taille des filtres (fréquence, temps)	(3, 11)
Tokenizeur : dimensionalité des filtres	(256, 362, 512)
Transformeur : dimensions cachées	512
Transformeur : nombre de couches	8
Transformeur : nombre de têtes	8
MLP : dimensionalité	1024
Dropout	0.2
Params (M)	27

L'adaptation du modèle à l'analyse de survie repose sur une approche par discrétisation temporelle (*discrete-time analysis*; Figure 6.2) [228], [229]. La durée du suivi est divisée en 7 périodes espacées logarithmiquement entre 28 jours et 2 ans (la dernière période correspondant à > 2 ans). Le modèle prédit le hasard (risque instantané) $h(t)$ de crise pour chaque période via une fonction logistique appliquée aux sorties du réseau (ϕ):

$$h_s(t) = \sigma\big(\phi_s(t)\big) = \frac{1}{1 + e^{-\phi_s(t)}},$$

où s représente le segment d'EEG analysé. Pour l'entraînement, chaque patient est représenté par deux valeurs: l'indice t_i de la période où survient une crise ou la dernière période de suivi disponible et un indicateur d'évènement δ distinguant les crises (1) des censures (0). La fonction de loss correspond à la log-vraisemblance négative (*negative log-likelihood*) du modèle de hasard:

$$\mathcal{L} = -\sum_{i=1}^{n}\left[\delta_i \log\big(h_s(t_i)\big) + \sum_{j=1}^{t_i-\delta_i} \log\big(1-h_s(j)\big)\right].$$

L'entraînement s'effectue sur des segments de 60 secondes avec chevauchement (1s), totalisant environ 1 800 segments par EEG de 30 minutes. Le modèle est entraîné sur deux GPUs A6000 durant 25 époques, chaque époque comptant environ 1.7M de segments. Une augmentation des données est appliquée selon la méthode TrivialAugment [230] avec une probabilité de 90%, ajoutant soit un bruit gaussien aléatoire, soit un masque aléatoire, soit une égalisation aléatoire. L'optimisation utilise un *cosine decay* du taux d'apprentissage après une période de réchauffement d'une époque. Les hyperparamètres (taux d'apprentissage maximal, probabilité de TrivialAugment, probabilité de *dropout* et *weight decay*) ont été optimisés sur l'ensemble de validation, avec les valeurs finales présentées dans le Table 6.2. La libraire *PyTorch* (version 2.6.0) est utilisée pour entraîner les modèles profonds.

Tableau 6.2 : Paramètres d'apprentissage pour EEGSurvNet

Paramètre	**Valeur**
Optimisateur	AdamW
Taux d'apprentissage maximal	5.0 x10^{-5}
Weight decay	0.05
Momentum	β1, β2=0.9, 0.999
Taille des lots effective	512
Nombre de GPUs	2
Époques	25
Horaire du taux d'apprentissage	*Cosine decay*
Époque de réchauffement	1
Horaire du réchauffement	Linéaire
Probabilité d'augmentation	0.9
Gradient clipping	Aucun

Pour l'inférence sur un EEG complet, nous effectuons deux niveaux d'agrégation. Premièrement, les hasards sont moyennés sur tous les segments:

$$\bar{h}(t) = \frac{1}{N}\sum_{s=1}^{N} h_s(t)\,.$$

Puis, pour régulariser la sortie, nous calculons un hasard constant h^* en moyennant les hasards sur toutes les périodes:

$$h^* = \frac{1}{T}\sum_{t=1}^{T} \bar{h}(t)\,.$$

À partir de ce hasard constant, nous calculons la fonction de survie:

$$S(t) = (1 - h^*)^t\,.$$

Le score de risque global R est égal au hasard constant h^* et permet une comparaison avec les modèles de risque proportionnel comme Cox.

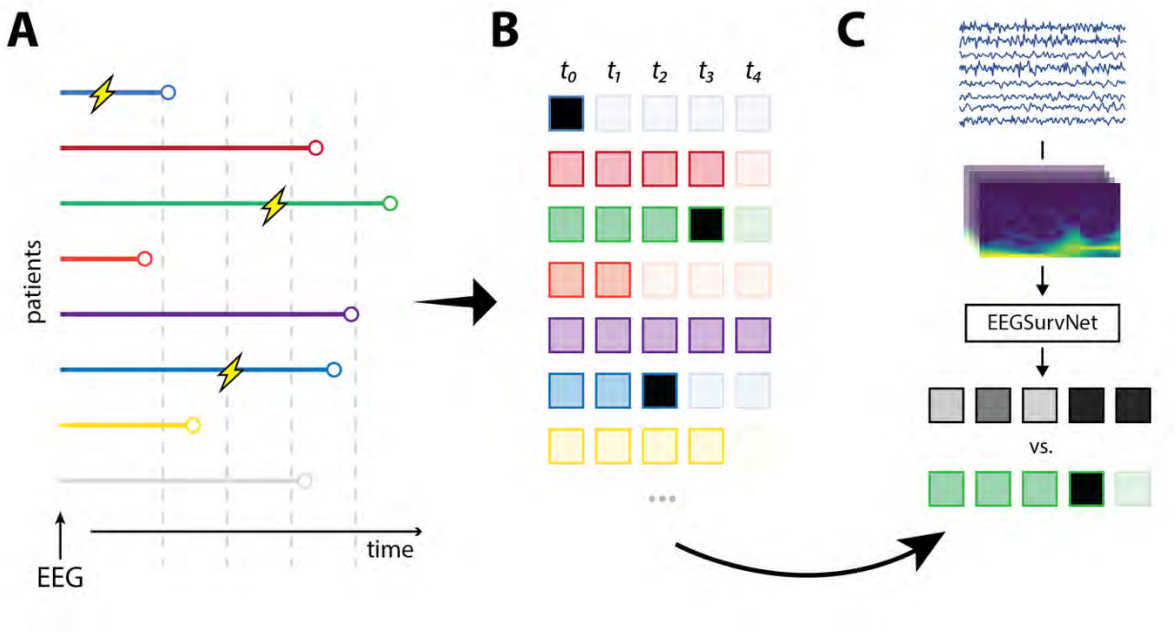


Figure 6.2 : Analyse de survie par discrétisation temporelle. A : Les séries temporelles originales pour chaque patient, certains subissant une crise lors du suivi (éclair jaune). B : La discrétisation temporelle du suivi de chaque patient, où chaque période temps est représentée par une valeur binaire (0: pas de crise pendant la période; 1: crise pendant la période). C : Le modèle EEGSurvNet prend en entrée les signaux EEG transformés en spectrogramme, puis prédit une valeur de risque h_t pour chaque période. L'optimisation utilise la log-vraisemblance négative du modèle de hasard. Les prédictions sont par la suite agrégées à travers les segments puis les périodes pour obtenir un score de risque global.

Modèle clinique Cox

À des fins de comparaison, nous avons développé un modèle de risque proportionnel de Cox utilisant les prédicteurs traditionnels de risque de crise: âge, sexe, histoire familiale d'épilepsie, présence de lésion focale en neuroimagerie et présence de pointes épileptiformes à l'EEG. Sur la base de travaux antérieurs [8], nous avons inclus des termes d'interaction entre la présence de pointes et les autres prédicteurs. Le modèle Cox prédit un risque proportionnel selon l'équation:

$$h(t|X) = h_0(t)\exp(\beta^T X)\ ,$$

où $h_0(t)$ est le risque de base, X le vecteur des prédicteurs et β les coefficients estimés. Le risque total est calculé en combinant le risque de base (estimé sur l'échantillon d'entraînement) avec le risque proportionnel, produisant une mesure analogue au score de risque global du modèle profond.

Combinaison entre le modèle clinique et le modèle profond

Une troisième approche combine les prédictions des modèles profond et clinique en multipliant leurs scores de risque respectifs:

$$R_{\text{combiné}} = R_{\text{EEGSurvNet}} \times \exp(\beta^T X)$$

Cette approche simple présente l'avantage d'une grande interprétabilité: les deux modèles restants indépendants, ils peuvent être analysés séparément sans avoir à considérer les interactions complexes entre données EEG et cliniques. Nous avons également exploré le développement d'un modèle profond multimodal intégrant directement les données cliniques et EEG. Cependant, malgré une optimisation extensive des hyperparamètres, ce modèle tendait systématiquement vers un surapprentissage des données cliniques en négligeant l'information EEG. Le développement d'une architecture multimodale efficace reste un axe de recherche prometteur à approfondir.

6.2.4 Taille d'échantillon et analyse de puissance

L'analyse de puissance a été réalisée avec la librairie R "powerSurvEpi" selon la méthode de Schmoor et al. [231] Nous avons estimé la taille d'échantillon requise pour un modèle de Cox avec un prédicteur binaire, représentant la présence ou l'absence d'anomalie à l'EEG. L'analyse assume un taux d'événements (crises) de 0.5, un rapport de risque (*hazards ratio*) de 2.0, un seuil de signification α de 0.05 et une puissance cible de 0.8. Selon ces paramètres, un échantillon de 131 patients est nécessaire pour détecter l'effet d'intérêt.

6.2.5 Analyses

L'évaluation des modèles repose sur deux aspects complémentaires: la discrimination et la calibration. La discrimination est évaluée par une extension des courbes ROC adaptée aux données de survie, appelée *cumulative/dynamic time-dependant ROC curve* [232], [233]. La sensibilité et la spécificité sont définies comme des mesures dépendantes du temps; les cas cumulatifs comprennent tous les individus ayant présenté un événement jusqu'au temps t, tandis que les contrôles dynamiques sont ceux qui présentent un événement après t. L'AUC cumulative/dynamique quantifie la capacité du modèle à distinguer les sujets qui auront une crise avant un temps donné de ceux qui en auront une après. Nous avons aussi appliqué une pondération par l'inverse de la probabilité de censure (IPCW) pour gérer le biais introduit par la censure [234]. Les poids IPCW sont estimés à partir de la distribution de censure de l'ensemble d'entraînement via l'estimateur de Kaplan-Meier, sous l'hypothèse d'une censure aléatoire indépendante des caractéristiques. Nous calculons l'AUROC à chaque période t en plus de l'AUROC intégrée sur deux ans (iAUROC). La calibration, elle, est évaluée par le score de Brier intégré sur deux ans, qui mesure l'exactitude des probabilités prédites [235]. Les intervalles de confiance à 95% sont calculés par *bootstrap* (1000 itérations).

Pour s'accorder avec la littérature des modèles prédictifs de récidive de crise [216], [236], [237], [238], [239], [240], nous avons aussi calculé les iAUROC et iBS sur 1 an, ainsi que l'index C de Harrell [241].

Nous avons aussi testé un modèle de référence *Baseline* qui prédit des probabilités centrées autour des probabilités de base observées dans l'ensemble d'entraînement. Spécifiquement, pour chaque patient de l'ensemble test, le modèle génère un score de risque tiré d'une distribution normale (μ = risque moyen d'entraînement, $\sigma = 0.2$). La performance des modèles est comparée à cette référence via le score de Brier intégré (iBS) calculé sur 2 ans, ainsi que le score de compétence de Brier (BSS) à chaque période. Le BSS quantifie l'amélioration relative de la prédiction par rapport au modèle *Baseline* [242]:

$$\mathrm{BSS}(t) = 1 - \frac{\mathrm{BS}_{\mathrm{model}}(t)}{\mathrm{BS}_{\mathrm{ref}}(t)},$$

où $\mathrm{BS}_{\mathrm{model}}$ et $\mathrm{BS}_{\mathrm{ref}}$ sont respectivement les scores de Brier du modèle évalué et du modèle de référence. Le BSS prend des valeurs entre $]-\infty, 1.0]$, une valeur positive indiquant une

performance supérieure à la référence. La calibration du modèle a aussi été évaluée visuellement par des courbes de calibration à différents horizons temporels. Pour chaque période d'intérêt, les patients sont regroupés selon leur probabilité prédite de crise (bins de 0.2), et le taux observé de crises dans chaque groupe est comparé à la prédiction moyenne.

Des analyses stratifiées ont été réalisées pour évaluer la performance dans des sous-populations cliniquement pertinentes: groupe d'âge (18–40, 40–60 et >60 ans), sexe, présence de lésion focale à l'imagerie et présence de ralentissement anormal à l'EEG et présence de décharges épileptiformes à l'EEG.

Pour interpréter le modèle, nous utilisons les valeurs Shapley estimées selon la méthode des gradients, représentant la contribution de chaque valeur d'entrée du modèle sur sa prédiction [243], [244]. Nous avons entraîné l'estimateur des valeurs Shapley sur 500 segments aléatoires de l'échantillon d'entraînement, puis extrait les valeurs des 50 segments avec les scores de risque les plus élevés et les 50 avec les scores les plus bas. Les valeurs sont ensuite intégrées sur le domaine temps-fréquence et sur le domaine spatial. La libraire Python *shap* a été utilisées pour cette analyse [243].

Une étude d'ablation systématique évalue l'impact de quatre paramètres clés du modèle: la durée du segment (30s vs. 60s), la résolution fréquentielle (16 vs. 32), la résolution temporelle (0.4s vs. 0.8s) et l'augmentation des données (présence vs. absence). Pour chaque configuration, le modèle est réentraîné sur l'ensemble d'entraînement avec des paramètres d'optimisation fixes (taux d'apprentissage, taille des lots, *scheduling*) et évalué sur l'échantillon test. En complément, le modèle DeepEpilepsy original [214] a été adapté à la prédiction de survie en modifiant sa couche de sortie pour générer les sept valeurs de risque temporel, puis réentraîné sur l'ensemble d'entraînement.

6.2.6 Approbation éthique et disponibilité du code source

Cette étude a reçu l'approbation du Comité d'éthique de la recherche du Centre de recherche du CHUM (Montréal, Canada, numéro de projet: 19.334). Le comité a accordé une dérogation au consentement éclairé en raison de l'absence d'intervention diagnostique/thérapeutique et du risque minimal pour les participants. Toutes les méthodes respectent l'Énoncé de politique des trois Conseils sur l'éthique de la recherche avec des êtres humains du Canada.

Le code source de l'étude sera disponible lors de la publication à l'adresse suivante: https://gitlab.com/chum-epilepsy/epi_surv.

6.3 Résultats

6.3.1 Caractéristiques de la population

Dans la période d'intérêt, 1 540 EEG ont été réalisés au CHUM chez 1 286 patients. Après exclusion, 1 014 EEG de 994 patients ont été inclus: 879 EEG de 786 patients dans l'échantillon d'entraînement et 135 EEG de 115 patients dans l'échantillon test. Le suivi médian de l'échantillon d'entraînement et de test étaient de 2.2 ans après l'EEG. Les caractéristiques cliniques détaillées des échantillon d'entraînement et de test sont présentées dans le Table 6.3.

Une crise est survenue après 295 des EEG d'entraînement (33.6%) et 40 des EEG de test (29.6%), avec un taux de survie sans crise à un an de 0.69% (95%CI: 0.66–0.73) dans l'ensemble d'entraînement et de 0.72% (0.66–0.79) dans l'ensemble test. Pour les prédicteurs cliniques du modèle Cox, les deux cohortes présentaient des distributions similaires (Table 6.3).

Tableau 6.3 : Description des cohortes d'entraînement et de test

	Cohorte d'entraînement	**Cohorte de test**
Nombre d'EEG (nombre de patients)	879 (786)	135 (115)
Récidive de crise au suivi (%)	295 (33.6)	40 (29.6)
Sexe = femme (%)	450 (51.3)	73 (54.1)
Âge (médiane [IQR])	49.00 [32.00, 62.00]	51.00 [30.00, 63.50]
Suivi total après l'EEG en semaines (médiane [IQR])	116.00 [52.00, 153.00]	113.00 [57.00, 194.50]
Type d'épilepsie (%)		
Focale	409 (46.5)	58 (43.0)
Généralisée	103 (11.7)	13 (9.6)
Inconnue	53 (6.0)	3 (2.2)
Pas d'épilepsie	314 (35.7)	61 (45.2)
Histoire de convulsion fébrile (%)	21 (2.4)	6 (4.4)
Histoire familiale d'épilepsie (%)	63 (7.2)	15 (11.1)
Nombre de jour depuis la dernière crise (médiane [IQR])	225.00 [52.00, 1130.00]	200.00 [73.00, 938.25]
Nombre de médicaments anticrises (%)		
0	327 (37.2)	67 (49.6)
1	338 (38.5)	37 (27.4)
2	145 (16.5)	18 (13.3)
3	54 (6.1)	8 (5.9)
4	12 (1.4)	5 (3.7)
5	3 (0.3)	0 (0.0)
Lésion focale à l'imagerie (%)	332 (37.8)	52 (38.5)
EEG de sommeil (%)	126 (14.3)	20 (14.8)
Pointes épileptiformes à l'EEG (%)		
Absence	654 (74.4)	103 (76.3)

Présence	156 (17.7)	26 (19.3)
Incertain	69 (7.8)	6 (4.4)
Ralentissement anormal à l'EEG (%)	266 (30.3)	42 (31.1)

6.3.2 Développement et entraînement du modèle

Pour le développement du modèle, la cohorte d'entraînement a été séparée en un échantillon d'entraînement (80%) et un échantillon de validation (20%) de façon aléatoire. L'échantillon d'entraînement comportait 1.4M de segments, et celui de validation, 0.3M. Cet ensemble de validation a permis d'optimiser les hyperparamètres suivants: taille des lots, taux d'apprentissage, *weight decay*, *dropout*, *gradient clipping* et probabilité d'augmentation. L'ensemble de données *TUH Abnormal Corpus* (3.7M de segments) a aussi été utilisé pour optimiser le format d'entrée du modèle et corriger les erreurs du code source [123].

Le modèle final a été entraîné sur l'ensemble de la cohorte d'entraînement (1.7M de segments de 60s pour 25 époques).

6.3.3 Performances

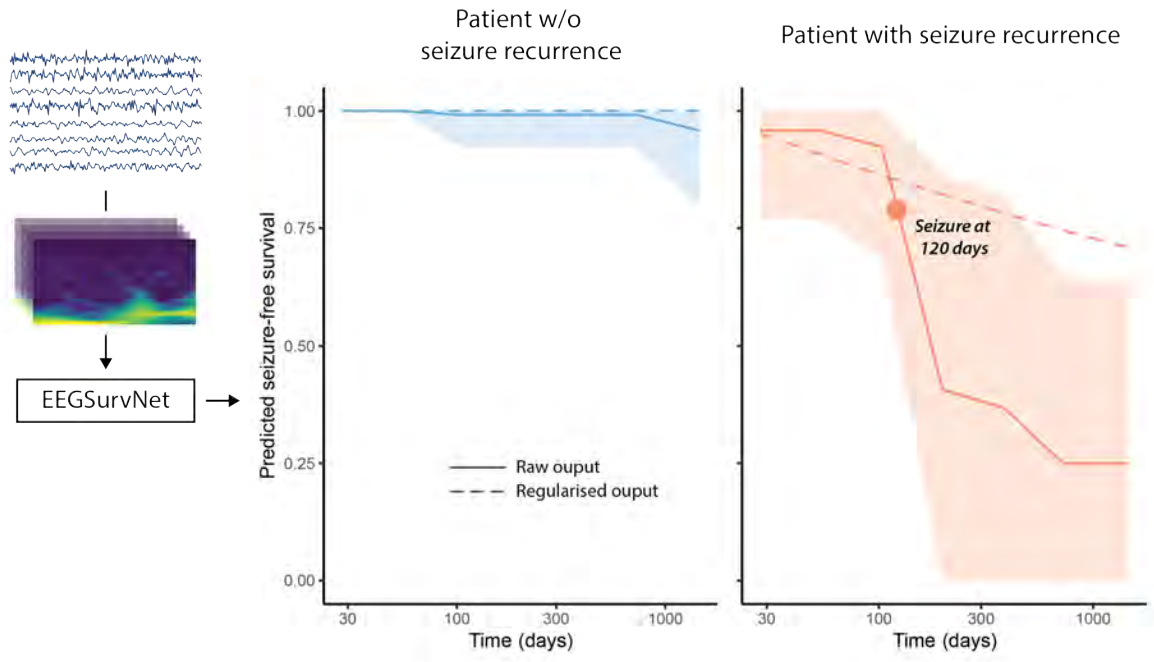


Figure 6.3 : Illustration des prédictions brutes du modèle EEGSurvNet sur deux patients de la cohorte test. Les EEG transformés en spectrogramme servent d'entrée au modèle. La sortie du modèle est ensuite convertie en fonction de survie qui caractérise la survie sans crise à chaque période ($S(t)$). La prédiction brute est basée sur le hasard moyenné sur les segments $\bar{h}(t)$, et la

prédiction régularisée, sur le risque constant h^*. Le ruban ombragé démontre la déviation standard des prédictions à travers les segments d'un même EEG.

Un exemple de prédictions du modèle EEGSurvNet est illustré à la Figure 6.3 et les résultats complets sont présentés dans le Table 6.4. Sur l'ensemble test, EEGSurvNet a obtenu un iAUROC sur deux ans de 0.69 (intervalle de confiance de 95%: 0.64–0.73) et un index C de 0.66 (0.60–0.73). Par comparaison, l'iAUROC du modèle Cox était de 0.61 (0.56–0.65; C = 0.61 [0.55–0.68]), et celui du modèle combiné, 0.70 (0.66–0.74; C = 0.69 [0.65–0.73]). Pour tous ces modèles, les performances sont supérieures au modèle de référence (iAUROC = 0.54 [0.49–0.59], C = 0.53 [0.45–0.61]).

L'AUROC en fonction du temps est présentée à la Figure 6.4. Pour EEGSurvNet, celles-ci atteignent un pic à ~2 mois (AUROC: 0.80 [0.72–0.88]) et sont statistiquement supérieure au modèle de référence entre les 3^{e} et 6^{e} mois.

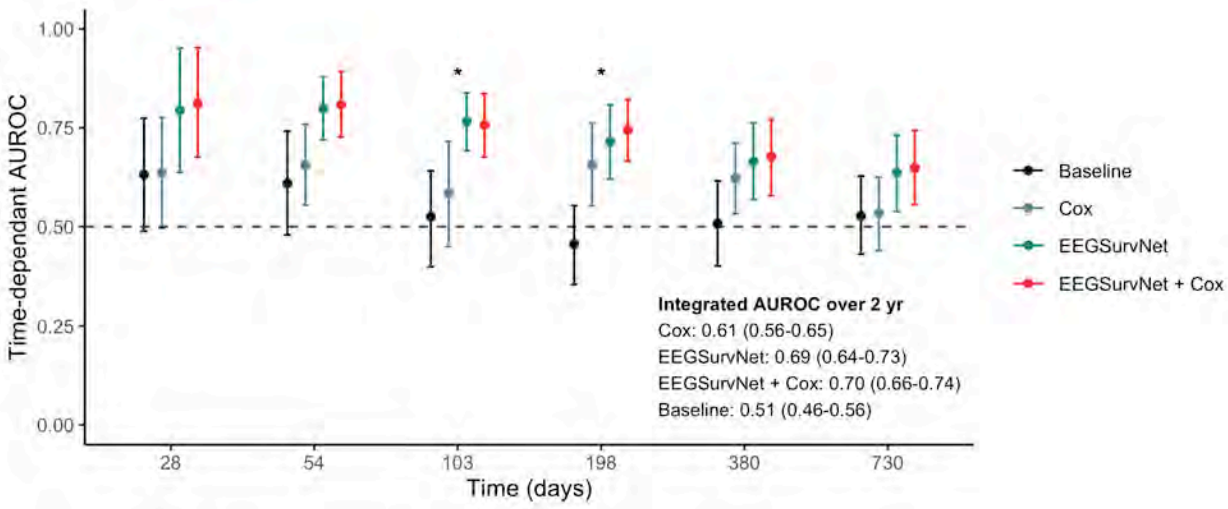


Figure 6.4 : Performance des différents modèles pour prédire la survie sans crise à travers le temps. Le modèle de référence (*Baseline*) correspond aux prédictions selon le risque de base de récidive de crise tiré de la cohorte d'entraînement. AUROC: Aire-sous-la-courbe ROC.

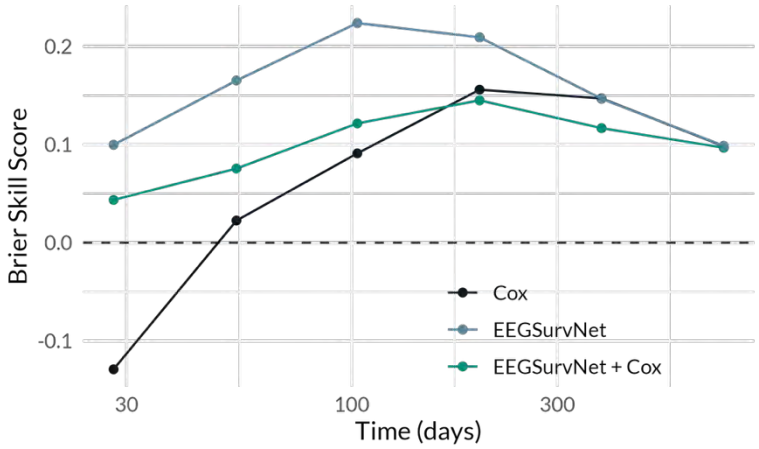


Figure 6.5 : *Brier Skill Score* des modèles de survie dans le temps. Le *Brier Skill Score* est une mesure de l'amélioration de la calibration contre le modèle de référence, où des valeurs entre 0 et 1 signifie un gain de calibration.

Pour la calibration, EEGSurvNet obtient un score de Brier intégré à 2 ans de 0.18 (0.15–0.20), statistiquement significativement supérieur au modèle de référence (0.24 [0.23–0.25])] (Table 6.4). Le *Brier Skill Score*, mesurant le gain de performance contre le modèle de référence, est positif sur l'ensemble des périodes et atteint un pic à 3 mois avec une valeur de 0.22 (Figure 6.4). La courbe de calibration en fonction du temps sont présentées à la Figure 6.6.

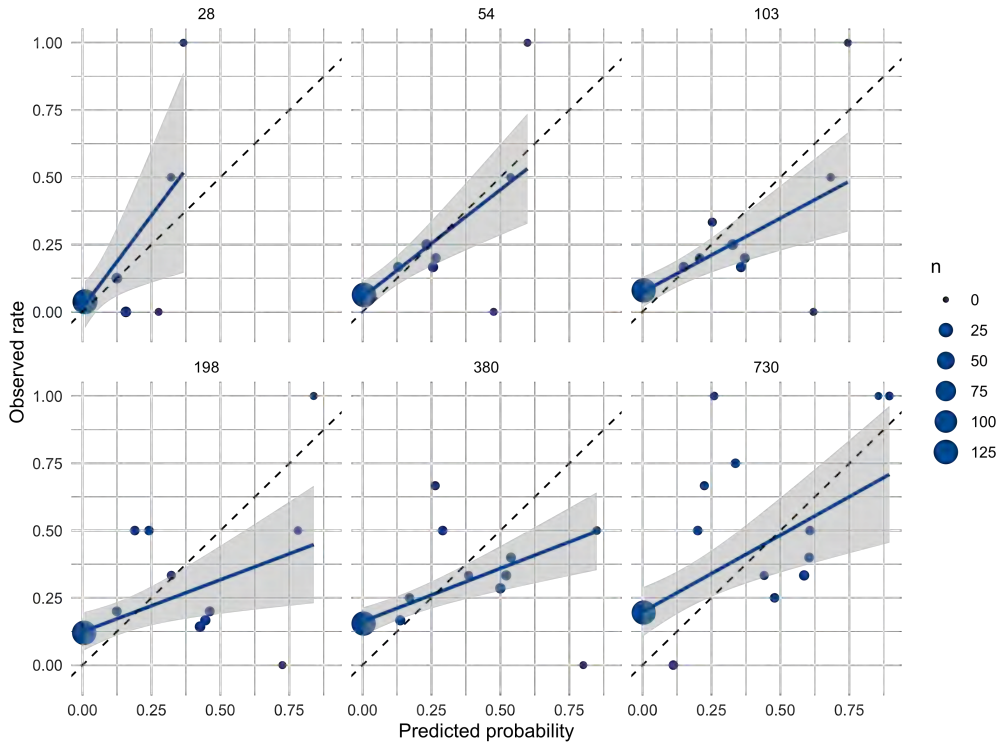


Figure 6.6 : Calibration du modèle EEGSurvNet en fonction du temps. Pour chaque période, les patients sont regroupés en bins de taille 0.2 selon leur probabilité prédite de crise. L'axe des x représente la probabilité moyenne prédite dans chaque bin, l'axe des y le taux observé de crises. La taille des points est proportionnelle au nombre de patients dans chaque bin. La ligne diagonale représente une calibration parfaite. La ligne bleue représente une régression linéaire pondérée avec intervalle de confiance à 95%.

Pour régulariser la sortie du modèle profond, le risque dans le temps $\bar{h}(t)$ est moyenné à travers les périodes pour un risque constant h^*. Sans cette opération, les performances d'EEGSurvNet sont les suivantes : iAUROC à 2 ans de 0.60 (0.55–0.65), un iBS à deux ans de 0.18 (0.16–0.20) et index C de 0.61 (0.53–0.69).

Tableau 6.4 : Discrimination et calibration des différents modèles sur la cohorte de test

	iAUROC à 1 an	**iAUROC à 2 ans**	**iBS à 1 an**	**iBS à 2 ans**	**Index C**
Baseline (risque de base)	0.53 (0.47–0.58)	0.54 (0.49–0.59)	0.17 (0.16–0.18)	0.24 (0.23–0.25)	0.53 (0.45–0.61)
Modèle clinique Cox	0.63 (0.58–0.68)	0.61 (0.56–0.65)	0.15 (0.14–0.16)	0.21 (0.20–0.23)	0.61 (0.55–0.68)
EEGSurvNet	0.72 (0.68–0.77)	0.69 (0.64–0.73)	0.13 (0.11–0.15)	0.18 (0.15–0.20)	0.66 (0.60–0.73)
EEGSurvNet + Cox	0.74 (0.70–0.78)	0.70 (0.66–0.74)	0.15 (0.12–0.17)	0.19 (0.17–0.21)	0.69 (0.65–0.73)

En stratifiant la cohorte test selon le score de survie moyen prédit, on obtient trois groupes de risque: « Faible risque », « Risque moyen » et « Haut risque ». La survie sans crise de ces patients est significativement associée avec le risque prédit par EEGSurvNet (p = 0.003, Figure 6.7). Chez

les patients à faible risque, le taux de récidive à deux ans est de 12%, contre 42% pour les patients à haut risque.

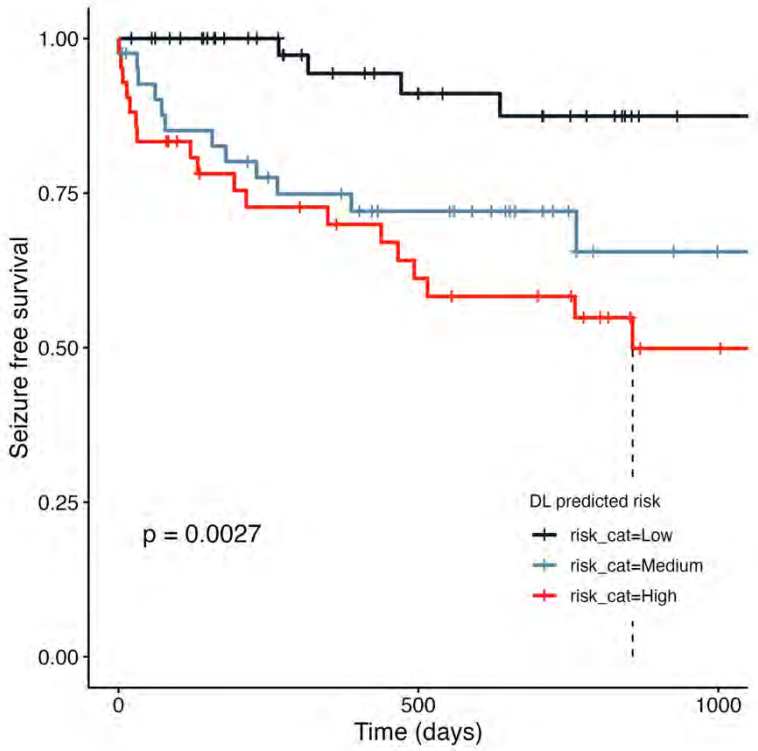


Figure 6.7 : Survie sans crise des patients de l'échantillon test, stratifié en fonction du risque prédit par le modèle EEGSurvNet.

6.3.4 Analyse par sous-groupe

Les analyses stratifiées révèlent des variations de performance selon les sous-groupes (Figure 6.8). Les jeunes adultes (<40 ans) présentent une meilleure discrimination avec un iAUROC à deux ans de 0.73 [0.68-0.79], comparativement aux 40–60 ans (0.64 [0.54-0.76]) et plus de 60 ans (0.67 [0.58-0.82]). Pour le type d'épilepsie, le modèle performe modérément en épilepsie focale (0.66 [0.60-0.72]) mais aléatoirement en épilepsie généralisée (0.50 [0.39-0.77]), bien que ce dernier résultat soit basé sur seulement 13 patients. La présence de lésion focale et l'absence de ralentissement anormal tendent à améliorer les performances. La seule différence statistiquement significative entre sous-groupes concerne la présence d'anomalies épileptiformes à l'EEG. Le modèle performe mieux en leur absence, avec un iAUROC de 0.78 (0.73-0.83), comparé à 0.53 (0.43-0.63) en leur présence.

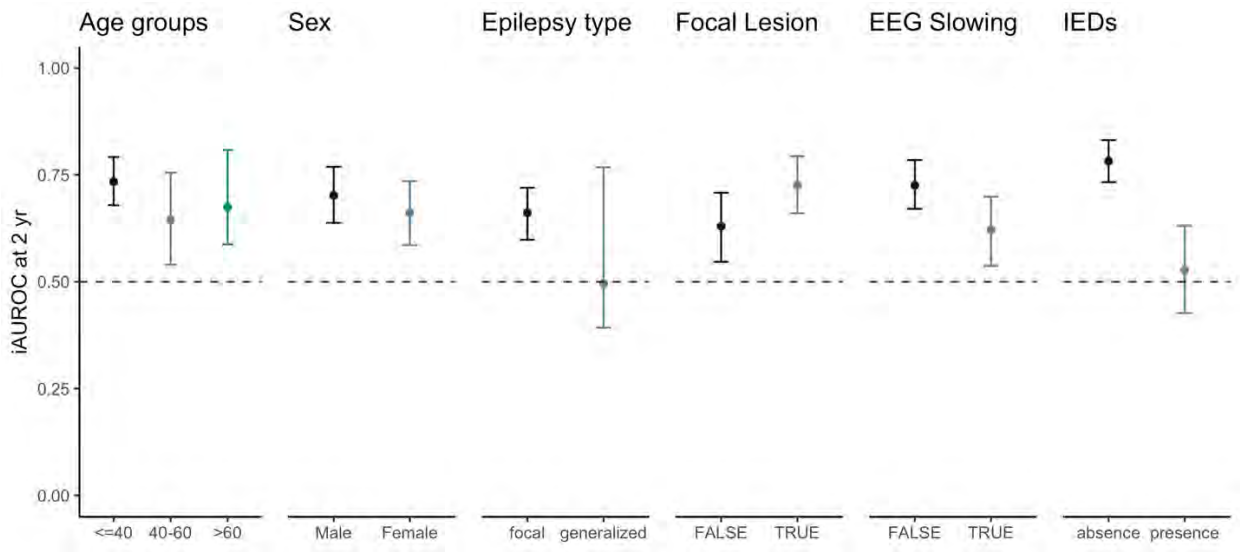


Figure 6.8 : Analyse des performances prédictives par sous-groupe. IED: *Interictal epileptiform discharges* (décharges épileptiformes intericales).

6.3.5 Interprétabilité

Pour interpréter le modèle profond, nous avons analysé les valeurs de Shapley par fréquence et par canal pour les 50 segments EEG où le modèle prédit un risque élevé, ainsi que pour les 50 segments où le modèle prédit un risque faible. Les résultats sont présentés à la Figure 6.9. Les valeurs de Shapley démontrent une importance élevée de la bande de fréquence de 6 à 15 Hz. En termes de localisation, les canaux situés au niveau temporal comportent la plus haute importance, suivi des canaux occipitaux. Nous observons aussi une asymétrie droite-gauche, avec des valeurs de Shapley plus élevées en temporal gauche que droit et un foyer de valeurs de Shapley basses en temporal

postérieur gauche.

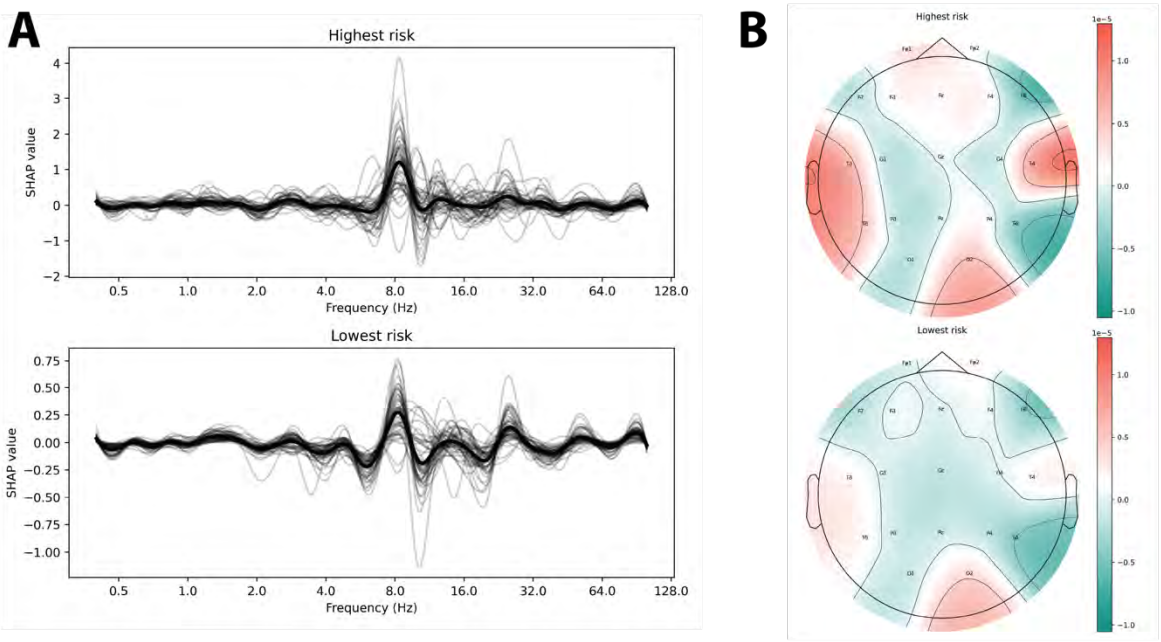


Figure 6.9 : Valeurs de Shapley pour les segments à plus haut et plus bas risque prédit. A : valeurs de Shapley moyennées par band de fréquence. La ligne plus sombre correspond à la moyenne sur les 50 segments. B : valeurs de Shapley moyennées par électrodes.

6.3.6 Étude d'ablation

Les résultats de l'étude d'ablation sont présentés dans le Table 6.5. Dans tous les cas, les performances sont inférieures à notre modèle. Le facteur ayant le plus d'impact est la résolution temporelle: lorsque celle-ci est diminuée de moitié, l'iAUROC à 2 ans passe de de 0.69 à 0.54. L'architecture originale de DeepEpilepsy, entraîné sur les signaux de base et non sur les spectrogrammes, n'a pas été transférable directement à l'analyse de survie, avec une iAUROC à deux ans de seulement 0.52.

Tableau 6.5 : Étude d'ablation

Modèle	Augmentation	Résolution temporelle (s)	Résolution fréquentielle	Durée des segments (s)	iAUROC*
DeepEpilepsy	Oui	0.05	–	30	0.52
No_augment	Non	0.2	32	60	0.58
16freq_8decim	Oui	0.4	16	60	0.53
8decim	Oui	0.4	32	60	0.54
16freq	Oui	0.2	16	60	0.62
30s	Oui	0.2	32	30	0.58
EEGSurvNet	**Oui**	**0.2**	**32**	**60**	**0.69**

**Aire sous la courbe ROC intégrée à 2 ans*

6.3.7 Discussion

Cette étude présente le développement et la validation d'un modèle de survie profond, EEGSurvNet, pour prédire le risque de crise à partir de l'EEG de routine. Le modèle démontre des performances supérieures aux prédicteurs traditionnels, autant en termes de discrimination (iAUROC à deux ans = 0.69) que de calibration (iBS à deux ans = 0.18). Les performances sont plus élevées dans les premiers mois suivants l'EEG, atteignant un AUROC de 0.80 à 2 mois. L'ajout de prédicteur cliniques améliore marginalement les performances, suggérant que l'EEG capte une part essentielle de l'information pronostique.

L'analyse de survie offre plusieurs avantages par rapport aux approches binaires traditionnelles. Elle modélise explicitement la composante temporelle du risque, tient compte des durées variables de suivi et permet une évaluation simultanée des effets à différents intervalles. Couplée à un modèle profond, elle permet de relier des données complexes comme l'image, le texte, ou les séries temporelles, avec une issue clinique plus granulaire. En oncologie, des modèles de survie profonds ont permis d'améliorer la prédiction du pronostic du chondrosarcome [221], du cancer du poumon [223] et du cancer de l'estomac [222], dépassant les performances des systèmes de classifications des stades traditionnels. En soins intensifs, Thorsen et al. ont développé un modèle intégrant données cliniques, texte libre et séries temporelles pour prédire la survie des patients avec un index de concordance de 0.73 [224]. Dans certains cas, cette approche peut être jumelée à une recommandation thérapeutique personnalisée [220].

En épilepsie, les modèles de survie traditionnels ont joué un rôle important dans la compréhension de l'histoire naturelle de la maladie et l'identification des facteurs pronostiques. Plusieurs études ont utilisé l'analyse de survie pour évaluer le risque de récidive après une première crise [25], caractériser la réponse au traitement [45], ou prédire la rémission à long terme après l'arrêt de la médication anti-crise [245], [246]. Certains modèles validés ont été spécifiquement développés pour guider la prise en charge clinique, notamment pour prédire la récidive de crise après une première crise aigüe symptomatique dans le contexte d'une hémorragie intracrânienne [216], [247], d'un accident vasculaire cérébral ischémique [217], [236], d'une thrombose veineuse cérébrale [238] ou d'un trauma crânien [237], ou pour prédire le succès du sevrage de la médication anticrise

[218]. Pourtant, malgré cette utilisation établie des modèles de survie en épilepsie et le succès récent des approches profondes dans d'autres domaines médicaux, aucune étude n'a encore exploré le potentiel des modèles de survie profonds pour l'analyse de l'EEG.

Les modèles de survie clinique existants ont des performances qui varient selon le contexte clinique, avec des index C allant de 0.67 pour la prédiction de récidive après le sevrage de la médication [218] à 0.89 pour le développement d'épilepsie post-traumatique [237]. Ces modèles, basés sur des variables cliniques facilement disponibles comme la présence de DEI à l'EEG, la sévérité de l'atteinte aigüe ou la durée de l'épilepsie, sont disponibles sous forme de nomogrammes (https://predictepilepsy.com) et sont utilisés dans la pratique clinique [248]. Dans notre cohorte moins sélectionnée, le modèle clinique obtient un index C de 0.61, reflétant la difficulté apportée d'une population plus hétérogène. Toutefois, EEGSurvNet atteint des performances comparables aux modèles cliniques validés, particulièrement lorsque combiné aux données cliniques (C = 0.69), suggérant sa pertinence pour une utilisation en pratique courante. Comme pour les autres modèles prédictifs en épilepsie, l'impact réel de ces prédictions sur les décisions cliniques et les issues des patients reste à démontrer par des études prospectives [248].

EEGSurvNet tend vers de meilleures performances chez certains sous-groupes, notamment les jeunes patients, chez les patients avec épilepsie focale, chez les patients sans ralentissement à l'EEG et chez les patients sans anomalie épileptiformes à l'EEG. Malheureusement, l'interprétation de cette analyse est complexifiée par la corrélation entre plusieurs de ces variables dans notre échantillon. Notamment, tous les patients de <40 ans présentaient une épilepsie focale (n = 11) ou indéterminée (n = 1), et la majorité des patients avec épilepsie généralisée montraient des DÉI à l'EEG (69% vs. 27% avec épilepsie focale). Il est donc difficile de départager l'effet de chacun de ces facteurs sur les performances. Plusieurs hypothèses peuvent tout de même être avancées. Les performances supérieures chez les jeunes patients pourraient s'expliquer par l'absence de facteurs confondants comme la polymédication ou comorbidités, plus fréquents chez les personnes âgées et susceptibles de masquer les marqueurs EEG de risque de crise. Concernant le type d'épilepsie, la meilleure performance dans l'épilepsie focale pourrait refléter des anomalies plus marquées du signal: 66% des patients présentaient une lésion à l'imagerie et 50% un ralentissement EEG anormal, contre seulement 8% et 15% respectivement dans l'épilepsie généralisée. Ces différences structurelles et fonctionnelles pourraient générer des signatures EEG

plus marquées. Tout de même, les EEG sans ralentissement étaient plus faciles à discriminer, suggérant que le ralentissement est un facteur confondant pour le modèle.

Un résultat particulièrement intéressant est la performance supérieure du modèle sur les EEG sans anomalies épileptiformes (iAUROC: 0.78 vs 0.53). Bien que cette différence puisse partiellement s'expliquer par des taux de récidive différents (23% sans pointes vs. 46% avec pointes) et le type d'épilepsie associé (28% des EEG en épilepsie focal présentait des DÉI vs. 69% en épilepsie généralisée), elle suggère surtout que le modèle utilise des caractéristiques du signal distinctes des marqueurs épileptiformes classiques. Ceci est cohérent avec la méthode de traitement des sorties du modèle, qui fait la moyenne des prédictions pour tout l'EEG. Cette méthode pourrait enlever du poids aux patrons paroxystiques comme les DÉI et favoriser les anomalies plus constantes du rythme de fond.

L'analyse des valeurs SHAP soulève plusieurs hypothèses quant aux patrons alternatifs détectés par EEGSurvNet. Une attention particulière semble être accordée aux régions temporales avec une asymétrie gauche-droite, ce qui pourrait s'expliquer par la prépondérance de l'épilepsie temporale dans notre cohorte (44% des cas, avec une latéralisation gauche prédominante). L'autre région importante est le lobe occipital. Combiné à l'importance accordée aux fréquences 6–15 Hz, ceci suggère que le modèle s'appuie sur le rythme postérieur dominant comme autre marqueur prédictif, ce qui concorde avec plusieurs travaux ayant révélé des altérations du rythme alpha chez les patients avec épilepsie [85], [152], [249], [250], [251]. D'autres patrons EEG de basse fréquence associés au risque de crise incluent le ralentissement temporal rythmique intermittent (TIRDA) [252] et les *Paroxysmal Slow Wave Events* [77], mais ces patrons sont intermittents et présents dans la bande 0–6 Hz, ce qui rend moins probable leur détection par le modèle. L'étude d'ablation souligne l'importance des longs segments, suggérant que les patrons pertinents évoluent sur une échelle temporelle d'une minute. Certaines trouvailles contrastent avec DeepEpilepsy, qui performait mieux avec des segments de 30s et s'appuyait sur les hautes fréquences (50–100 Hz) [214]. Cette différence dans les caractéristiques exploitées pourrait expliquer la meilleure robustesse d'EEGSurvNet, les fréquences 6–15 Hz étant généralement moins sensibles aux artéfacts [253].

Afin de combiner l'information clinique et l'EEG, nous avons simplement multiplié le rapport de risque (*hazards ratio*) avec le hasard prédit par le modèle profond. Cette approche a permis

d'améliorer légèrement les performances d'EEGSurvNet avec un iAUROC à 2 ans de 0.69 à 70. Cependant, elle ne permet pas d'interactions entre les deux modalités, les modèles profonds et Cox demeurant indépendants. L'apprentissage multimodal en EEG demeure un sujet peu exploré, particulièrement en épilepsie. Une revue de littérature récente [254] a identifié quatre articles en épilepsie combinant l'EEG à une autre modalité (EMG, spectroscopie proche infra-rouge, ECG) principalement pour améliorer la détection de crises d'épilepsie [255], [256], [257], [258]. Dans notre cas, la difficulté était de combiner des données tabulaires avec des séries temporelles. Une approche intéressante à envisager serait l'apprentissage auto-supervisé, qui a démontré sa faisabilité en maladie d'Alzheimer et en cardiologie en apprenant une représentation conjointe de l'IRM et des données cliniques [259], [260].

Cette étude comporte plusieurs limitations. Premièrement, les patients proviennent tous du même centre. Malgré la standardisation de l'EEG clinique, certains facteurs peuvent varier d'un centre à l'autre, autant en termes d'enregistrement que de pratique clinique. Par exemple, nos trouvailles pourraient mal se généraliser à une pratique de neurologie générale avec peu de cas d'épilepsie réfractaire, ou encore un centre où l'EEG ambulatoire est plus souvent utilisé que l'EEG de routine. Deuxièmement, le suivi clinique diffère entre les patients avec et sans épilepsie. Les patients avec épilepsie ont tendance à être suivi à plus long terme, ce qui affecte la probabilité de censure de l'analyse de survie. Afin de mitiger l'impact de ce biais, nous avons utilisé une pondération basée sur la censure (*inverse probability of censoring weights*: IPCW), mais son effet réel est difficile à quantifier étant donné que nos données sont issues de la pratique clinique réelle. Troisièmement, notre taille d'échantillon reste modeste pour les analyses par sous-groupes, pour lesquelles une erreur de type II ne peut être exclue.

6.4 Prochaine étape: validation externe

La validation externe d'EEGSurvNet est en cours en collaboration avec l'Université de Pennsylvanie (UPenn) et la Harvard Medical School (HMS). Ces centres disposent chacun d'un jeu de données d'environ 1 000 EEG avec notes cliniques correspondantes, offrant l'opportunité de tester la généralisabilité du modèle dans d'autres centres hospitaliers.

La première étape, presque achevée, consiste en la standardisation des données selon le format BIDS [227] pour faciliter leur partage et leur analyse. Pour extraire les données de suivi clinique, nous utiliserons un algorithme de traitement du langage naturel développé par UPenn qui permet

d'identifier automatiquement le nombre de crises entre les visites cliniques dans les notes médicales [261]. Cette approche automatisée permettra de déterminer systématiquement soit la date de la première crise post-EEG, soit la date de censure pour chaque patient.

Une fois ces données cliniques extraites, nous appliquerons EEGSurvNet aux EEG de ces cohortes externes. La performance du modèle sera évaluée selon les mêmes métriques que dans notre étude initiale: le score de Brier, l'AUROC dynamique/cumulatif et l'index C. Les analyses stratifiées seront également reproduites pour évaluer la consistance des patrons de performance observés dans notre cohorte, notamment concernant l'âge, le type d'épilepsie et la présence d'anomalies épileptiformes.

6.5 Conclusion

En conclusion, cette étude démontre la performance d'un modèle de survie profond pour prédire le risque de crise jusqu'à deux ans après un EEG de routine. Sa capacité à prédire le risque sur les EEG sans pointes renforce la suspicion de biomarqueurs neurophysiologiques invisibles à l'œil nu, possiblement dans les fréquences 6 à 15 Hz. Cette approche pourrait grandement améliorer le diagnostic et le suivi de l'épilepsie, qui sont actuellement grandement limités par le manque de biomarqueurs quantifiables du risque de crise. La prochaine étape sera de valider les performances sur des échantillons indépendants. Une collaboration est établie avec les Université de Harvard (Boston) et de Pennsylvanie (Philadelphie), où la collecte de données est présentement en cours. Par la suite, une étude prospective de ce modèle permettrait d'évaluer son utilité clinique réelle lors du diagnostic ou du suivi de patients avec épilepsie.

CHAPITRE 7 DISCUSSION GÉNÉRALE

Dans cette thèse, nous avons démontré l'applicabilité de l'analyse automatisée de l'EEG pour quantifier le risque de récidive de crise d'épilepsie. En premier lieu, nous avons exploré les performances diagnostiques de marqueurs computationnels décrits précédemment, appliqués à une cohorte de patients consécutifs ayant eu un EEG de routine au CHUM. Cette première expérience a démontré que ces marqueurs permettaient de distinguer les EEG de patients à haut risque de crise au-delà de la chance de façon statistiquement significative, mais avec des performances modestes. Pour améliorer la précision diagnostique, nous nous sommes ensuite tournés vers l'apprentissage profond afin d'augmenter la complexité de la représentation extraite du signal. Les modèles profonds, et plus spécifiquement les ViT dont DeepEpilepsy, ont permis d'augmenter la performance diagnostique comparativement aux marqueurs computationnels. Finalement, nous avons adapté DeepEpilepsy afin de prédire le risque de crise à travers le temps à la manière d'un modèle de survie. Cette approche permet de prédire une issue clinique actionnable, potentiellement plus pertinente comparativement au diagnostic seul. Notre modèle final, EEGSurvNet, peut prédire le risque de crise à travers le temps jusqu'à deux ans après un EEG de routine, mieux qu'un modèle basé sur les données cliniques seules.

Comme décrit dans la revue de littérature, la majorité des travaux antérieurs sur la détection automatisée de l'épilepsie à l'EEG de routine porte sur des marqueurs computationnels. Ces marqueurs incluent des variations subtiles dans les bandes de fréquences [152], [249], [250], [262], une diminution de la régularité du signal et de son entropie [203], [263], ou bien des anomalies de connectivité fonctionnelle [88], [100], [264]. Cependant, ces changements ont souvent été détectés dans des études cas-contrôles, où des sujets provenant d'une certaine population (e.g., patients avec épilepsie) sont comparés avec des sujets d'une autre population (contrôles sains, patients avec crises non-épileptiques, etc.) [67]. Dans ce type d'étude, les différences entre les deux populations vont au-delà du diagnostic d'épilepsie et du risque de crise, et incluent les comorbidités, la prévalence de lésions cérébrales, la charge médicamenteuse et possiblement d'autres facteurs confondants non-mesurables. Ainsi, les études cas-témoins ont tendance à surévaluer les performances diagnostiques [170]. Dans notre cas, nous avons mis sur pied une cohorte qui comprend tous les patients consécutifs s'étant présentés pour un EEG de routine au CHUM et y ayant eu un suivi. Cette cohorte est représentative des patients chez qui nous déploierions un outil

d'analyse automatisée de l'EEG dans la vraie vie, et nous permet donc d'évaluer plus robustement les performances d'un tel outil [265]. De plus, la révision complète des dossiers médicaux de chaque patient avec un long suivi clinique (médiane de 2.2 ans), augmente la confiance dans les étiquettes utilisées pour l'apprentissage et permet des analyses stratifiées par sous-groupe. Ces détails méthodologiques visent à surmonter la principale limite des études précédentes : la sélection des sujets de l'étude [67].

Sur cette cohorte, nous avons premièrement démontré que les marqueurs computationnels étaient bien capables d'identifier des différences dans l'EEG de patients avec épilepsie. Le marqueur le plus performant était la puissance des bandes spectrales, dont la capacité discriminative atteignait presque celles de tous les marqueurs combinés. Il est intéressant de noter que le modèle pouvait identifier les EEG des patients avec épilepsie même dans sans ralentissement anormal à la lecture par un neurologue, suggérant qu'un processus plus complexe est à l'origine des différences détectées par cette approche. Une hypothèse est celle d'une interaction pathologique entre les différentes bandes de fréquence [266]. En effet, les patients avec épilepsie présentent un couplage anormal entre les hautes et basses fréquences dans les moments précédant une crise [267], [268]. Bien que notre méthode ne permette pas de détecter explicitement les couplages à travers les fréquences, ce type d'évènement pourrait avoir une répercussion sur la puissance des bandes pendant l'EEG interictal [269]. La seconde catégorie de marqueurs qui atteignaient des performances statistiquement au-delà de la chance étaient les marqueurs non-linéaires de régularité du signal, incluant l'entropie et la longueur de ligne (*Line Length*). Dans les années 1990, l'application de l'analyse dynamique non-linéaire à l'EEG démontrait une organisation anormale du signal lors d'une crise et dans les moments qui la précédaient [270], [271], [272], [273]. D'autres travaux ont dénoté cette anomalie à distance des crises [80], [203], [263]. Notre premier travail corrobore l'hypothèse de changements dans l'entropie des patients avec épilepsie en interictal, mais la nature exacte de ces changements est difficile à identifier. En effet, chaque marqueur est extrait à plusieurs échelles temporelles et plusieurs localisations, puis est amené à interagir avec les autres marqueurs de façon non-linéaire. Ainsi, bien que notre travail démontre des différences dans le signal EEG, il ne permet pas de faire le lien avec des processus physiopathologiques précis, ce qui requerrait une étude dédiée.

En utilisant les marqueurs computationnels comme entrée aux modèles, nous effectuons une forme de « régularisation » et contraignons la représentation de l'EEG à des changements observables,

ou du moins basés sur des hypothèses neurophysiologiques. En réalité, notre compréhension des processus générant le signal est limitée [21], [22], et dans ce contexte, l'apprentissage profond est très pertinent. Son utilisation permet de faire table rase sur les hypothèses qui sous-tendent les changements interictaux liés à l'épilepsie et au seuil convulsif, donnant (presque) carte blanche au modèle profond pour trouver les caractéristiques pertinentes. Le choix de l'architecture du modèle permet de diriger faiblement la nature des caractéristiques extraites.

L'application de l'apprentissage profond à l'EEG a beaucoup évolué depuis les modèles ShallowConvNet [119] et EEGNet [118]. Ces modèles relativement simple (1.5K–300K paramètres) s'inspiraient d'algorithmes de traitement de signal comme le *Filter Bank Common Spatial Patterns* pour extraire des caractéristiques locales, et dont les performances de pointes plafonnent à une certaine taille de modèle et d'échantillon d'entraînement [122]. Plus récemment, les modèles SpikeNet et Score-AI, adaptés d'architecture plus modernes comme le ResNet [114] et DenseNet [274], ont démontré l'applicabilité de modèles plus profonds (~300K et 20M de paramètres) et entraînés sur de plus grandes bases de données (10K et 30K EEG) pour identifier des DÉI ou des périodes de ralentissement à l'EEG [103], [104].

DeepEpilepsy et EEGSurvNet se distinguent par une architecture hybride qui combine les convolutions au mécanisme d'attention du Transformeur [131], [132]. Les convolutions compriment efficacement le signal en intégrant rapidement l'information spatiale à courte échelle temporelle (0.2–1s), puis le Transformeur modélise l'information entre ces états sur des périodes plus longues (30s pour DeepEpilepsy, 60s pour EEGSurvNet). Cette architecture permet d'identifier des processus plus subtils avec dynamiques temporelles plus étendues, comme le démontre la performance supérieure de DeepEpilepsy comparée aux marqueurs computationnels et au ShallowConvNet. Bien que DeepEpilepsy surpasse aussi un CNN moderne de type ResNet (ConvNeXt, [116]), cette différence est probablement marginale, comme le démontrent des travaux en vision par ordinateur où CNN et ViT atteignent des performances similaires lorsque correctement optimisés [115]. Par rapport aux autres modèles profonds utilisés pour le diagnostic d'épilepsie, DeepEpilepsy et EEGSurvNet se distinguent aussi par leur taille. Avec 10M et 27M de paramètres, ils sont 300 à 9 000 fois plus grands que leurs prédécesseurs [73], [74], [87], [98]. Cette augmentation de complexité est contrebalancée par des méthodes de régularisation (*weigth decay*, *dropout*, augmentation de données) qui préviennent le surapprentissage [133].

Les résultats de notre deuxième étude suggèrent que des modèles de plus grandes tailles mènent à des meilleures performances, avec des performances corrélées à la taille de l'échantillon d'entraînement. Ceci est contraire aux conclusions de Kiessner et al., qui ont identifié une saturation des performances entre 3 500 et 11 000 paramètres pour divers CNN sur un ensemble de données de 10 000 EEG classifiés comme normaux vs. anormaux [122]. Premièrement, il est possible que leurs architectures (basés sur EEGNet et ShallowConvNet) possèdent un biais inductif qui empêche d'apprendre des caractéristiques assez complexes [112]. Ceci concorde avec nos trouvailles, où la performance de ShallowConvNet saturait assez rapidement avec la taille d'entraînement. Deuxièmement, il existe une variabilité inter-observateurs non-négligeable dans l'interprétation d'un EEG [103], [213]. Le jeu de données utilisé comporte donc un taux d'erreur minimal qui est inconnu, plafonnant les performances. Ces trouvailles justifient l'investissement dans des infrastructures de collecte de données à grande échelle et des ressources computationnelles plus puissantes. Avec l'émergence de bases de données EEG en épilepsie telles que le *Temple University Hospital EEG Corpus* [123] et, plus récemment, le *Harvard EEG Database* [275], les modèles profonds prendront sûrement plus d'espace dans ce domaine. Ces données ouvrent la voie à des modèles encore plus ambitieux et l'application de techniques comme l'apprentissage auto-supervisée [276], dont l'application à l'EEG est à ses balbutiements [277], [278].

La nature exacte des patrons détectés par les modèles profonds demeure inconnue. Parmi les quatre études antérieures utilisant l'apprentissage profond pour détecter l'épilepsie à l'EEG, deux ont utilisé des techniques d'interprétabilité pour comprendre les caractéristiques extraites par les modèles [74], [87]. Dans Uyttenhove et al., les auteurs ont utilisé Grad-CAM pour détecter les segments d'EEG associés avec la classe « épilepsie », et ont démontré que le modèle avait appris à détecter les anomalies épileptiformes comme les pointes et les crises. Cette étude a utilisé le *Temple University Dataset*, qui comporte des EEG hautement anormaux incluant des tracés des soins intensifs. Dans Rijnders et al., les EEG sont présentés au CNN après extraction des matrices de connectivité, restreignant fortement la capacité du modèle à découvrir des patrons au sein du signal [87].

Une des nouveautés des modèles présentés dans cette thèse est qu'ils utilisent le signal brut ou minimalement transformé. De plus, leur performance est maintenue en absence d'anomalies visibles. Ces modèles offrent donc une nouvelle fenêtre sur les répercussions

électroencéphalographiques de l'épilepsie. Pour DeepEpilepsy, notre analyse d'interprétabilité consistait à mesurer la variance de deux caractéristiques connues, l'entropie et la puissance de bande spectrale, au sein de l'espace latent appris par le modèle. Cette approche a révélé que les puissances des bandes de haute-fréquence avait une plus grande variance, et donc que les patrons détectés par le modèle y sont situés. Pour EEGSurvNet, nous avons plutôt estimé les valeurs Shapley [243], [244], ce qui a démontré une sensibilité du modèle à la bande thêta haute-alpha, ainsi qu'aux canaux temporaux et occipitaux. Ces contradictions entre les deux modèles s'expliquent possiblement par les changements du format d'entrée. En effet, pour EEGSurvNet, les données sont transformées en spectrogrammes et sont légèrement filtrées (filtre passe-haut à 0.2 Hz et filtre Notch à 60 Hz). Ces modifications permettent possiblement de régulariser l'apprentissage et ont un impact positif sur les performances. Sans ces modifications, DeepEpilepsy était incapable de s'adapter à la tâche de l'analyse de survie, qui est plus complexe que la classification simple. Les hautes fréquences contiennent beaucoup d'artéfacts (surtout myogéniques et de contamination électrique), qui pourraient avoir confondues DeepEpilepsy. L'attention portée par EEGSurvNet aux fréquences 6 à 15 Hz est particulièrement intéressante car elle suggère une relation entre des anomalies du rythme alpha et le risque de crise, une des premières anomalies relevées par les analyses quantitatives de l'EEG dans les années 1940 [249].

La détection de l'épilepsie ou l'identification de patrons anormaux sont des issues cliniques pertinentes, mais la question fondamentale en épilepsie demeure le risque de crise chez un patient donné [2]. L'utilisation d'issue binaire (épilepsie vs. absence d'épilepsie, récidive de crise vs. absence de récidive) apporte plusieurs limitations. Premièrement, l'épilepsie comporte un spectre très large de présentation: certains patients auront des crises quotidiennes, alors que d'autres n'auront jamais de récidive. Deuxièmement, la prise de décision clinique ne s'arrête pas au diagnostic d'épilepsie; l'épilepsie est une maladie chronique avec une dynamique complexe, et une fois le diagnostic posé, l'ajustement du traitement passe par l'estimation du risque de crise. Troisièmement, par définition, les patients sans épilepsie aussi ont un risque théorique de crise, qui est inférieur à 60% à 4 ans [2]. Il est donc utile de modéliser ce risque théorique dans le contexte de la définition actuelle de l'épilepsie [2]. L'analyse de survie permet donc de réconcilier ces limitations en attribuant un risque à tous les patients, avec et sans épilepsie, une approche qui est analogue à la quantification du seuil convulsif.

Notre modèle final, EEGSurvNet, pourrait être déployée dans le contexte clinique et offrir une quantification du risque à tous les patients qui ont un EEG de routine. Ce serait ensuite au neurologue traitant d'intégrer cette information à la prise en charge clinique. Les impacts potentiels de cette information sont multiples. Premièrement, chez un patient se présentant avec une incertitude diagnostique (par exemple, ayant eu une crise épileptique unique ou avec épisodes neurologiques répétés de nature incertaine), un risque de crise prédit comme élevé pourrait mener à un suivi plus serré, une admission à l'unité de monitoring d'épilepsie, ou même un essai de médication anticrise. Au contraire, un risque prédit faible pourrait justifier une absence de suivi, étant donné que les patients à faibles risques n'ont que 12% de chance d'avoir une convulsion à 2 ans. Par comparaison, sans utiliser de modèle profond, les patients avec un EEG sans anomalies épileptiformes et une crise non-provoquée ont un risque de récidive à 2 ans d'environ 25% [24]. Une autre utilité clinique potentielle est la sélection de patients pour le sevrage de médication anticrise. Certains patients ont un bon contrôle de l'épilepsie avec la médication, mais celle-ci comporte des effets secondaires nuisibles comme la somnolence, étourdissements et troubles de l'humeur [279]. Chez les patients sans crise depuis plus d'un an, l'échec de sevrage de la médication est d'environ 15–45% [174], [280]. EEGSurvNet pourrait donc s'ajouter aux outils cliniques existants [218], [280] pour diminuer le risque de récidive lors du sevrage. Finalement, les restrictions quant à la conduite automobile sont un enjeu majeur dans la qualité de vie des patients avec épilepsie [281]. Au Québec et ailleurs dans le monde, la réglementation quant à la conduite automobile est presqu'entièrement déterminée par la date de la dernière crise [239], [282]. Un outil plus précis, basé sur la quantification du risque de crise, pourrait améliorer la sécurité automobile et la qualité de vie des patients [283]. Dans tous ces cas, des études prospectives seront nécessaires pour évaluer l'impact réel sur la prise de décision clinique.

CHAPITRE 8 CONCLUSION

L'objectif principal de cette thèse était de développer des méthodes computationnelles pour extraire de l'EEG des marqueurs quantifiables du risque de crise. À travers trois études complémentaires et appuyés par une base de données cliniques de haute qualité, nous avons progressé d'une classification binaire basée sur des caractéristiques prédéfinies vers une modélisation temporelle complexe du risque épileptique tirant avantage de l'apprentissage profond.

La première implication de ce travail est l'établissement de nouveaux standards méthodologiques pour le développement d'algorithmes appliqués à l'EEG en épilepsie. Tel que démontré par notre revue systématique, le principal facteur qui freine la translation clinique dans ce domaine est la présence de biais méthodologiques dans la sélection des sujets et l'évaluation des algorithmes [67]. Au centre de cette thèse se situe une cohorte de patients consécutifs ayant eu un EEG de routine, représentant le spectre complet des patients chez qui cet examen est effectué et, ultimement, chez qui cette technologie serait déployée. À travers nos études, une emphase particulière est attribuée à la validation chez une cohorte de patients indépendants avec un décalage temporel. Malgré cela, la principale limitation de cette thèse est que les patients proviennent tous du même centre hospitalier. La validation de notre approche sur des cohortes similaires recrutés dans d'autres centres hospitaliers est présentement amorcée.

L'autre implication importante est la découverte d'une nouvelle fenêtre pour investiguer la neurophysiologie de l'épilepsie. Les algorithmes profonds permettent une représentation du signal EEG distincte de celle perçue par l'analyse visuelle. Les caractéristiques captées par les modèles ont une échelle temporelle de l'ordre d'au moins une minute, et semblent impliquer principalement la bande de fréquence 6 à 15 Hz. Et plus important encore, ces marqueurs semblent être indépendants de la présence de décharges épileptiformes intericales. Ces patrons pourraient relever d'anomalies de plus haut niveau comme des altérations structurelles ou fonctionnelles de réseaux neuronaux, ou même des couplages entre les fréquences d'oscillation. De plus, les marqueurs pourraient varier selon le type d'épilepsie, de l'étiologie sous-jacente, de l'état d'éveil ou du nombre de médicaments. Nos trouvailles sont toutefois limitées par le nombre restreint d'EEG. En effet, la qualité de la représentation apprise par les ViT est corrélée avec le nombre d'échantillon [284], et bien que notre méthode de segmentation permette de présenter 1.7M segments aux modèles, ces segments ne proviennent en réalité que de 900 EEG différents. Une

étude approfondie de la représentation profonde sur un plus grand jeu de données pourrait mener à une meilleure compréhension de la physiopathologie de l'épilepsie.

Finalement, l'implication majeure est la meilleure caractérisation du risque de crise chez les patients. La définition même de l'épilepsie repose sur le risque de crise, mais sa quantification est imprécise et fondée sur quelques éléments cliniques parfois subjectifs, telle que la certitude que les épisodes antérieurs constituent des crises et la présence de pointes épileptiformes à l'EEG [2]. La découverte de marqueurs quantifiables du risque de crise révolutionnerait la prise en charge de ces patients [49], [50], [157]. La certitude diagnostique serait améliorée, réduisant d'une part les délais diagnostiques mais surtout le surdiagnostic, dont les conséquences sont encore plus sévères [12], [13]. Lors du suivi, cette information permettrait un ajustement plus juste de la médication anticrise, une amélioration de la sécurité automobile, une référence plus précoce pour une évaluation chirurgicale et un sevrage plus sécuritaire des médicaments chez les patients sélectionnés [248]. Cette quantification du risque pourrait augmenter l'efficacité des unités de monitoring vidéo-EEG en présélectionnant les patients à haut risque de crise dans les prochains jours. Il en est de même pour les essais cliniques, qui pourraient sélectionner plus efficacement les patients avec un seuil convulsif abaissé. Bref, cela redéfinirait le rôle de l'EEG en épilepsie, le transformant en examen périodique qui offrirait une information à jour du contrôle de la maladie. Ces hypothèses nécessitent des études cliniques prospectives, pendant lesquelles sera évaluée l'interaction entre l'humain et la technologie proposée, qui sera un élément crucial de son utilité.

À cette fin, une autre limitation importante de la thèse est que l'évaluation des modèles ne tient pas comptes des biais socioculturels présents dans les données. La nature rétrospective du jeu de données limite la quantité de variable collectée, et certaines informations comme l'origine ethnique, le statut socio-économique et le genre ne sont pas disponibles d'emblée dans les dossiers médicaux. Malheureusement, ces facteurs cliniques peuvent influencer les décisions cliniques et engendrer des biais dans la prise en charge des patients [285], [286], [287]. Les algorithmes entraînés sur ces données biaisées ont tendance à reconduire ces biais dans leurs prédictions [288]. Il est donc important d'incorporer un processus de reconnaissance et mitigation des biais lors du déploiement des modèles dans de futures études cliniques [289], [290].

L'EEG est un outil fondamental dans la prise en charge de l'épilepsie, mais son interprétation est basée sur l'analyse visuelle. L'apprentissage profond a révolutionné notre société par sa capacité à

modéliser des signaux complexes comme les images, le son et le langage. L'application de l'apprentissage profond à l'EEG est certainement une étape clé de l'évolution de la neurologie clinique. L'impact réel de cette nouvelle technologie dépendra d'une validation robuste et d'une réflexion approfondie sur l'interaction entre les humains (patients et cliniciens) et l'intelligence artificielle.

RÉFÉRENCES

[1] A. K. Ngugi, C. Bottomley, I. Kleinschmidt, J. W. Sander, and C. R. Newton, 'Estimation of the burden of active and life-time epilepsy: a meta-analytic approach.', *Epilepsia*, vol. 51, no. 5, pp. 883–890, May 2010, doi: 10.1111/j.1528-1167.2009.02481.x.

[2] R. S. Fisher *et al.*, 'ILAE Official Report: A practical clinical definition of epilepsy', *Epilepsia*, vol. 55, no. 4, pp. 475–482, Apr. 2014, doi: 10.1111/epi.12550.

[3] R. S. Fisher *et al.*, 'Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE).', *Epilepsia*, vol. 46, no. 4, pp. 470–472, Apr. 2005, doi: 10.1111/j.0013-9580.2005.66104.x.

[4] O. Devinsky, D. C. Hesdorffer, D. J. Thurman, S. Lhatoo, and G. Richerson, 'Sudden unexpected death in epilepsy: epidemiology, mechanisms, and prevention', *The Lancet Neurology*, vol. 15, no. 10, pp. 1075–1088, Sep. 2016, doi: 10.1016/S1474-4422(16)30158-2.

[5] O. Devinsky, T. Spruill, D. Thurman, and D. Friedman, 'Recognizing and preventing epilepsy-related mortality: A call for action.', *Neurology*, vol. 86, no. 8, pp. 779–786, Feb. 2016, doi: 10.1212/WNL.0000000000002253.

[6] D. L. Schomer and F. H. Lopes da Silva, *Niedermeyer's ElectroencephalographyBasic Principles, Clinical Applications, and Related Fields: Basic Principles, Clinical Applications, and Related Fields*. Oxford, UK: Oxford University Press, 2017. doi: 10.1093/med/9780190228484.001.0001.

[7] E. W. Lothman, 'The Neurobiology of Epileptiform Discharges', *American Journal of EEG Technology*, vol. 33, no. 2, pp. 93–112, Jun. 1993, doi: 10.1080/00029238.1993.11080438.

[8] É. Lemoine *et al.*, 'Prognostic value of interictal epileptiform discharges on routine EEG in adults with epilepsy', *Epilepsia*, vol. n/a, no. n/a, 2025, doi: 10.1111/epi.18318.

[9] R. N. McGinty, D. J. Costello, P. Kinirons, and B. McNamara, 'Diagnostic Yield of Routine EEG in Adults with Active Epilepsy', *Ir Med J*, vol. 112, no. 1, p. 851, Jan. 2019.

[10] E. Baldin, W. A. Hauser, J. R. Buchhalter, D. C. Hesdorffer, and R. Ottman, 'Yield of epileptiform electroencephalogram abnormalities in incident unprovoked seizures: A population-based study', *Epilepsia*, vol. 55, no. 9, pp. 1389–1398, 2014, doi: 10.1111/epi.12720.

[11] A. Aschner, C. Kowal, O. Arski, J. A. G. Crispo, N. Farhat, and E. Donner, 'Prevalence of epileptiform electroencephalographic abnormalities in people without a history of seizures: A systematic review and meta-analysis', *Epilepsia*, vol. 65, no. 3, pp. 583–599, 2024, doi: 10.1111/epi.17864.

[12] S. R. Benbadis and W. O. Tatum, 'Overintepretation of EEGs and Misdiagnosis of Epilepsy', *Journal of Clinical Neurophysiology*, vol. 20, no. 1, p. 42, Feb. 2003.

[13] F. A. Chowdhury, L. Nashef, and R. D. C. Elwes, 'Misdiagnosis in epilepsy: a review and recognition of diagnostic uncertainty', *European Journal of Neurology*, vol. 15, no. 10, pp. 1034–1042, 2008, doi: 10.1111/j.1468-1331.2008.02260.x.

[14] J. Y. Kang and G. L. Krauss, ‘Normal Variants Are Commonly Overread as Interictal Epileptiform Abnormalities.’, *J Clin Neurophysiol*, vol. 36, no. 4, pp. 257–263, Jul. 2019, doi: 10.1097/WNP.0000000000000613.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘ImageNet Classification with Deep Convolutional Neural Networks’, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2012. Accessed: Jun. 19, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[16] G. Hinton *et al.*, ‘Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups’, *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.

[17] A. Vaswani *et al.*, ‘Attention Is All You Need’, Jun. 2017, Accessed: Feb. 15, 2022. [Online]. Available: https://arxiv.org/abs/1706.03762v5

[18] J. Vamathevan *et al.*, ‘Applications of machine learning in drug discovery and development’, *Nat Rev Drug Discov*, vol. 18, no. 6, pp. 463–477, Jun. 2019, doi: 10.1038/s41573-019-0024-5.

[19] J. Jumper *et al.*, ‘Highly accurate protein structure prediction with AlphaFold’, *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.

[20] J. Abramson *et al.*, ‘Accurate structure prediction of biomolecular interactions with AlphaFold 3’, *Nature*, vol. 630, no. 8016, pp. 493–500, Jun. 2024, doi: 10.1038/s41586-024-07487-w.

[21] J. Chvojka *et al.*, ‘The role of interictal discharges in ictogenesis — A dynamical perspective’, *Epilepsy & Behavior*, vol. 121, p. 106591, Aug. 2021, doi: 10.1016/j.yebeh.2019.106591.

[22] X.-J. Wang, ‘Neurophysiological and Computational Principles of Cortical Rhythms in Cognition’, *Physiological Reviews*, vol. 90, no. 3, pp. 1195–1268, Jul. 2010, doi: 10.1152/physrev.00035.2008.

[23] W. A. Hauser and E. Beghi, ‘First seizure definitions and worldwide incidence and mortality.’, *Epilepsia*, vol. 49 Suppl 1, pp. 8–12, 2008, doi: 10.1111/j.1528-1167.2008.01443.x.

[24] A. T. Berg, ‘Risk of recurrence after a first unprovoked seizure’, *Epilepsia*, vol. 49, no. s1, pp. 13–18, Jan. 2008, doi: 10.1111/j.1528-1167.2008.01444.x.

[25] W. A. Hauser, S. S. Rich, J. R.-J. Lee, J. F. Annegers, and V. E. Anderson, ‘Risk of Recurrent Seizures after Two Unprovoked Seizures’, *N Engl J Med*, vol. 338, no. 7, pp. 429–434, Feb. 1998, doi: 10.1056/NEJM199802123380704.

[26] A. Krumholz *et al.*, ‘Practice Parameter: evaluating an apparent unprovoked first seizure in adults (an evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology and the American Epilepsy Society.’, *Neurology*, vol. 69, no. 21, pp. 1996–2007, Nov. 2007, doi: 10.1212/01.wnl.0000285084.93652.43.

[27] D. Hirtz *et al.*, ‘Practice parameter: evaluating a first nonfebrile seizure in children: report of the quality standards subcommittee of the American Academy of Neurology, The Child

Neurology Society, and The American Epilepsy Society.', *Neurology*, vol. 55, no. 5, pp. 616–623, Sep. 2000, doi: 10.1212/wnl.55.5.616.

[28] J. Pillai and M. R. Sperling, 'Interictal EEG and the Diagnosis of Epilepsy', *Epilepsia*, vol. 47, no. s1, pp. 14–22, Oct. 2006, doi: 10.1111/j.1528-1167.2006.00654.x.

[29] M. A. Kural *et al.*, 'Criteria for defining interictal epileptiform discharges in EEG: A clinical validation study', *Neurology*, vol. 94, no. 20, pp. e2139–e2147, May 2020, doi: 10.1212/WNL.0000000000009439.

[30] H. K. Bouma, C. Labos, G. C. Gore, C. Wolfson, and M. R. Keezer, 'The diagnostic accuracy of routine electroencephalography after a first unprovoked seizure', *European Journal of Neurology*, vol. 23, no. 3, pp. 455–463, Mar. 2016, doi: 10.1111/ene.12739.

[31] W. O. Tatum *et al.*, 'Clinical utility of EEG in diagnosing and monitoring epilepsy in adults', *Clinical Neurophysiology*, vol. 129, no. 5, pp. 1056–1082, May 2018, doi: 10.1016/j.clinph.2018.01.019.

[32] S. Gallotto and M. Seeck, 'EEG biomarker candidates for the identification of epilepsy', *Clin Neurophysiol Pract*, vol. 8, pp. 32–41, Dec. 2022, doi: 10.1016/j.cnp.2022.11.004.

[33] S. J. M. Smith, 'EEG in the diagnosis, classification, and management of patients with epilepsy', *J Neurol Neurosurg Psychiatry*, vol. 76, no. suppl 2, pp. ii2–ii7, 2005, doi: 10.1136/jnnp.2005.069245.

[34] DouglasS. Goodin and MichaelJ. Aminoff, 'Does the interictal EEG have a role in the diagnosis of epilepsy?', *The Lancet*, vol. 323, no. 8381, pp. 837–839, Apr. 1984, doi: 10.1016/S0140-6736(84)92281-5.

[35] F. A. Nascimento *et al.*, 'Focal epilepsies: Update on diagnosis and classification', *Epileptic Disorders*, vol. 25, no. 1, pp. 1–17, 2023, doi: 10.1002/epd2.20045.

[36] M. H. Libenson, '2 - Visual Analysis of the EEG: Wakefulness, Drowsiness, and Sleep', in *Practical Approach to Electroencephalography (Second Edition)*, M. H. Libenson, Ed., Philadelphia: Elsevier, 2025, pp. 7–34. doi: 10.1016/B978-0-323-39229-7.00002-0.

[37] J. Jing *et al.*, 'Interrater Reliability of Experts in Identifying Interictal Epileptiform Discharges in Electroencephalograms', *JAMA Neurology*, vol. 77, no. 1, pp. 49–57, Jan. 2020, doi: 10.1001/jamaneurol.2019.3531.

[38] F. A. Nascimento *et al.*, 'One EEG, one read – A manifesto towards reducing interrater variability among experts', *Clinical Neurophysiology*, vol. 133, pp. 68–70, Jan. 2022, doi: 10.1016/j.clinph.2021.10.007.

[39] A. S. Greenblatt, S. Beniczky, and F. A. Nascimento, 'Pitfalls in scalp EEG: Current obstacles and future directions', *Epilepsy & Behavior*, vol. 149, p. 109500, Dec. 2023, doi: 10.1016/j.yebeh.2023.109500.

[40] G. Slinger *et al.*, 'Clinical characteristics and diagnoses of 1213 children referred to a first seizure clinic', *Epilepsia Open*, vol. 9, no. 2, pp. 548–557, 2024, doi: 10.1002/epi4.12883.

[41] Y. Li *et al.*, 'Early Access to First-Seizure Clinics, Subsequent Outcomes, and Factors Associated With Attendance', *JAMA Neurology*, vol. 81, no. 7, pp. 732–740, Jul. 2024, doi: 10.1001/jamaneurol.2024.1187.

[42] A. M. McIntosh *et al.*, 'Newly diagnosed seizures assessed at two established first seizure clinics: Clinic characteristics, investigations, and findings over 11 years.', *Epilepsia Open*, vol. 6, no. 1, pp. 171–180, Mar. 2021, doi: 10.1002/epi4.12460.

[43] A. Jackson, L. Teo, and U. Seneviratne, 'Challenges in the first seizure clinic for adult patients with epilepsy', *Epileptic Disorders*, vol. 18, no. 3, pp. 305–314, Sep. 2016, doi: 10.1684/epd.2016.0853.

[44] P. Kwan and M. J. Brodie, 'Early Identification of Refractory Epilepsy', *New England Journal of Medicine*, vol. 342, no. 5, pp. 314–319, Feb. 2000, doi: 10.1056/NEJM200002033420503.

[45] L. G. Kim, T. L. Johnson, A. G. Marson, and D. W. Chadwick, 'Prediction of risk of seizure recurrence after a single seizure and early epilepsy: further results from the MESS trial', *The Lancet Neurology*, vol. 5, no. 4, pp. 317–322, Apr. 2006, doi: 10.1016/S1474-4422(06)70383-0.

[46] B. Sultana *et al.*, 'Incidence and Prevalence of Drug-Resistant Epilepsy', *Neurology*, vol. 96, no. 17, p. 805, Apr. 2021, doi: 10.1212/WNL.0000000000011839.

[47] P. J. Karoly *et al.*, 'Cycles in epilepsy', *Nature Reviews Neurology*, vol. 17, no. 5, pp. 267–284, May 2021, doi: 10.1038/s41582-021-00464-1.

[48] M. G. Leguia *et al.*, 'Seizure Cycles in Focal Epilepsy', *JAMA Neurology*, vol. 78, no. 4, pp. 454–463, Apr. 2021, doi: 10.1001/jamaneurol.2020.5370.

[49] R. S. Fisher, 'Bad information in epilepsy care', *Epilepsy Behav*, vol. 67, pp. 133–134, Feb. 2017, doi: 10.1016/j.yebeh.2016.10.022.

[50] J. Engel Jr. *et al.*, 'Epilepsy biomarkers', *Epilepsia*, vol. 54, no. s4, pp. 61–69, 2013, doi: 10.1111/epi.12299.

[51] J. Buchhalter *et al.*, 'EEG parameters as endpoints in epilepsy clinical trials - An expert panel opinion paper', *Epilepsy Research*, vol. 187, p. 107028, Nov. 2022, doi: 10.1016/j.eplepsyres.2022.107028.

[52] K. M. Sauro *et al.*, 'The current state of epilepsy guidelines: A systematic review', *Epilepsia*, vol. 57, no. 1, pp. 13–23, Jan. 2016, doi: 10.1111/epi.13273.

[53] A. Pitkänen *et al.*, 'Advances in the development of biomarkers for epilepsy', *The Lancet Neurology*, vol. 15, no. 8, pp. 843–856, Jul. 2016, doi: 10.1016/S1474-4422(16)00112-5.

[54] R. A. B. Badawy, A. S. Harvey, and R. A. L. Macdonell, 'Cortical hyperexcitability and epileptogenesis: Understanding the mechanisms of epilepsy – Part 1', *Journal of Clinical Neuroscience*, vol. 16, no. 3, pp. 355–365, Mar. 2009, doi: 10.1016/j.jocn.2008.08.026.

[55] 'Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies', *The Lancet Neurology*, vol. 13, no. 9, pp. 893–903, Sep. 2014, doi: 10.1016/S1474-4422(14)70171-1.

[56] R. Stevelink *et al.*, 'GWAS meta-analysis of over 29,000 people with epilepsy identifies 26 risk loci and subtype-specific genetic architecture', *Nat Genet*, vol. 55, no. 9, pp. 1471–1482, Sep. 2023, doi: 10.1038/s41588-023-01485-w.

[57] H. O. Heyne *et al.*, 'Polygenic risk scores as a marker for epilepsy risk across lifetime and after unspecified seizure events', *Nat Commun*, vol. 15, no. 1, p. 6277, Jul. 2024, doi: 10.1038/s41467-024-50295-z.

[58] J. Wang *et al.*, 'Circulating microRNAs are promising novel biomarkers for drug-resistant epilepsy', *Sci Rep*, vol. 5, p. 10201, May 2015, doi: 10.1038/srep10201.

[59] J. Wang *et al.*, 'Genome-wide circulating microRNA expression profiling indicates biomarkers for epilepsy', *Sci Rep*, vol. 5, p. 9522, Mar. 2015, doi: 10.1038/srep09522.

[60] W. Lai, D. Du, and L. Chen, 'Metabolomics Provides Novel Insights into Epilepsy Diagnosis and Treatment: A Review', *Neurochem Res*, vol. 47, no. 4, pp. 844–859, Apr. 2022, doi: 10.1007/s11064-021-03510-y.

[61] S. Larivière, A. Bernasconi, N. Bernasconi, and B. C. Bernhardt, 'Connectome biomarkers of drug-resistant epilepsy', *Epilepsia*, vol. 62, no. 1, pp. 6–24, 2021, doi: 10.1111/epi.16753.

[62] E. A. van Vliet *et al.*, 'WONOEP appraisal: Imaging biomarkers in epilepsy', *Epilepsia*, vol. 58, no. 3, pp. 315–330, 2017, doi: 10.1111/epi.13621.

[63] R. Mito, J. Cole, S. Genc, G. Jackson, and A. Zalesky, 'Towards precision MRI biomarkers in epilepsy with normative modelling', *Brain*, p. awaf090, Mar. 2025, doi: 10.1093/brain/awaf090.

[64] J. DellaBadia Jr, W. L. Bell, J. W. Keyes Jr, V. P. Mathews, and S. S. Glazier, 'Assessment and cost comparison of sleep-deprived EEG, MRI and PET in the prediction of surgical treatment for epilepsy', *Seizure*, vol. 11, no. 5, pp. 303–309, Jul. 2002, doi: 10.1053/seiz.2001.0648.

[65] S. Abdelhady, H. Shokri, M. Fathy, and mona M. wahid el din, 'Evaluation of the direct costs of epilepsy in a sample of Egyptian patients following up in Ain Shams University Hospital', *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, vol. 56, no. 1, p. 112, Nov. 2020, doi: 10.1186/s41983-020-00247-3.

[66] D. Velis, P. Plouin, J. Gotman, F. L. Da Silva, and members of the ILAE DMC Subcommittee on Neurophysiology, 'Recommendations Regarding the Requirements and Applications for Long-term Recordings in Epilepsy', *Epilepsia*, vol. 48, no. 2, pp. 379–384, Feb. 2007, doi: 10.1111/j.1528-1167.2007.00920.x.

[67] É. Lemoine *et al.*, 'Computer-assisted analysis of routine EEG to identify hidden biomarkers of epilepsy: A systematic review', *Computational and Structural Biotechnology Journal*, vol. 24, pp. 66–86, Dec. 2024, doi: 10.1016/j.csbj.2023.12.006.

[68] É. Lemoine *et al.*, 'Computer-assisted analysis of routine electroencephalogram to identify hidden biomarkers of epilepsy: protocol for a systematic review', *BMJ Open*, vol. 13, no. 1, p. e066932, Jan. 2023, doi: 10.1136/bmjopen-2022-066932.

[69] L. Douw *et al.*, '"Functional connectivity" is a sensitive predictor of epilepsy diagnosis after the first seizure', *PLoS ONE [Electronic Resource]*, vol. 5, no. 5, p. e10839, 2010, doi: 10.1371/journal.pone.0010839.

[70] A. Tripathi, G. Srivastava, and P. K. Maurya, 'Epileptic Data Classification Using Frequency Power Estimation of Channel (FP1-F7) in Children', in *2018 7th International Conference on*

Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2018, pp. 501–505. doi: 10.1109/icrito.2018.8748624.

[71] M. Cabrerizo, M. Ayala, M. Goryawala, P. Jayakar, and M. Adjouadi, 'A new parametric feature descriptor for the classification of epileptic and control EEG records in pediatric population', *Int J Neural Syst*, vol. 22, no. 2, p. 1250001, Apr. 2012, doi: 10.1142/S0129065712500013.

[72] T. A. Cassar, K. P. Camilleri, S. G. Fabri, M. Zervakis, and S. Micheloyannis, 'ARMA modeling for the diagnosis of controlled epileptic activity in young children', in *2008 3rd International Symposium on Communications, Control and Signal Processing*, Mar. 2008, pp. 25–30. doi: 10.1109/ISCCSP.2008.4537186.

[73] M. C. Guerrero, J. S. Parada, and H. E. Espitia, 'EEG signal analysis using classification techniques: Logistic regression, artificial neural networks, support vector machines, and convolutional neural networks', *Heliyon*, vol. 7, no. 6, Jun. 2021, doi: 10.1016/j.heliyon.2021.e07258.

[74] T. Uyttenhove, A. Maes, T. V. Steenkiste, D. Deschrijver, and T. Dhaene, 'Interpretable Epilepsy Detection in Routine, Interictal EEG Data using Deep Learning', in *Proceedings of the Machine Learning for Health NeurIPS Workshop*, PMLR, Nov. 2020, pp. 355–366. Accessed: Feb. 16, 2022. [Online]. Available: https://proceedings.mlr.press/v136/uyttenhove20a.html

[75] F. S. Bao, J. Gao, J. Hu, D. Y. C. Lie, Y. Zhang, and K. J. Oommen, 'Automated epilepsy diagnosis using interictal scalp EEG', in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Sep. 2009, pp. 6603–6607. doi: 10.1109/IEMBS.2009.5332550.

[76] B. Hjorth, 'EEG analysis based on time domain properties', *Electroencephalography and Clinical Neurophysiology*, vol. 29, no. 3, pp. 306–310, Sep. 1970, doi: 10.1016/0013-4694(70)90143-4.

[77] D. Zelig *et al.*, 'Paroxysmal slow wave events predict epilepsy following a first seizure', *Epilepsia*, vol. 63, no. 1, pp. 190–198, 2022, doi: 10.1111/epi.17110.

[78] N. Ahmadi, Y. Pei, E. Carrette, A. P. Aldenkamp, and M. Pechenizkiy, 'EEG-based classification of epilepsy and PNES: EEG microstate and functional brain network features', *Brain Informatics*, vol. 7, no. 1, p. 6, 2020, doi: 10.1186/s40708-020-00107-z.

[79] K. Luo and D. Luo, 'An EEG feature-based diagnosis model for epilepsy', in *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, Oct. 2010, pp. V8-592-V8-594. doi: 10.1109/ICCASM.2010.5619259.

[80] J. A. Urigüen, B. García-Zapirain, J. Artieda, J. Iriarte, and M. Valencia, 'Comparison of background EEG activity of different groups of patients with idiopathic epilepsy using Shannon spectral entropy and cluster-based permutation statistical testing', *PLOS ONE*, vol. 12, no. 9, p. e0184044, Sep. 2017, doi: 10.1371/journal.pone.0184044.

[81] S. Prabhu and R. J. Martis, 'Computer Aided Tool for diagnosing Epilepsy using Kolmogorov Complexity and Approximate Entropy', in *2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, 2020, pp. 69–73. doi: 10.1109/discover50404.2020.9278044.

[82] Z. Yang, Y. Wang, and G. Ouyang, 'Adaptive neuro-fuzzy inference system for classification of background EEG signals from ESES patients and controls', *Thescientificworldjournal*, vol. 2014, p. 140863, 2014, doi: 10.1155/2014/140863.

[83] W. J. Bosl, T. Loddenkemper, and C. A. Nelson, 'Nonlinear EEG biomarker profiles for autism and absence epilepsy', *Neuropsychiatric Electrophysiology*, vol. 3, no. 1, p. 1, Mar. 2017, doi: 10.1186/s40810-017-0023-x.

[84] M. N. Tibdewal and S. A. Tale, 'Extraction and performance analysis of multi-domain novel features for classification and detection of epileptic EEG', in *2017 International Conference on Big Data, IoT and Data Science (BID)*, 2017, pp. 91–96. doi: 10.1109/bid.2017.8336579.

[85] J. Pyrzowski, M. Sieminski, A. Sarnowska, J. Jedrzejczak, and W. M. Nyka, 'Interval analysis of interictal EEG: pathology of the alpha rhythm in focal epilepsy', *Scientific Reports*, vol. 5, Art. no. 1, 2015, doi: 10.1038/srep16230.

[86] J. Cao *et al.*, 'Using interictal seizure-free EEG data to recognise patients with epilepsy based on machine learning of brain functional connectivity', *Biomedical Signal Processing and Control*, vol. 67 (no pagination), 2021, doi: 10.1016/j.bspc.2021.102554.

[87] B. Rijnders, E. E. Korkmaz, and F. Yildirim, 'CNN for a Connectivity Based Epilepsy Diagnosis with Resting-state EEG', Sep. 27, 2021. doi: 10.1101/2021.09.24.21264004.

[88] H. Schmidt *et al.*, 'A computational biomarker of idiopathic generalized epilepsy from resting state EEG', *Epilepsia*, vol. 57, no. 10, pp. e200–e204, Oct. 2016, doi: 10.1111/epi.13481.

[89] A. Dasgupta, R. Das, L. Nayak, and R. K. De, 'Analyzing epileptogenic brain connectivity networks using clinical EEG data', in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2015, pp. 815–821. doi: 10.1109/BIBM.2015.7359791.

[90] S. Sargolzaei *et al.*, 'A probabilistic approach for pediatric epilepsy diagnosis using brain functional connectivity networks', *BMC Bioinformatics*, vol. 16 Suppl 7, p. S9, 2015, doi: 10.1186/1471-2105-16-S7-S9.

[91] S. Sargolzaei, M. Cabrerizo, M. Goryawala, A. S. Eddin, and M. Adjouadi, 'Scalp EEG brain functional connectivity networks in pediatric epilepsy', *Computers in Biology & Medicine*, vol. 56, pp. 158–166, 2015, doi: 10.1016/j.compbiomed.2014.10.018.

[92] S. Sargolzaei, M. Cabrerizo, M. Goryawala, A. S. Eddin, and M. Adjouadi, 'Functional connectivity network based on graph analysis of scalp EEG for epileptic classification', in *2013 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2013, pp. 1–4. doi: 10.1109/spmb.2013.6736779.

[93] H. Schmidt, G. Petkov, M. P. Richardson, and J. R. Terry, 'Dynamics on networks: the role of local dynamics and global networks on the emergence of hypersynchronous neural activity', *PLoS Comput Biol*, vol. 10, no. 11, pp. e1003947–e1003947, Nov. 2014, doi: 10.1371/journal.pcbi.1003947.

[94] W. A. Chaovalitwongse, R. S. Pottenger, S. Wang, Y. Fan, and L. D. Iasemidis, 'Pattern- and Network-Based Classification Techniques for Multichannel Medical Data Signals to Improve Brain Diagnosis', *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 41, no. 5, pp. 977–988, 2011, doi: 10.1109/tsmca.2011.2106118.

[95] Y. J. Fan, W. A. Chaovalitwongse, C. C. Liu, R. C. Sachdeo, L. Iasemidis, and P. Pardalos, 'Optimisation and data mining techniques for the screening of epileptic patients', *International Journal of Bioinformatics Research & Applications*, vol. 5, no. 2, pp. 187–196, 2009.

[96] E. Mazzucchi *et al.*, 'Hyperventilation in patients with focal epilepsy: Electromagnetic tomography, functional connectivity and graph theory: A possible tool in epilepsy diagnosis?', *Journal of Clinical Neurophysiology*, vol. 34(1), pp. 92–99, 2017, doi: 10.1097/WNP.0000000000000329.

[97] C. Song *et al.*, 'A Feature Tensor-Based Epileptic Detection Model Based on Improved Edge Removal Approach for Directed Brain Networks', *Frontiers in Neuroscience*, vol. 14, p. 557095, 2020, doi: 10.3389/fnins.2020.557095.

[98] L.-C. Lin, C.-S. Ouyang, R.-C. Wu, R.-C. Yang, and C.-T. Chiang, 'Alternative Diagnosis of Epilepsy in Children Without Epileptiform Discharges Using Deep Convolutional Neural Networks', *Int. J. Neur. Syst.*, vol. 30, no. 05, p. 1850060, May 2020, doi: 10.1142/S0129065718500600.

[99] P. F. Whiting *et al.*, 'QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies.', *Ann Intern Med*, vol. 155, no. 8, pp. 529–536, Oct. 2011, doi: 10.7326/0003-4819-155-8-201110180-00009.

[100] P. Myers *et al.*, 'Diagnosing Epilepsy with Normal Interictal EEG Using Dynamic Network Models', *Ann Neurol*, vol. 97, no. 5, pp. 907–918, May 2025, doi: 10.1002/ana.27168.

[101] L. Tait *et al.*, 'Estimating the likelihood of epilepsy from clinically noncontributory electroencephalograms using computational analysis: A retrospective, multisite case–control study', *Epilepsia*, vol. 65, no. 8, pp. 2459–2469, 2024, doi: 10.1111/epi.18024.

[102] I. Faiman *et al.*, 'Limited clinical validity of univariate resting-state EEG markers for classifying seizure disorders', *Brain Communications*, vol. 5, no. 6, p. fcad330, Dec. 2023, doi: 10.1093/braincomms/fcad330.

[103] J. Tveit *et al.*, 'Automated Interpretation of Clinical Electroencephalograms Using Artificial Intelligence', *JAMA Neurology*, Jun. 2023, doi: 10.1001/jamaneurol.2023.1645.

[104] J. Jing *et al.*, 'Development of Expert-Level Automated Detection of Epileptiform Discharges During Electroencephalogram Interpretation', *JAMA Neurol*, Oct. 2019, doi: 10.1001/jamaneurol.2019.3485.

[105] 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, doi: 10.1088/1741-2552/ab260c.

[106] A. Craik, Y. He, and J. L. Contreras-Vidal, 'Deep learning for electroencephalogram (EEG) classification tasks: a review', *J. Neural Eng.*, vol. 16, no. 3, p. 031001, Apr. 2019, doi: 10.1088/1741-2552/ab0ab5.

[107] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into deep learning*. Cambridge University Press, 2023.

[108] M. Jafari *et al.*, 'Emotion recognition in EEG signals using deep learning methods: A review', *Computers in Biology and Medicine*, vol. 165, p. 107450, Oct. 2023, doi: 10.1016/j.compbiomed.2023.107450.

[109] H. Altaheri *et al.*, 'Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: a review', *Neural Comput & Applic*, vol. 35, no. 20, pp. 14681–14722, Jul. 2023, doi: 10.1007/s00521-021-06352-5.

[110] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[111] M. D. Zeiler and R. Fergus, 'Visualizing and Understanding Convolutional Networks', in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 818–833. doi: 10.1007/978-3-319-10590-1_53.

[112] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[113] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *CoRR*, Sep. 2014, Accessed: Jun. 19, 2025. [Online]. Available: https://www.semanticscholar.org/paper/Very-Deep-Convolutional-Networks-for-Large-Scale-Simonyan-Zisserman/eb42cf88027de515750f230b23b1a057dc782108

[114] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[115] S. L. Smith, A. Brock, L. Berrada, and S. De, 'ConvNets Match Vision Transformers at Scale', Oct. 25, 2023, *arXiv*: arXiv:2310.16764. doi: 10.48550/arXiv.2310.16764.

[116] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, 'A ConvNet for the 2020s', *arXiv:2201.03545 [cs]*, Mar. 2022, Accessed: Mar. 27, 2022. [Online]. Available: http://arxiv.org/abs/2201.03545

[117] M. Tan and Q. Le, 'EfficientNetV2: Smaller Models and Faster Training', in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 10096–10106. Accessed: Jun. 19, 2025. [Online]. Available: https://proceedings.mlr.press/v139/tan21a.html

[118] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, 'EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces', *J. Neural Eng.*, vol. 15, no. 5, p. 056013, Jul. 2018, doi: 10.1088/1741-2552/aace8c.

[119] R. T. Schirrmeister *et al.*, 'Deep learning with convolutional neural networks for EEG decoding and visualization', *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017, doi: 10.1002/hbm.23730.

[120] L. A. W. Gemein *et al.*, 'Machine-learning-based diagnostics of EEG pathology', *NeuroImage*, vol. 220, p. 117021, Oct. 2020, doi: 10.1016/j.neuroimage.2020.117021.

[121] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, 'Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface', in *2008 IEEE International Joint Conference on*

Neural Networks (IEEE World Congress on Computational Intelligence), Jun. 2008, pp. 2390–2397. doi: 10.1109/IJCNN.2008.4634130.

[122] A.-K. Kiessner, R. T. Schirrmeister, J. Boedecker, and T. Ball, 'Reaching the ceiling? Empirical scaling behaviour for deep EEG pathology classification', *Comput Biol Med*, vol. 178, p. 108681, Aug. 2024, doi: 10.1016/j.compbiomed.2024.108681.

[123] I. Obeid and J. Picone, 'The Temple University Hospital EEG Data Corpus', *Frontiers in Neuroscience*, vol. 10, p. 196, 2016, doi: 10.3389/fnins.2016.00196.

[124] A.-K. Kiessner, R. T. Schirrmeister, L. A. W. Gemein, J. Boedecker, and T. Ball, 'An extended clinical EEG dataset with 15,300 automatically labelled recordings for pathology decoding', *NeuroImage: Clinical*, vol. 39, p. 103482, Jan. 2023, doi: 10.1016/j.nicl.2023.103482.

[125] R. Schirrmeister, L. Gemein, K. Eggensperger, F. Hutter, and T. Ball, 'Deep learning with convolutional neural networks for decoding and visualization of EEG pathology', in *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, Dec. 2017, pp. 1–7. doi: 10.1109/SPMB.2017.8257015.

[126] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, 'Scaling Vision Transformers', Jun. 20, 2022, *arXiv*: arXiv:2106.04560. doi: 10.48550/arXiv.2106.04560.

[127] Z. Wan, M. Li, S. Liu, J. Huang, H. Tan, and W. Duan, 'EEGformer: A transformer–based brain activity classification method using EEG signal', *Frontiers in Neuroscience*, vol. 17, 2023, Accessed: Oct. 11, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2023.1148855

[128] R. Yang and E. Modesitt, 'ViT2EEG: Leveraging Hybrid Pretrained Vision Transformers for EEG Data', Aug. 01, 2023, *arXiv*: arXiv:2308.00454. Accessed: Aug. 22, 2024. [Online]. Available: http://arxiv.org/abs/2308.00454

[129] Y. Song, Q. Zheng, B. Liu, and X. Gao, 'EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2023, doi: 10.1109/TNSRE.2022.3230250.

[130] A. Dosovitskiy *et al.*, 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', *arXiv:2010.11929 [cs]*, Jun. 2021, Accessed: Jan. 29, 2022. [Online]. Available: http://arxiv.org/abs/2010.11929

[131] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, 'Early Convolutions Help Transformers See Better', Jun. 28, 2021. Accessed: Mar. 28, 2024. [Online]. Available: https://arxiv.org/abs/2106.14881v3

[132] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, 'Escaping the Big Data Paradigm with Compact Transformers', Jun. 07, 2022, *arXiv*: arXiv:2104.05704. Accessed: Sep. 30, 2023. [Online]. Available: http://arxiv.org/abs/2104.05704

[133] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, 'How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers', *arXiv:2106.10270 [cs]*, Jun. 2021, Accessed: Mar. 03, 2022. [Online]. Available: http://arxiv.org/abs/2106.10270

[134] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, 'RandAugment: Practical automated data augmentation with a reduced search space', Nov. 13, 2019, *arXiv*: arXiv:1909.13719. doi: 10.48550/arXiv.1909.13719.

[135] Y.-H. Cao, H. Yu, and J. Wu, 'Training Vision Transformers with Only 2040 Images', Jan. 26, 2022, arXiv:2201.10728. doi: 10.48550/arXiv.2201.10728.

[136] R. Hussein, S. Lee, and R. Ward, 'Multi-Channel Vision Transformer for Epileptic Seizure Prediction', *Biomedicines*, vol. 10, no. 7, p. 1551, Jun. 2022, doi: 10.3390/biomedicines10071551.

[137] Z. Deng, C. Li, R. Song, X. Liu, R. Qian, and X. Chen, 'EEG-based seizure prediction via hybrid vision transformer and data uncertainty learning', *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106401, Aug. 2023, doi: 10.1016/j.engappai.2023.106401.

[138] M. J. D. Bayazi *et al.*, 'General-Purpose Brain Foundation Models for Time-Series Neuroimaging Data'.

[139] U. Amin and S. R. Benbadis, 'The Role of EEG in the Erroneous Diagnosis of Epilepsy', *J Clin Neurophysiol*, vol. 36, no. 4, pp. 294–297, 2019.

[140] D. Chadwick and D. Smith, 'The misdiagnosis of epilepsy', *BMJ*, vol. 324, no. 7336, pp. 495–496, Mar. 2002.

[141] U. Seneviratne, M. Cook, and W. D'Souza, 'The electroencephalogram of idiopathic generalized epilepsy', *Epilepsia*, vol. 53, no. 2, pp. 234–248, 2012, doi: 10.1111/j.1528-1167.2011.03344.x.

[142] U. Seneviratne, R. C. Boston, M. Cook, and W. D'Souza, 'EEG correlates of seizure freedom in genetic generalized epilepsies', *Neurology: Clinical Practice*, vol. 7, no. 1, pp. 35–44, Feb. 2017, doi: 10.1212/CPJ.0000000000000323.

[143] M. Guida, A. Iudice, E. Bonanni, and F. S. Giorgi, 'Effects of antiepileptic drugs on interictal epileptiform discharges in focal epilepsies: an update on current evidence', *Expert Review of Neurotherapeutics*, vol. 15, no. 8, pp. 947–959, Aug. 2015, doi: 10.1586/14737175.2015.1065180.

[144] V. Arntsen, T. Sand, M. R. Syvertsen, and E. Brodtkorb, 'Prolonged epileptiform EEG runs are associated with persistent seizures in juvenile myoclonic epilepsy', *Epilepsy Research*, vol. 134, pp. 26–32, Aug. 2017, doi: 10.1016/j.eplepsyres.2017.05.003.

[145] U. R. Acharya, S. Vinitha Sree, G. Swapna, R. J. Martis, and J. S. Suri, 'Automated EEG analysis of epilepsy: A review', *Knowledge-Based Systems*, vol. 45, pp. 147–165, Jun. 2013, doi: 10.1016/j.knosys.2013.02.014.

[146] W. Woldman *et al.*, 'Dynamic network properties of the interictal brain determine whether seizures appear focal or generalised', *Scientific Reports*, vol. 10, no. 1, Art. no. 1, 2020, doi: 10.1038/s41598-020-63430-9.

[147] F. A. Chowdhury *et al.*, 'Revealing a Brain Network Endophenotype in Families with Idiopathic Generalised Epilepsy', *PLOS ONE*, vol. 9, no. 10, p. e110136, Oct. 2014, doi: 10.1371/journal.pone.0110136.

[148] Y. Varatharajah *et al.*, 'Quantitative analysis of visually reviewed normal scalp EEG predicts seizure freedom following anterior temporal lobectomy', *Epilepsia*, vol. 63, no. 7, pp. 1630–1642, Jul. 2022, doi: 10.1111/epi.17257.

[149] E. Abela *et al.*, 'Slower alpha rhythm associates with poorer seizure control in epilepsy', *Annals of Clinical and Translational Neurology*, vol. 6(2), pp. 333–343, 2019, doi: 10.1002/acn3.710.

[150] P. G. Larsson and H. Kostov, 'Lower frequency variability in the alpha activity in EEG among patients with epilepsy.', *Clin Neurophysiol*, vol. 116, no. 11, pp. 2701–2706, Nov. 2005, doi: 10.1016/j.clinph.2005.07.019.

[151] E. J. Pegg, J. R. Taylor, and R. Mohanraj, 'Spectral power of interictal EEG in the diagnosis and prognosis of idiopathic generalized epilepsies', *Epilepsy & Behavior*, vol. 112, p. 107427, Nov. 2020, doi: 10.1016/j.yebeh.2020.107427.

[152] P. G. Larsson, O. Eeg-Olofsson, and G. Lantz, 'Alpha frequency estimation in patients with epilepsy', *Clinical EEG and Neuroscience*, vol. 43(2), pp. 97–104, 2012, doi: 10.1177/1550059411433611.

[153] T. Miyauchi, K. Endo, T. Yamaguchi, and H. Hagimoto, 'Computerized analysis of EEG background activity in epileptic patients', *Epilepsia*, vol. 32, no. 6, pp. 870–881, 1991.

[154] G. F. Diaz *et al.*, 'Generalized background qEEG abnormalities in localized symptomatic epilepsy', *Electroencephalography and Clinical Neurophysiology*, vol. 106(6), pp. 501–507, 1998, doi: 10.1016/S0013-4694%2898%2900026-1.

[155] A. Sathyanarayana *et al.*, 'Measuring the effects of sleep on epileptogenicity with multifrequency entropy', *Clinical Neurophysiology*, vol. 132, no. 9, pp. 2012–2018, Sep. 2021, doi: 10.1016/j.clinph.2021.06.001.

[156] I. Faiman, S. Smith, J. Hodsoll, A. H. Young, and P. Shotbolt, 'Resting-state EEG for the diagnosis of idiopathic epilepsy and psychogenic nonepileptic seizures: A systematic review', *Epilepsy & Behavior*, vol. 121, p. 108047, Aug. 2021, doi: 10.1016/j.yebeh.2021.108047.

[157] J. Engel Jr, A. Bragin, and R. Staba, 'Nonictal EEG biomarkers for diagnosis and treatment', *Epilepsia Open*, vol. 3, no. S2, pp. 120–126, Dec. 2018, doi: 10.1002/epi4.12233.

[158] D. Dash *et al.*, 'Update on Minimal Standards for Electroencephalography in Canada: A Review by the Canadian Society of Clinical Neurophysiologists', *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques*, vol. 44, no. 6, pp. 631–642, 2017, doi: 10.1017/cjn.2017.217.

[159] M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, and A. Gramfort, 'Autoreject: Automated artifact rejection for MEG and EEG data', *NeuroImage*, vol. 159, pp. 417–429, Oct. 2017, doi: 10.1016/j.neuroimage.2017.06.030.

[160] T. Gandhi, B. K. Panigrahi, and S. Anand, 'A comparative study of wavelet families for EEG signal classification', *Neurocomputing*, vol. 74, no. 17, pp. 3051–3057, Oct. 2011, doi: 10.1016/j.neucom.2011.04.029.

[161] H. Zou and T. Hastie, 'Regularization and Variable Selection via the Elastic Net', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[162] G. Ke *et al.*, 'LightGBM: a highly efficient gradient boosting decision tree', in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 3149–3157.

[163] G. C. Cawley and N. L. C. Talbot, 'On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation', *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, Aug. 2010.

[164] Erin LeDell, Maya Petersen, and Mark van der Laan, 'Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates', *Electronic Journal of Statistics*, vol. 9, no. 1, pp. 1583–1607, Jan. 2015, doi: 10.1214/15-EJS1035.

[165] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, 'Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach', *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988, doi: 10.2307/2531595.

[166] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G.M. Moons, 'Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement', *Annals of Internal Medicine*, Jan. 2015, doi: 10.7326/M14-0697.

[167] S. Clarke *et al.*, 'Computer-assisted EEG diagnostic review for idiopathic generalized epilepsy', *Epilepsy & Behavior*, p. 106556, Oct. 2019, doi: 10.1016/j.yebeh.2019.106556.

[168] M. E. Drake, H. Padamadan, and S. A. Newell, 'Interictal quantitative EEG in epilepsy', *Seizure - European Journal of Epilepsy*, vol. 7, no. 1, pp. 39–42, Feb. 1998, doi: 10.1016/S1059-1311(98)90006-1.

[169] N. Mammone and F. C. Morabito, 'Analysis of absence seizure EEG via Permutation Entropy spatio-temporal clustering', *The 2011 International Joint Conference on Neural Networks*, pp. 1417–1422, 2011, doi: 10.1109/ijcnn.2011.6033390.

[170] J. G. Lijmer *et al.*, 'Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests', *JAMA*, vol. 282, no. 11, pp. 1061–1066, Sep. 1999, doi: 10.1001/jama.282.11.1061.

[171] M. S. Pepe, Z. Feng, H. Janes, P. M. Bossuyt, and J. D. Potter, 'Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design', *J Natl Cancer Inst*, vol. 100, no. 20, pp. 1432–1438, Oct. 2008, doi: 10.1093/jnci/djn326.

[172] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi, 'The myth of generalisability in clinical research and machine learning in health care', *The Lancet Digital Health*, vol. 2, no. 9, pp. e489–e492, Sep. 2020, doi: 10.1016/S2589-7500(20)30186-2.

[173] Allan Krumholz *et al.*, 'Evidence-based guideline: Management of an unprovoked first seizure in adults', *Neurology*, vol. 84, no. 16, p. 1705, Apr. 2015, doi: 10.1212/WNL.0000000000001487.

[174] D. Gloss *et al.*, 'Antiseizure Medication Withdrawal in Seizure-Free Patients: Practice Advisory Update Summary: Report of the AAN Guideline Subcommittee', *Neurology*, vol. 97, no. 23, pp. 1072–1081, Dec. 2021, doi: 10.1212/WNL.0000000000012944.

[175] M. F. Selvitelli, L. M. Walker, D. L. Schomer, and B. S. Chang, 'The relationship of interictal epileptiform discharges to clinical epilepsy severity: a study of routine electroencephalograms and review of the literature', *J Clin Neurophysiol*, vol. 27, no. 2, pp. 87–92, Apr. 2010, doi: 10.1097/WNP.0b013e3181d64b1e.

[176] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, and C. Lin, 'Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images', *NeuroImage*, vol. 60, no. 1, pp. 59–70, Mar. 2012, doi: 10.1016/j.neuroimage.2011.11.066.

[177] L. Jollans *et al.*, 'Quantifying performance of machine learning methods for neuroimaging data', *Neuroimage*, vol. 199, pp. 351–365, Oct. 2019, doi: 10.1016/j.neuroimage.2019.05.082.

[178] A. Jabès, G. Klencklen, P. Ruggeri, J.-P. Antonietti, P. Banta Lavenex, and P. Lavenex, 'Age-Related Differences in Resting-State EEG and Allocentric Spatial Working Memory Performance', *Frontiers in Aging Neuroscience*, vol. 13, 2021, Accessed: Nov. 27, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnagi.2021.704362

[179] W. T. Blume, 'Drug Effects on EEG', *Journal of Clinical Neurophysiology*, vol. 23, no. 4, pp. 306–311, Aug. 2006, doi: 10.1097/01.wnp.0000229137.94384.fa.

[180] V.-H. Nguyen Michel *et al.*, 'The yield of routine EEG in geriatric patients: A prospective hospital-based study', *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 40, no. 5, pp. 249–254, Nov. 2010, doi: 10.1016/j.neucli.2010.08.003.

[181] B. Bučková, M. Brunovský, M. Bareš, and J. Hlinka, 'Predicting Sex From EEG: Validity and Generalizability of Deep-Learning-Based Interpretable Classifier', *Frontiers in Neuroscience*, vol. 14, 2020, Accessed: Jan. 31, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2020.589303

[182] K. R. V *et al.*, 'Machine learning detects EEG microstate alterations in patients living with temporal lobe epilepsy', *Seizure*, vol. 61, pp. 8–13, Oct. 2018, doi: 10.1016/j.seizure.2018.07.007.

[183] T. Verhoeven *et al.*, 'Automated diagnosis of temporal lobe epilepsy in the absence of interictal spikes', *NeuroImage: Clinical*, vol. 17, pp. 10–15, Jan. 2018, doi: 10.1016/j.nicl.2017.09.021.

[184] C.-S. Ouyang, R.-C. Yang, R.-C. Wu, C.-T. Chiang, and L.-C. Lin, 'Determination of Antiepileptic Drugs Withdrawal Through EEG Hjorth Parameter Analysis', *Int. J. Neur. Syst.*, vol. 30, no. 11, p. 2050036, Aug. 2020, doi: 10.1142/S0129065720500367.

[185] M. A. Kural *et al.*, 'Accurate identification of EEG recordings with interictal epileptiform discharges using a hybrid approach: Artificial intelligence supervised by human experts', *Epilepsia*, vol. 63, no. 5, pp. 1064–1073, 2022, doi: 10.1111/epi.17206.

[186] T. J. Gelety, R. J. Burgess, M. E. Drake Jr., C. E. Ford, and M. E. Brown, 'Computerized spectral analysis of the interictal EEG in epilepsy', *Clinical Electroencephalography*, vol. 16, no. 2, pp. 94–97, 1985.

[187] Y. Varatharajah *et al.*, 'Characterizing the electrophysiological abnormalities in visually reviewed normal EEGs of drug-resistant focal epilepsy patients', *Brain Communications*, vol. 3, no. 2, p. fcab102, Apr. 2021, doi: 10.1093/braincomms/fcab102.

[188] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, 'Indications of nonlinear deterministic and finite-dimensional structures in time series of brain

electrical activity: Dependence on recording region and brain state', *Phys. Rev. E*, vol. 64, no. 6, p. 061907, Nov. 2001, doi: 10.1103/PhysRevE.64.061907.

[189] U. R. Acharya, H. Fujita, V. K. Sudarshan, S. Bhat, and J. E. W. Koh, 'Application of entropies for automated diagnosis of epilepsy using EEG signals: A review', *Knowledge-Based Systems*, vol. 88, pp. 85–96, Nov. 2015, doi: 10.1016/j.knosys.2015.08.004.

[190] B. Scheepers, P. Clough, and C. Pickles, 'The misdiagnosis of epilepsy: findings of a population study', *Seizure*, vol. 7, no. 5, pp. 403–406, Oct. 1998, doi: 10.1016/S1059-1311(05)80010-X.

[191] J. P. Leach, R. Lauder, A. Nicolson, and D. F. Smith, 'Epilepsy in the UK: Misdiagnosis, mistreatment, and undertreatment?: The Wrexham area epilepsy project', *Seizure*, vol. 14, no. 7, pp. 514–520, Oct. 2005, doi: 10.1016/j.seizure.2005.08.008.

[192] A. Aschner, C. Kowal, O. Arski, J. A. G. Crispo, N. Farhat, and E. Donner, 'Prevalence of epileptiform electroencephalographic abnormalities in people without a history of seizures: A systematic review and meta-analysis', *Epilepsia*, vol. 65, no. 3, pp. 583–599, 2024, doi: 10.1111/epi.17864.

[193] M. E. Peltola *et al.*, 'Routine and sleep EEG: Minimum recording standards of the International Federation of Clinical Neurophysiology and the International League Against Epilepsy', *Epilepsia*, vol. 64, no. 3, pp. 602–618, 2023, doi: 10.1111/epi.17448.

[194] G. H. Klem, H. O. Lüders, H. H. Jasper, and C. Elger, 'The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology', *Electroencephalogr Clin Neurophysiol Suppl*, vol. 52, pp. 3–6, 1999.

[195] J. J. Halford *et al.*, 'Characteristics of EEG Interpreters Associated With Higher Interrater Agreement', *Journal of Clinical Neurophysiology*, vol. 34, no. 2, 2017, [Online]. Available: https://journals.lww.com/clinicalneurophys/Fulltext/2017/03000/Characteristics_of_EEG_Interpreters_Associated.12.aspx

[196] E. Bagheri, J. Dauwels, B. C. Dean, C. G. Waters, M. B. Westover, and J. J. Halford, 'Interictal epileptiform discharge characteristics underlying expert interrater agreement', *Clinical Neurophysiology*, vol. 128, no. 10, pp. 1994–2005, Oct. 2017, doi: 10.1016/j.clinph.2017.06.252.

[197] L. Tait *et al.*, 'Estimating the likelihood of epilepsy from clinically noncontributory electroencephalograms using computational analysis: A retrospective, multisite case–control study', *Epilepsia*, vol. 65, no. 8, pp. 2459–2469, 2024, doi: 10.1111/epi.18024.

[198] É. Lemoine *et al.*, 'Machine-learning for the prediction of one-year seizure recurrence based on routine electroencephalography', *Scientific Reports*, vol. 13, no. 1, p. 12650, Aug. 2023, doi: 10.1038/s41598-023-39799-8.

[199] P. Thangavel *et al.*, 'Improving automated diagnosis of epilepsy from EEGs beyond IEDs', *J. Neural Eng.*, vol. 19, no. 6, p. 066017, Nov. 2022, doi: 10.1088/1741-2552/ac9c93.

[200] J. Thomas *et al.*, 'Automated Adult Epilepsy Diagnostic Tool Based on Interictal Scalp Electroencephalogram Characteristics: A Six-Center Study.', *Int J Neural Syst*, p. 2050074, Jan. 2021, doi: 10.1142/S0129065720500744.

[201] M. A. Mulkey, H. Huang, T. Albanese, S. Kim, and B. Yang, 'Supervised deep learning with vision transformer predicts delirium using limited lead EEG', *Sci Rep*, vol. 13, no. 1, p. 7890, May 2023, doi: 10.1038/s41598-023-35004-y.

[202] N. A. Obuchowski and D. K. McCLISH, 'Sample Size Determination for Diagnostic Accuracy Studies Involving Binormal Roc Curve Indices', *Statistics in Medicine*, vol. 16, no. 13, pp. 1529–1542, 1997, doi: 10.1002/(SICI)1097-0258(19970715)16:13<1529::AID-SIM565>3.0.CO;2-H.

[203] N. Burioka *et al.*, 'Approximate entropy of the electroencephalogram in healthy awake subjects and absence epilepsy patients', *Clin EEG Neurosci*, vol. 36, no. 3, pp. 188–193, Jul. 2005, doi: 10.1177/155005940503600309.

[204] S. Qian, Y. Zhu, W. Li, M. Li, and J. Jia, 'What Makes for Good Tokenizers in Vision Transformer?', vol. 14, no. 8, 2015.

[205] Z. Wang, J. Guo, M. van 't Klooster, S. Hoogteijling, J. Jacobs, and M. Zijlmans, 'Prognostic Value of Complete Resection of the High-Frequency Oscillation Area in Intracranial EEG: A Systematic Review and Meta-Analysis', *Neurology*, vol. 102, no. 9, p. e209216, May 2024, doi: 10.1212/WNL.0000000000209216.

[206] W. Zweiphenning *et al.*, 'Intraoperative electrocorticography using high-frequency oscillations or spikes to tailor epilepsy surgery in the Netherlands (the HFO trial): a randomised, single-blind, adaptive non-inferiority trial', *The Lancet Neurology*, vol. 21, no. 11, pp. 982–993, Nov. 2022, doi: 10.1016/S1474-4422(22)00311-8.

[207] L. Noorlag, N. E. C. van Klink, K. Kobayashi, J. Gotman, K. P. J. Braun, and M. Zijlmans, 'High-frequency oscillations in scalp EEG: A systematic review of methodological choices and clinical findings', *Clinical Neurophysiology*, vol. 137, pp. 46–58, May 2022, doi: 10.1016/j.clinph.2021.12.017.

[208] K. A. Klotz, Y. Sag, J. Schönberger, and J. Jacobs, 'Scalp Ripples Can Predict Development of Epilepsy After First Unprovoked Seizure in Childhood', *Annals of Neurology*, vol. 89, no. 1, pp. 134–142, 2021, doi: 10.1002/ana.25939.

[209] N. E. C. van Klink, M. A. van 't Klooster, F. S. S. Leijten, J. Jacobs, K. P. J. Braun, and M. Zijlmans, 'Ripples on rolandic spikes: A marker of epilepsy severity', *Epilepsia*, vol. 57, no. 7, pp. 1179–1189, 2016, doi: 10.1111/epi.13423.

[210] H. O. Heyne *et al.*, 'Polygenic risk scores as a marker for epilepsy risk across lifetime and after unspecified seizure events', *Nat Commun*, vol. 15, no. 1, p. 6277, Jul. 2024, doi: 10.1038/s41467-024-50295-z.

[211] S. Ghosh *et al.*, 'NeuroMorphix: A Novel Brain MRI Asymmetry-specific Feature Construction Approach For Seizure Recurrence Prediction', Apr. 16, 2024, *arXiv*: arXiv:2404.10290. doi: 10.48550/arXiv.2404.10290.

[212] B. K. Beaulieu-Jones *et al.*, 'Predicting seizure recurrence after an initial seizure-like episode from routine clinical notes using large language models: a retrospective cohort study', *The Lancet Digital Health*, vol. 5, no. 12, pp. e882–e894, Dec. 2023, doi: 10.1016/S2589-7500(23)00179-6.

[213] D. Mansilla *et al.*, 'Generalizability of electroencephalographic interpretation using artificial intelligence: An external validation study', *Epilepsia*, vol. 65, no. 10, pp. 3028–3037, 2024, doi: 10.1111/epi.18082.

[214] É. Lemoine *et al.*, 'Improving Diagnostic Accuracy of Routine EEG for Epilepsy using Deep Learning', Jan. 14, 2025, *medRxiv*. doi: 10.1101/2025.01.13.25320425.

[215] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, 'Survival Analysis Part I: Basic concepts and first analyses', *Br J Cancer*, vol. 89, no. 2, pp. 232–238, Jul. 2003, doi: 10.1038/sj.bjc.6601118.

[216] E. Haapaniemi *et al.*, 'The CAVE Score for Predicting Late Seizures After Intracerebral Hemorrhage', *Stroke*, vol. 45, no. 7, pp. 1971–1976, Jul. 2014, doi: 10.1161/STROKEAHA.114.004686.

[217] M. Galovic *et al.*, 'Prediction of late seizures after ischaemic stroke with a novel prognostic model (the SeLECT score): a multivariable prediction model development and validation study', *The Lancet Neurology*, vol. 17, no. 2, pp. 143–152, Feb. 2018, doi: 10.1016/S1474-4422(17)30404-0.

[218] C. Ferreira-Atuesta *et al.*, 'Predictive models for starting antiseizure medication withdrawal following epilepsy surgery in adults', *Brain*, vol. 146, no. 6, pp. 2389–2398, Jun. 2023, doi: 10.1093/brain/awac437.

[219] S. Wiegrebe, P. Kopper, R. Sonabend, B. Bischl, and A. Bender, 'Deep Learning for Survival Analysis: A Review', Jul. 12, 2023, *arXiv*: arXiv:2305.14961. Accessed: Oct. 11, 2023. [Online]. Available: http://arxiv.org/abs/2305.14961

[220] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, 'DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network', *BMC Medical Research Methodology*, vol. 18, no. 1, p. 24, Feb. 2018, doi: 10.1186/s12874-018-0482-1.

[221] L. Yan *et al.*, 'Deep learning models for predicting the survival of patients with chondrosarcoma based on a surveillance, epidemiology, and end results analysis', *Front Oncol*, vol. 12, p. 967758, Aug. 2022, doi: 10.3389/fonc.2022.967758.

[222] L. Xu and C. Guo, 'CoxNAM: An interpretable deep survival analysis model', *Expert Systems with Applications*, vol. 227, p. 120218, Oct. 2023, doi: 10.1016/j.eswa.2023.120218.

[223] Y. She *et al.*, 'Development and Validation of a Deep Learning Model for Non–Small Cell Lung Cancer Survival', *JAMA Network Open*, vol. 3, no. 6, p. e205842, Jun. 2020, doi: 10.1001/jamanetworkopen.2020.5842.

[224] H.-C. Thorsen-Meyer *et al.*, 'Discrete-time survival analysis in the critically ill: a deep learning approach using heterogeneous data', *npj Digit. Med.*, vol. 5, no. 1, pp. 1–10, Sep. 2022, doi: 10.1038/s41746-022-00679-6.

[225] D. Dash *et al.*, 'Update on Minimal Standards for Electroencephalography in Canada: A Review by the Canadian Society of Clinical Neurophysiologists', *Canadian Journal of Neurological Sciences*, vol. 44, no. 6, pp. 631–642, Nov. 2017, doi: 10.1017/cjn.2017.217.

[226] H. Jasper, 'Ten-twenty electrode system of the international federation', *Electroencephalogr Clin Neurophysiol*, vol. 10, pp. 371–375, 1958.

[227] C. R. Pernet *et al.*, 'EEG-BIDS, an extension to the brain imaging data structure for electroencephalography', *Sci Data*, vol. 6, no. 1, Art. no. 1, Jun. 2019, doi: 10.1038/s41597-019-0104-8.

[228] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini, 'Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach', *Statistics in Medicine*, vol. 17, no. 10, pp. 1169–1186, 1998, doi: 10.1002/(SICI)1097-0258(19980530)17:10<1169::AID-SIM796>3.0.CO;2-D.

[229] H. Kvamme and Ø. Borgan, 'Continuous and Discrete-Time Survival Prediction with Neural Networks', Oct. 15, 2019, *arXiv*: arXiv:1910.06724. doi: 10.48550/arXiv.1910.06724.

[230] S. G. Müller and F. Hutter, 'TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation', Aug. 17, 2021, *arXiv*: arXiv:2103.10158. doi: 10.48550/arXiv.2103.10158.

[231] C. Schmoor, W. Sauerbrei, and M. Schumacher, 'Sample size considerations for the evaluation of prognostic factors in survival analysis', *Statistics in Medicine*, vol. 19, no. 4, pp. 441–452, Feb. 2000, doi: 10.1002/(SICI)1097-0258(20000229)19:4<441::AID-SIM349>3.0.CO;2-N.

[232] A. N. Kamarudin, T. Cox, and R. Kolamunnage-Dona, 'Time-dependent ROC curve analysis in medical research: current methods and applications', *BMC Medical Research Methodology*, vol. 17, no. 1, p. 53, Apr. 2017, doi: 10.1186/s12874-017-0332-6.

[233] H. Hung and C.-T. Chiang, 'Estimation methods for time-dependent AUC models with survival data', *Canadian Journal of Statistics*, vol. 38, no. 1, pp. 8–26, 2010, doi: 10.1002/cjs.10046.

[234] J. Lambert and S. Chevret, 'Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves', *Stat Methods Med Res*, vol. 25, no. 5, pp. 2088–2102, Oct. 2016, doi: 10.1177/0962280213515571.

[235] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, 'Assessment and comparison of prognostic classification schemes for survival data', *Statistics in Medicine*, vol. 18, no. 17–18, pp. 2529–2545, 1999, doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5.

[236] L. Sinka *et al.*, 'Association of Mortality and Risk of Epilepsy With Type of Acute Symptomatic Seizure After Ischemic Stroke and an Updated Prognostic Model', *JAMA Neurology*, vol. 80, no. 6, pp. 605–613, Jun. 2023, doi: 10.1001/jamaneurol.2023.0611.

[237] X. Wang *et al.*, 'Development and external validation of a predictive nomogram model of posttraumatic epilepsy: A retrospective analysis', *Seizure - European Journal of Epilepsy*, vol. 88, pp. 36–44, May 2021, doi: 10.1016/j.seizure.2021.03.023.

[238] E. Lindgren *et al.*, 'Development and Validation of a Clinical Score to Predict Epilepsy After Cerebral Venous Thrombosis', *JAMA Neurology*, vol. 81, no. 12, pp. 1274–1283, Dec. 2024, doi: 10.1001/jamaneurol.2024.3481.

[239] L. Möller, G. Krämer, L. Habermehl, K. Menzler, and S. Knake, 'Driving regulations for epilepsy in Europe', *Seizure: European Journal of Epilepsy*, vol. 109, pp. 83–91, Jul. 2023, doi: 10.1016/j.seizure.2023.05.016.

[240] L. J. Bonnett, C. Tudur-Smith, P. R. Williamson, and A. G. Marson, 'Risk of recurrence after a first seizure and implications for driving: further analysis of the Multicentre study of early Epilepsy and Single Seizures', *BMJ*, vol. 341, p. c6477, Dec. 2010, doi: 10.1136/bmj.c6477.

[241] F. E. Harrell Jr, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, 'Evaluating the Yield of Medical Tests', *JAMA*, vol. 247, no. 18, pp. 2543–2546, May 1982, doi: 10.1001/jama.1982.03320430047030.

[242] P. J. Karoly *et al.*, 'The circadian profile of epilepsy improves seizure forecasting', *Brain*, vol. 140, no. 8, pp. 2169–2182, Aug. 2017, doi: 10.1093/brain/awx173.

[243] S. Lundberg and S.-I. Lee, 'A Unified Approach to Interpreting Model Predictions', Nov. 25, 2017, *arXiv*: arXiv:1705.07874. doi: 10.48550/arXiv.1705.07874.

[244] M. Sundararajan, A. Taly, and Q. Yan, 'Axiomatic Attribution for Deep Networks', Jun. 13, 2017, *arXiv*: arXiv:1703.01365. doi: 10.48550/arXiv.1703.01365.

[245] E. Beghi *et al.*, 'Discontinuation of antiseizure medications in seizure-free patients with long-term follow-up: Patients' profile, seizure recurrence, and risk factors', *Epilepsy & Behavior*, vol. 117, p. 107871, Apr. 2021, doi: 10.1016/j.yebeh.2021.107871.

[246] M. Contento *et al.*, 'Prediction of seizure recurrence risk following discontinuation of antiepileptic drugs', *Epilepsia*, vol. 62, no. 9, pp. 2159–2170, 2021, doi: 10.1111/epi.16993.

[247] D. Campos-Fernandez *et al.*, 'Predictive Model for Estimating the Risk of Epilepsy After Aneurysmal Subarachnoid Hemorrhage', *Neurology*, vol. 102, no. 8, p. e209221, Apr. 2024, doi: 10.1212/WNL.0000000000209221.

[248] S. Sheikh and L. Jehi, 'Predictive models of epilepsy outcomes', *Current Opinion in Neurology*, vol. 37, no. 2, p. 115, Apr. 2024, doi: 10.1097/WCO.0000000000001241.

[249] A. Stoller, 'Slowing of the alpha-rhythm of the electroencephalogram and its association with mental deterioration and epilepsy', *J Ment Sci*, vol. 95, no. 401, pp. 972–984, Oct. 1949, doi: 10.1192/bjp.95.401.972.

[250] V. Janiukstyte *et al.*, 'Alpha rhythm slowing in temporal lobe epilepsy across scalp EEG and MEG', *Brain Communications*, vol. 6, no. 6, p. fcae439, Dec. 2024, doi: 10.1093/braincomms/fcae439.

[251] V. M. Uritskii, V. B. Slezin, E. A. Korsakova, S. K. Khorshev, and N. I. Muzalevskaia, 'Fractal diagnostics of disturbances in the alpha-rhythm dynamics in patients with epilepsy', *Biofizika*, vol. 44, no. 6, pp. 1109–1114, 1999.

[252] J. Reiher, M. Beaudry, and C. P. Leduc, 'Temporal intermittent rhythmic delta activity (TIRDA) in the diagnosis of complex partial epilepsy: sensitivity, specificity and predictive value', *Can J Neurol Sci*, vol. 16, no. 4, pp. 398–401, Nov. 1989, doi: 10.1017/s0317167100029450.

[253] U. Amin, F. Nascimento, I. Karakis, D. Schomer, and S. Benbadis, 'Normal variants and artifacts: importance in EEG interpretation', *Epileptic Disorders*, vol. n/a, no. n/a, 2023, doi: 10.1002/epd2.20040.

[254] S. Zhao *et al.*, 'A Systematic Review of Machine Learning Methods for Multimodal EEG Data in Clinical Application', Dec. 31, 2024, *arXiv*: arXiv:2501.08585. doi: 10.48550/arXiv.2501.08585.

[255] Y. Liu *et al.*, 'Epileptic Seizure Detection Using Convolutional Neural Network: A Multi-Biosignal study', in *Proceedings of the Australasian Computer Science Week Multiconference*, in ACSW '20. New York, NY, USA: Association for Computing Machinery, Feb. 2020, pp. 1–8. doi: 10.1145/3373017.3373055.

[256] P. Sirpal, A. Kassab, P. Pouliot, D. K. Nguyen, and F. Lesage, 'fNIRS improves seizure detection in multimodal EEG-fNIRS recordings', *JBO*, vol. 24, no. 5, p. 051408, Feb. 2019, doi: 10.1117/1.JBO.24.5.051408.

[257] J. Zhang, L. Swinnen, C. Chatzichristos, W. Van Paesschen, and M. De Vos, 'Learning Robust Representations of Tonic-Clonic Seizures With Cyclic Transformer', *IEEE J Biomed Health Inform*, vol. 28, no. 6, pp. 3721–3731, Jun. 2024, doi: 10.1109/JBHI.2024.3375123.

[258] M.-P. Hosseini, T. X. Tran, D. Pompili, K. Elisevich, and H. Soltanian-Zadeh, 'Multimodal data analysis of epileptic EEG and rs-fMRI via deep learning and edge computing', *Artif Intell Med*, vol. 104, p. 101813, Apr. 2020, doi: 10.1016/j.artmed.2020.101813.

[259] P. Hager, M. J. Menten, and D. Rueckert, 'Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data', Mar. 30, 2023, *arXiv*: arXiv:2303.14080. doi: 10.48550/arXiv.2303.14080.

[260] W. Huang, 'Multimodal Contrastive Learning and Tabular Attention for Automated Alzheimer's Disease Prediction', presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2473–2482. Accessed: Jun. 29, 2025. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023W/CVAMD/html/Huang_Multimodal_Contrastive_Learning_and_Tabular_Attention_for_Automated_Alzheimers_Disease_ICCVW_2023_paper.html

[261] K. Xie *et al.*, 'Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes', *Epilepsia*, vol. 64, no. 7, pp. 1900–1909, 2023, doi: 10.1111/epi.17633.

[262] G. F. Díaz *et al.*, 'Generalized background qEEG abnormalities in localized symptomatic epilepsy', *Electroencephalogr Clin Neurophysiol*, vol. 106, no. 6, pp. 501–507, Jun. 1998, doi: 10.1016/s0013-4694(98)00026-1.

[263] R. Hornero, P. Espino, A. Alonso, and M. Lopez, 'Estimating complexity from EEG background activity of epileptic patients', *IEEE Engineering in Medicine and Biology Magazine*, vol. 18, no. 6, pp. 73–79, 1999, doi: 10.1109/51.805149.

[264] A. Coito *et al.*, 'Altered directed functional connectivity in temporal lobe epilepsy in the absence of interictal spikes: A high density EEG study.', *Epilepsia*, vol. 57, no. 3, pp. 402–411, Mar. 2016, doi: 10.1111/epi.13308.

[265] J. Reitsma, A. Rutjes, P. Whiting, V. Vlassov, M. Leeflang, and J. Deeks, 'Chapter 9: Assessing methodological quality', in *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*, The Cochrane Collaboration, 2009.

[266] M. A. Belluscio, K. Mizuseki, R. Schmidt, R. Kempter, and G. Buzsáki, 'Cross-Frequency Phase–Phase Coupling between Theta and Gamma Oscillations in the Hippocampus', *J Neurosci*, vol. 32, no. 2, pp. 423–435, Jan. 2012, doi: 10.1523/JNEUROSCI.4122-11.2012.

[267] L. Gagliano, E. Bou Assi, D. K. Nguyen, S. Rihana, and M. Sawan, 'Bilateral preictal signature of phase-amplitude coupling in canine epilepsy', *Epilepsy Res*, vol. 139, pp. 123–128, Jan. 2018, doi: 10.1016/j.eplepsyres.2017.11.009.

[268] E. Bou Assi, L. Gagliano, S. Rihana, D. K. Nguyen, and M. Sawan, 'Bispectrum Features and Multilayer Perceptron Classifier to Enhance Seizure Prediction', *Sci Rep*, vol. 8, no. 1, Art. no. 1, Oct. 2018, doi: 10.1038/s41598-018-33969-9.

[269] H. Motoi *et al.*, 'Phase-amplitude coupling between interictal high-frequency activity and slow wave in epilepsy surgery', *Epilepsia*, vol. 59, no. 10, pp. 1954–1965, Oct. 2018, doi: 10.1111/epi.14544.

[270] L. D. Iasemidis, J. C. Sackellares, H. P. Zaveri, and W. J. Williams, 'Phase space topography and the Lyapunov exponent of electrocorticograms in partial seizures', *Brain Topogr*, vol. 2, no. 3, pp. 187–201, 1990, doi: 10.1007/BF01140588.

[271] D. E. Lerner, 'Monitoring changing dynamics with correlation integrals: Case study of an epileptic seizure', *Physica D: Nonlinear Phenomena*, vol. 97, no. 4, pp. 563–576, Oct. 1996, doi: 10.1016/0167-2789(96)00085-1.

[272] J. Martinerie *et al.*, 'Epileptic seizures can be anticipated by non-linear analysis', *Nat Med*, vol. 4, no. 10, pp. 1173–1176, Oct. 1998, doi: 10.1038/2667.

[273] H. R. Moser, B. Weber, H. G. Wieser, and P. F. Meier, 'Electroencephalograms in epilepsy: analysis and seizure prediction within the framework of Lyapunov theory', *Physica D: Nonlinear Phenomena*, vol. 130, no. 3, pp. 291–305, Jun. 1999, doi: 10.1016/S0167-2789(99)00043-3.

[274] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, 'Densely connected convolutional networks', presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[275] S. Zafar *et al.*, 'Harvard Electroencephalography Database'. BDSP. doi: 10.60508/K85B-FC87.

[276] R. Balestriero *et al.*, 'A Cookbook of Self-Supervised Learning', Jun. 28, 2023, arXiv:2304.12210. doi: 10.48550/arXiv.2304.12210.

[277] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort, 'Uncovering the structure of clinical EEG signals with self-supervised learning', *J. Neural Eng.*, vol. 18, no. 4, p. 046020, Mar. 2021, doi: 10.1088/1741-2552/abca18.

[278] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, 'BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data', *Frontiers in Human Neuroscience*, vol. 15, 2021, Accessed: Feb. 08, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2021.653659

[279] B. A. A. Alsfouk, M. J. Brodie, M. Walters, P. Kwan, and Z. Chen, 'Tolerability of Antiseizure Medications in Individuals With Newly Diagnosed Epilepsy', *JAMA Neurology*, vol. 77, no. 5, pp. 574–581, May 2020, doi: 10.1001/jamaneurol.2020.0032.

[280] H. J. Lamberink *et al.*, 'Individualised prediction model of seizure recurrence and long-term outcomes after withdrawal of antiepileptic drugs in seizure-free patients: a systematic review and individual participant data meta-analysis.', *Lancet Neurol*, vol. 16, no. 7, pp. 523–531, Jul. 2017, doi: 10.1016/S1474-4422(17)30114-X.

[281] A. Strzelczyk *et al.*, 'The impact of epilepsy on quality of life: Findings from a European survey', *Epilepsy & Behavior*, vol. 142, p. 109179, May 2023, doi: 10.1016/j.yebeh.2023.109179.

[282] Loi du Québec, 'Code de la Sécurité Routière du Québec C-24.2, r. 40.1 - Règlement relatif à la santé des conducteurs'. Accessed: Jul. 05, 2025. [Online]. Available: https://www.legisquebec.gouv.qc.ca/fr/document/rc/C-24.2,%20r.%2040.1

[283] K. M. Schubert *et al.*, 'Implications for driving based on the risk of seizures after ischaemic stroke', *J Neurol Neurosurg Psychiatry*, vol. 95, no. 9, pp. 833–837, Sep. 2024, doi: 10.1136/jnnp-2024-333505.

[284] H. Zhu, B. Chen, and C. Yang, 'Understanding Why ViT Trains Badly on Small Datasets: An Intuitive Perspective', Feb. 07, 2023, arXiv:2302.03751. doi: 10.48550/arXiv.2302.03751.

[285] S. McClelland, H. Guo, and K. S. Okuyemi, 'Racial disparities in the surgical management of intractable temporal lobe epilepsy in the United States: a population-based analysis', *Arch Neurol*, vol. 67, no. 5, pp. 577–583, May 2010, doi: 10.1001/archneurol.2010.86.

[286] J. P. Betjemann, A. C. Thompson, C. Santos-Sánchez, P. A. Garcia, and S. L. Ivey, 'Distinguishing language and race disparities in epilepsy surgery', *Epilepsy Behav*, vol. 28, no. 3, pp. 444–449, Sep. 2013, doi: 10.1016/j.yebeh.2013.06.020.

[287] S. Kandregula *et al.*, 'Racial and socioeconomic disparities in the advanced treatment of medically intractable pediatric epilepsy', *Neurosurgical Focus*, vol. 53, no. 4, p. E2, Oct. 2022, doi: 10.3171/2022.7.FOCUS22338.

[288] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, 'Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations', *Nat Med*, vol. 27, no. 12, Art. no. 12, Dec. 2021, doi: 10.1038/s41591-021-01595-0.

[289] J. Yang, A. A. S. Soltan, D. W. Eyre, and D. A. Clifton, 'Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning', *Nat Mach Intell*, vol. 5, no. 8, Art. no. 8, Aug. 2023, doi: 10.1038/s42256-023-00697-3.

[290] A. Chohlas-Wood, M. Coots, S. Goel, and J. Nyarko, 'Designing equitable algorithms', *Nat Comput Sci*, vol. 3, no. 7, Art. no. 7, Jul. 2023, doi: 10.1038/s43588-023-00485-4.

[291] K. M. Fiest *et al.*, 'Prevalence and incidence of epilepsy: A systematic review and meta-analysis of international studies', *Neurology*, vol. 88, no. 3, pp. 296–303, Jan. 2017, doi: 10.1212/WNL.0000000000003509.

[292] M. A. Lopes, S. Perani, S. N. Yaakub, M. P. Richardson, M. Goodfellow, and J. R. Terry, 'Revealing epilepsy type using a computational analysis of interictal EEG', *Scientific Reports*, vol. 9, no. 1, p. 10169, Jul. 2019, doi: 10.1038/s41598-019-46633-7.

[293] J.-H. Zhang *et al.*, 'Personalized prediction model for seizure-free epilepsy with levetiracetam therapy: a retrospective data analysis using support vector machine', *Br J Clin Pharmacol*, vol. 84, no. 11, pp. 2615–2624, Nov. 2018, doi: 10.1111/bcp.13720.

[294] J. T. Oliva and J. L. G. Rosa, 'Differentiation between Normal and Interictal EEG Using Multitaper Spectral Classifiers', in *2018 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2018, pp. 1–8. doi: 10.1109/IJCNN.2018.8489503.

[295] J. E. Jacob, V. V. Sreelatha, T. Iype, G. K. Nair, and D. G. Yohannan, 'Diagnosis of epilepsy from interictal EEGs based on chaotic and wavelet transformation', *Analog Integrated Circuits and Signal Processing*, vol. 89, no. 1, pp. 131–138, Oct. 2016, doi: 10.1007/s10470-016-0810-5.

[296] B. Wahl, A. Cossy-Gantner, S. Germann, and N. R. Schwalbe, 'Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings?', *BMJ Global Health*, vol. 3, no. 4, p. e000798, Aug. 2018, doi: 10.1136/bmjgh-2018-000798.

[297] Y. Si, 'Machine learning applications for electroencephalograph signals in epilepsy: a quick review', *Acta Epileptologica*, vol. 2, no. 1, p. 5, Dec. 2020, doi: 10.1186/s42494-020-00014-0.

[298] C. West, W. Woldman, K. Oak, B. McLean, and R. Shankar, 'A Review of Network and Computer Analysis of Epileptiform Discharge Free EEG to Characterize and Detect Epilepsy', *Clin EEG Neurosci*, vol. 53, no. 1, pp. 74–78, Jan. 2022, doi: 10.1177/15500594211008285.

[299] M. D. F. McInnes, D. Moher, B. D. Thombs, T. A. McGrath, P. M. Bossuyt, and and the PRISMA-DTA Group, 'Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement', *JAMA*, vol. 319, no. 4, pp. 388–396, Jan. 2018, doi: 10.1001/jama.2017.19163.

[300] J. November, 'Biomedical computing: Digitizing life in the United States', *Biomedical Computing: Digitizing Life in the United States*, pp. 1–344, Jan. 2012.

[301] S. Supriya, S. Siuly, H. Wang, and Y. Zhang, 'Automated epilepsy detection techniques from electroencephalogram signals: a review study', *Health Information Science and Systems*, vol. 8, no. 1, Art. no. 1, Oct. 2020, doi: 10.1007/s13755-020-00129-1.

[302] H. J. Schünemann *et al.*, 'GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy', *Journal of Clinical Epidemiology*, vol. 122, pp. 129–141, Jun. 2020, doi: 10.1016/j.jclinepi.2019.12.020.

[303] Y. Varatharajah *et al.*, 'Electrophysiological Correlates of Brain Health Help Diagnose Epilepsy and Lateralize Seizure Focus', in *Annual International Conference Of The IEEE Engineering In Medicine And Biology Society*, 2020, pp. 3460–3464. doi: 10.1109/EMBC44109.2020.9176668.

[304] F. D. Yağmur and A. Sertbaş, 'Automatic Diagnosis of Epilepsy from EEG Signals using Discrete Cosine Transform', in *2020 28th Signal Processing and Communications Applications Conference (SIU)*, 2020, pp. 1–4. doi: 10.1109/siu49456.2020.9302300.

[305] S. Panwar, S. D. Joshi, A. Gupta, and P. Agarwal, 'Automated Epilepsy Diagnosis Using EEG With Test Set Evaluation', *IEEE Transactions on Neural Systems & Rehabilitation Engineering*, vol. 27, no. 6, pp. 1106–1116, 2019, doi: 10.1109/TNSRE.2019.2914603.

[306] H. Rajaei, M. Cabrerizo, S. Sargolzaei, A. Pinzon-Ardila, S. Gonzalez-Arias, and M. Adjouadi, 'Pediatric epilepsy: Clustering by functional connectivity using phase synchronization', in *2015 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Oct. 2015, pp. 1–4. doi: 10.1109/BioCAS.2015.7348368.

[307] R. Ruseckaitė, 'Automated Epilepsy Prediction by Means of Visual Perception Testing and Digital EEG Processing Data', *Informatica*, vol. 12, no. 3, pp. 455–468, Jan. 2001, doi: 10.3233/INF-2001-12307.

[308] M. Poulos, F. Georgiacodis, V. Chrissikopoulos, and A. Evagelou, 'Diagnostic Test for the Discrimination between Interictal Epileptic and Non-Epileptic Pathological EEG Events using Auto-Cross-Correlation Methods', *Neurodiagnostic Journal*, vol. 43(4), pp. 228–240, 2003, doi: 10.1080/1086508x.2003.11079445.

[309] A. M. Bastos and J.-M. Schoffelen, 'A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls', *Frontiers in Systems Neuroscience*, vol. 9, 2016, Accessed: Apr. 11, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnsys.2015.00175

[310] O. Sporns, 'Graph theory methods: applications in brain networks', *Dialogues Clin Neurosci*, vol. 20, no. 2, pp. 111–121, Jun. 2018.

[311] S. Jonas, A. O. Rossetti, M. Oddo, S. Jenni, P. Favaro, and F. Zubler, 'EEG-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features', *Hum Brain Mapp*, vol. 40, no. 16, pp. 4606–4617, Nov. 2019, doi: 10.1002/hbm.24724.

[312] A. Alwosheel, S. van Cranenburgh, and C. G. Chorus, 'Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis', *Journal of Choice Modelling*, vol. 28, pp. 167–182, Sep. 2018, doi: 10.1016/j.jocm.2018.07.002.

[313] B. Jiao *et al.*, 'Neural biomarker diagnosis and prediction to mild cognitive impairment and Alzheimer's disease using EEG technology', *Alzheimer's Research & Therapy*, vol. 15, no. 1, p. 32, Feb. 2023, doi: 10.1186/s13195-023-01181-1.

[314] S. J. M. Smith, 'EEG in neurological conditions other than epilepsy: when does it help, what does it add?', *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. suppl 2, pp. ii8–ii12, Jun. 2005, doi: 10.1136/jnnp.2005.068486.

[315] J. J. Newson and T. C. Thiagarajan, 'EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies', *Frontiers in Human Neuroscience*, vol. 12, 2019, Accessed: Apr. 11, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2018.00521

[316] A. W. Rutjes, J. B. Reitsma, J. P. Vandenbroucke, A. S. Glas, and P. M. Bossuyt, 'Case–Control and Two-Gate Designs in Diagnostic Accuracy Studies', *Clinical Chemistry*, vol. 51, no. 8, pp. 1335–1341, Aug. 2005, doi: 10.1373/clinchem.2005.048595.

[317] R. Aggarwal *et al.*, 'Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis', *npj Digital Medicine*, vol. 4, no. 1, p. 65, Apr. 2021, doi: 10.1038/s41746-021-00438-z.

[318] X. Liu *et al.*, 'A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis', *The Lancet Digital Health*, vol. 1, no. 6, pp. e271–e297, Oct. 2019, doi: 10.1016/S2589-7500(19)30123-2.

[319] M. Nagendran *et al.*, 'Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies', *BMJ*, vol. 368. p. m689, 2020.

[320] S. Kaufman, S. Rosset, and C. Perlich, *Leakage in Data Mining: Formulation, Detection, and Avoidance*, vol. 6. 2011, p. 563. doi: 10.1145/2020408.2020496.

[321] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. in Springer series in statistics. Springer, 2009. [Online]. Available: https://books.google.ca/books?id=eBSgoAEACAAJ

[322] A. L. Beam and I. S. Kohane, 'Big Data and Machine Learning in Health Care', *JAMA*, vol. 319, no. 13, pp. 1317–1318, Apr. 2018, doi: 10.1001/jama.2017.18391.

[323] T. Rad, 'Automated EEG artifact elimination by applying machine learning algorithms to ICA-based features', *J. Neural Eng.*, p. 9, 2017.

[324] F. Farzan *et al.*, 'Standardization of electroencephalography for multi-site, multi-platform and multi-investigator studies: insights from the canadian biomarker integration network in depression', *Sci Rep*, vol. 7, no. 1, Art. no. 1, Aug. 2017, doi: 10.1038/s41598-017-07613-x.

[325] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, and K. A. Robbins, 'The PREP pipeline: standardized preprocessing for large-scale EEG analysis', *Frontiers in Neuroinformatics*, vol. 9, 2015, Accessed: Nov. 12, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fninf.2015.00016

[326] A. Melnik *et al.*, 'Systems, Subjects, Sessions: To What Extent Do These Factors Influence EEG Data?', *Frontiers in Human Neuroscience*, vol. 11, 2017, Accessed: Nov. 12, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2017.00150

[327] L. Hernandez-Ronquillo *et al.*, 'Diagnostic Accuracy of Ambulatory EEG vs Routine EEG in Patients With First Single Unprovoked Seizure', *Neurology: Clinical Practice*, vol. 13, no. 3, Jun. 2023, doi: 10.1212/CPJ.0000000000200160.

[328] S. Saba-Sadiya, E. Chantland, T. Alhanai, T. Liu, and M. M. Ghassemi, 'Unsupervised EEG Artifact Detection and Correction', *Front. Digit. Health*, vol. 2, p. 608920, Jan. 2021, doi: 10.3389/fdgth.2020.608920.

[329] P. Nejedly *et al.*, 'Intracerebral EEG Artifact Identification Using Convolutional Neural Networks', *Neuroinform*, vol. 17, no. 2, pp. 225–234, Apr. 2019, doi: 10.1007/s12021-018-9397-6.

[330] A. Delorme and J. A. Martin, 'Automated Data Cleaning for the Muse EEG', in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2021, pp. 1–5. doi: 10.1109/BIBM52615.2021.9669415.

[331] A. Delorme, 'EEG is better left alone', *Sci Rep*, vol. 13, no. 1, Art. no. 1, Feb. 2023, doi: 10.1038/s41598-023-27528-0.

[332] V. K. Jirsa, W. C. Stacey, P. P. Quilichini, A. I. Ivanov, and C. Bernard, 'On the nature of seizure dynamics', *Brain*, vol. 137, no. 8, pp. 2210–2230, Aug. 2014, doi: 10.1093/brain/awu133.

[333] A. Zalesky, A. Fornito, and E. T. Bullmore, 'Network-based statistic: Identifying differences in brain networks', *NeuroImage*, vol. 53, no. 4, pp. 1197–1207, Dec. 2010, doi: 10.1016/j.neuroimage.2010.06.041.

[334] M. Rubinov and O. Sporns, 'Complex network measures of brain connectivity: Uses and interpretations', *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, Sep. 2010, doi: 10.1016/j.neuroimage.2009.10.003.

[335] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, 'Deep Neural Networks and Tabular Data: A Survey', *arXiv:2110.01889 [cs]*, Feb. 2022, Accessed: Apr. 09, 2022. [Online]. Available: http://arxiv.org/abs/2110.01889

[336] T. Chen and C. Guestrin, 'XGBoost: A Scalable Tree Boosting System', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[337] E. B. Assi, 'Towards accurate prediction of epileptic seizures: A review', *Biomedical Signal Processing and Control*, p. 14, 2017.

[338] A. Bhattacharya, T. Baweja, and S. P. K. Karri, 'Epileptic Seizure Prediction Using Deep Transformer Model', *Int. J. Neur. Syst.*, vol. 32, no. 02, p. 2150058, Feb. 2022, doi: 10.1142/S0129065721500581.

[339] J. Yan, J. Li, H. Xu, Y. Yu, and T. Xu, 'Seizure Prediction Based on Transformer Using Scalp Electroencephalogram', *Applied Sciences*, vol. 12, no. 9, Art. no. 9, Jan. 2022, doi: 10.3390/app12094158.

[340] C. Li, X. Huang, R. Song, R. Qian, X. Liu, and X. Chen, 'EEG-based seizure prediction via Transformer guided CNN', *Measurement*, vol. 203, p. 111948, Nov. 2022, doi: 10.1016/j.measurement.2022.111948.

[341] S. Ahmed, I. E. Nielsen, A. Tripathi, S. Siddiqui, R. P. Ramachandran, and G. Rasool, 'Transformers in Time-Series Analysis: A Tutorial', *Circuits Syst Signal Process*, vol. 42, no. 12, pp. 7433–7466, Dec. 2023, doi: 10.1007/s00034-023-02454-8.

[342] J. P. A. Ioannidis, B. Y. S. Kim, and A. Trounson, 'How to design preclinical studies in nanomedicine and cell therapy to maximize the prospects of clinical translation', *Nat Biomed Eng*, vol. 2, no. 11, pp. 797–809, Nov. 2018, doi: 10.1038/s41551-018-0314-y.

[343] C. Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead', *arXiv:1811.10154 [cs, stat]*, Sep. 2019, Accessed: Mar. 27, 2022. [Online]. Available: http://arxiv.org/abs/1811.10154

[344] R. Desislavov, F. Martínez-Plumed, and J. Hernández-Orallo, 'Compute and Energy Consumption Trends in Deep Learning Inference', Mar. 29, 2023. doi: 10.1016/j.suscom.2023.100857.

[345] J. A. Diao, L. Wedlund, and J. Kvedar, ‘Beyond performance metrics: modeling outcomes and cost for clinical machine learning’, *npj Digit. Med.*, vol. 4, no. 1, Art. no. 1, Aug. 2021, doi: 10.1038/s41746-021-00495-4.

[346] D. B. Burkholder *et al.*, ‘Routine vs extended outpatient EEG for the detection of interictal epileptiform discharges’, *Neurology*, vol. 86, no. 16, pp. 1524–1530, Apr. 2016, doi: 10.1212/WNL.0000000000002592.

[347] P. J. Allen, G. Polizzi, K. Krakow, D. R. Fish, and L. Lemieux, ‘Identification of EEG events in the MR scanner: the problem of pulse artifact and a method for its subtraction’, *Neuroimage*, vol. 8, no. 3, pp. 229–239, Oct. 1998, doi: 10.1006/nimg.1998.0361.

[348] T.-P. Jung *et al.*, ‘Removing electroencephalographic artifacts by blind source separation’, *Psychophysiology*, vol. 37, no. 2, pp. 163–178, Mar. 2000, doi: 10.1111/1469-8986.3720163.

[349] S. Cruces, A. Cichocki, and L. Castedo, ‘Blind source extraction in Gaussian noise’, 2000. Accessed: Apr. 12, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Blind-source-extraction-in-Gaussian-noise-Cruces-Cichocki/d283e748d98350d48ff658f9b78fe6d64c6f6d23

[350] A. F. Jackson and D. J. Bolger, ‘The neurophysiological bases of EEG and EEG measurement: A review for the rest of us’, *Psychophysiology*, vol. 51, no. 11, pp. 1061–1071, 2014, doi: 10.1111/psyp.12283.

[351] Oliverfaust and M. Bairy, ‘NONLINEAR ANALYSIS OF PHYSIOLOGICAL SIGNALS: A REVIEW’, *Journal of Mechanics in Medicine and Biology*, vol. 12, Oct. 2012, doi: 10.1142/S0219519412400155.

[352] D. Lehmann, H. Ozaki, and I. Pal, ‘EEG alpha map series: brain micro-states by space-oriented adaptive segmentation’, *Electroencephalography and Clinical Neurophysiology*, vol. 67, no. 3, pp. 271–288, Sep. 1987, doi: 10.1016/0013-4694(87)90025-3.

[353] Y. LeCun, Y. Bengio, and G. Hinton, ‘Deep learning’, *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

ANNEXE A ARTICLE 3: COMPUTER-ASSISTED ANALYSIS OF ROUTINE EEG TO IDENTIFY HIDDEN BIOMARKERS OF EPILEPSY: A SYSTEMATIC REVIEW

Émile Lemoine[1, 2, 3], Joel Neves Briard[1, 3], Bastien Rioux[4], Oumayma Gharbi[1,3], Renata Podbielski[3], Bénédicte Nauche[3], Denahin Toffa[1,3], Mark Keezer[1,5,6], Frédéric Lesage[2], Dang K. Nguyen[1,3], Elie Bou Assi[1,3]

[1]Department of Neurosciences, University of Montreal, Canada

[2]Institute of biomedical engineering, Polytechnique Montreal, Canada

[3]University of Montreal Hospital Center's Research Center, Canada

[4]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom

[5]School of Public Health, University of Montreal, Canada

[6]Stichting Epilepsie Instellingen Nederland (SEIN), Heemstede, The Netherlands

Ma contribution à cet article comprend la conception du protocole de revue systématique, l'encadrement de la recherche bibliographique, l'évaluation critique des articles, l'extraction et la synthèse des données sur les biomarqueurs EEG, l'analyse des différentes méthodes computationnelles, l'interprétation des résultats, ainsi que la rédaction et la révision du manuscrit. Cet article a été publié dans la revue *Computational and Structural Biotechnology Journal* (*CSBJ*) le 5 décembre 2023.

A.1 Abstract

Background: Computational analysis of routine electroencephalogram (rEEG) could improve the accuracy of epilepsy diagnosis. We aim to systematically assess the diagnostic performances of computed biomarkers for epilepsy in individuals undergoing rEEG.

Methods: We searched MEDLINE, EMBASE, EBM reviews, IEEE Explore and the grey literature for studies published between January 1961 and December 2022. We included studies reporting a computational method to diagnose epilepsy based on rEEG without relying on the identification of interictal epileptiform discharges or seizures. Diagnosis of epilepsy as per a treating physician was the reference standard. We assessed the risk of bias using an adapted QUADAS-2 tool.

Results: We screened 10 166 studies, and 37 were included. The sample size ranged from 8–192 (mean=54). The computed biomarkers were based on linear (43%), non-linear (27%), connectivity (38%), and convolutional neural networks (10%) models. The risk of bias was high or unclear in

all studies, more commonly from spectrum effect and data leakage. Diagnostic accuracy ranged between 64–100%. We observed high methodological heterogeneity, preventing pooling of accuracy measures.

Conclusion: The current literature provides insufficient evidence to reliably assess the diagnostic yield of computational analysis of rEEG.

Significance: We provide guidelines regarding patient selection, reference standard, algorithms, and performance validation.

Systematic review registration: PROSPERO #292261

A.2 Highlights

1) There is insufficient evidence to reliably assess the diagnostic accuracy of computational analysis of routine EEG for epilepsy.
2) Studies are at high risk of bias, mostly due to issues in patient selection and performance validation.
3) We suggest guidelines for future studies regarding patient selection, reference standard, algorithms, and performance validation.

A.3 Introduction

Epilepsy is characterized by a chronic propensity towards epileptic seizures [2]. It is a common neurological condition, with an estimated period (lifetime) prevalence of 1% in the general population [291]. Diagnosing epilepsy poses a serious clinical challenge, with a ~20% misdiagnosis rate [190], [191]. A false positive diagnosis can lead to unnecessary employment and lifestyle restrictions, adverse effects from medications, and social stigma, often for several years [139]. On the contrary, a delay in diagnosis and treatment can put the patient at risk for seizure-related injuries, road accidents, and death [5].

According to the International League Against Epilepsy (ILAE), the diagnosis of epilepsy requires at least two unprovoked epileptic seizures or a single unprovoked seizure with a risk of recurrence ≥60% over 10 years [2]. A short term (20- to 60-minute) scalp electroencephalogram (EEG), or routine EEG, can support a diagnosis after a first single unprovoked seizure. Interictal epileptiform discharges (IEDs) on routine EEG double the risk of recurrent seizures, thus allowing a diagnosis of epilepsy and generally warranting antiseizure medication (ASM) therapy [2], [30], [31].

While they are considered a hallmark of epilepsy, IEDs have limitations that impact the diagnostic utility of routine EEG for epilepsy. On the one hand, overinterpretation of EEG waveforms as IEDs can lead to an erroneous diagnosis of epilepsy [139]. Although the diagnosis of epilepsy is clinical and depends on a clear history of at least one unprovoked seizure [2], in reality, physicians often face an unreliable recounting of the suspected seizure event, and several paroxysmal disorders such as syncope can masquerade as seizures [13], [140]. In these situations, the moderate interrater reliability of IEDs (even among fellowship-trained neurophysiologists) can lead to epilepsy overdiagnosis [37], [253]. On the other hand, IEDs are elusive[10], [28]. In a systematic review of diagnostic accuracy studies assessing routine EEG after a first unprovoked seizure, the sensitivity of EEG was only 17% in adults [30]. Computer-assisted analysis has been proposed as an alternative to increase the test performance of EEG.

Several characteristics of brain activity on EEG may help identify people with epilepsy, including connectivity [88], [183], [292], signal predictability and complexity [184], [293], spectral power [151], [294], and chaoticity[295]. Discovering new, non-visible markers of epilepsy could increase the diagnostic yield of the EEG, improve its accessibility, and reduce costs, especially in settings where the expertise of a fellowship-trained neurophysiologist is unavailable [157], [296]. In spite of this, none of the proposed non-visible markers of epilepsy have translated into clinical practice [2], [31], [52], [53], [157]. Several narrative reviews have described potential biomarkers and EEG processing techniques [32], [297], [298], but there lacks a systematic review evaluating the population and methodological quality of these studies, and summarizing the diagnostic performance of these tools.

We performed a systematic review of diagnostic test accuracy of computational biomarkers (other than IEDs or electrographic seizures) extracted from routine EEG for the diagnosis of epilepsy.

A.4 Methods

We complied with our published protocol to conduct this study [68].

A.4.1 Study design

This study follows guidance from the Cochrane Diagnostic Test Accuracy group. We follow reporting standards set forth by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement for diagnostic test accuracy (PRISMA-DTA) [299]. We considered studies in

all languages published after 1961 (the first use of digital EEG [300]) up to the last review update (December 2022).

A.4.2 Study selection criteria

Type of studies

We included retrospective or prospective diagnostic studies comparing at least one computed biomarker for the diagnosis of epilepsy on <24h scalp EEG (either in the inpatient or outpatient setting) between people with and without epilepsy that did not explicitly rely on the identification of IEDs or ictal activity (seizures). We excluded studies without human participants, studies that used long-term (>24 hours), intracranial, or critical care recordings, studies that focused solely on seizure/spike detection or on short-term (<24h) seizure prediction, as well as studies that did not include both individuals with and without epilepsy. For studies that included multiple EEG recoding settings (e.g., routine and critical care settings) and electrode location (e.g., both surface and intracranial), we only extracted data that met the inclusion criteria.

Population

Our population of interest was individuals undergoing routine EEG in a clinical or research setting. We did not restrict the population to patients undergoing EEG after a first unprovoked seizure. Routine EEG was defined as a <24h scalp recording using the international 10–20 electrodes system, with or without prior sleep deprivation. There was no restriction on age, medication use, or comorbidities.

Reference standard

We defined the reference standard as the diagnosis of epilepsy, as determined by a physician, based on criteria specified by the study authors (clinical or para-clinical), so long as those criteria respected the definition of epilepsy by the International League Against Epilepsy (i.e., had at least one seizure and long-term enduring predisposition to other unprovoked seizures) [2], [3]. Alternative definitions (which do not rely on the presence of at least one seizure) were accepted for the qualitative analysis but excluded from meta-analyses.

Index test

The index test is a characteristic or feature that is computationally extracted from the EEG signal to identify patients with epilepsy, without relying on the detection of IEDs or seizures. These include measures of connectivity, entropy, chaoticity, and power spectrum density[301], as well as statistical models that combine several features or models that directly use the raw EEG signal as their input. We included studies that computed the biomarkers from the same EEG used to diagnose epilepsy, although this was considered in the evaluation of the risk of bias (see **Risk of Bias**).

A.4.3 Search strategy

The search strategy (**Appendix 1**) was developed by two medical librarians specialized in knowledge synthesis (BN and RP). We searched MEDLINE (Ovid), EMBASE (Ovid), EBM reviews (Ovid), IEEE Explore along with grey literature (see **Appendix 1** for details) for articles, conference papers and conference abstracts published between December 1961 and December 2022. We used the Covidence platform (Melbourne, Australia) to manage study selection and data collection. Two independent, mutually blinded reviewers (EL, and either JNB or BR) screened the records for eligibility by title and abstract. Any item deemed relevant by any reviewer was independently assessed for final inclusion from its full text by the same reviewers. Conflicts regarding inclusion were resolved by consensus.

A.4.4 Data collection

Two independent reviewers (EL and OG) extracted pre-specified data while blinded to the verdict of the other reviewer using a custom extraction form tested on the first five articles. Any conflicting data were re-assessed and resolved by consensus. Corresponding authors were contacted through their electronic address if data of interest were not available in the original publication. Data collection included the following information: 1) Title, authors, country of sampling, year of publication; 2) Study type (retrospective vs. prospective, design); 3) Study sample (inclusion/exclusion criteria, number of screened/included patients); 4) Data collection (number of patients and EEGs, duration of EEGs, recording protocol, participants characteristics); 5) Reference standard (definition, application to all patients, time-lapse with EEG); 6) Index test (preprocessing, segment selection, feature extraction and selection, classification algorithm and methodology, reporting of performance); and 7) Measurements of diagnostic test validity (e.g., accuracy, sensitivity, specificity). These items are further detailed in the pre-published protocol[68].

A.4.5 Study reproducibility

Two independent reviewers (EL and OG) assessed study reproducibility. A study was judged reproducible when, given access to the data, the processing methodology and machine learning (ML) methods were sufficiently detailed such that the experiment could be fully reproduced. More specifically, the following items were assessed: objective criteria for selection of EEG segments, code and data availability, and reporting of key methodological details (preprocessing [filtering, channel selection, artifact detection and removal, segmentation], ML optimization [feature extraction and selection, choice of ML model, hyperparameter tuning], and ML evaluation).

A.4.6 Risk of bias

The risk of bias of all included studies was assessed through a version of the QUADAS-2 tool adapted for the characteristics of this review [68], [99]. Two independent and mutually blinded reviewers (EL and OG) assessed the risk of bias for each of the following four elements as low, high, or unclear: 1) Patient selection (representativeness of clinical practice, identical inclusion/exclusion criteria for all participants, exclusion of individual EEG/EEG segments); 2) Index test (identical EEG protocols for all patients, validation of the index test on an independent sample); 3) Reference standard (specified criteria for the diagnosis of epilepsy, independence of the diagnosis to the index test); and 4) Flow and timing (whether the whole sample underwent the same reference standard, timing between index test and epilepsy diagnosis, exclusion of EEG or EEG segments during the evaluation). Any conflicting interpretations were resolved by consensus. These criteria are further detailed in the pre-published protocol [68].

A.4.7 Data synthesis

We planned to report the pooled sensitivity and specificity estimates for studies providing the number of true/false positives/negatives, and the area under the receiver operating characteristic curve (AUROC) for studies that provided a varying threshold. We planned a meta-analysis of diagnostic performances, a quantitative assessment of heterogeneity, and subgroup analyses [68]. However, due to excessive methodological heterogeneity among included studies, we concluded a meta-analysis would not help interpret our results and decided to report a qualitative assessment only (see **Results: Risk of bias and applicability**).

A.4.8 Quality of evidence

The quality of evidence for the primary outcome was evaluated by two authors (EL and OM) based on the GRADE criteria for diagnostic test accuracy [302], recognizing that the GRADE approach is designed for pooled estimates. Data from cross-sectional or cohort study which included patients with diagnostic uncertainty for epilepsy started at "high quality", while data from other observational designs started at "low quality". We downgraded the evidence by one level for high risk of bias, indirectness, inconsistency, imprecision, and high probability of publication bias, and we upgraded the quality by one level for large effect size.

A.5 Results

A.5.1 Study selection

The study selection flow diagram is presented in Figure A.1. Our initial search yielded 10 166 items. After removal of duplicates, title and abstract screening, and full text review, we included 37 studies. The most common reasons for exclusion pertained to study outcome (e.g., seizure or interictal spike detection) in 164 studies (45% of final exclusions), study design (e.g., no diagnostic accuracy testing) in 97 studies (27%), and EEG type (e.g., intracranial, critical care, or long-term monitoring) in 67 studies (19%).

A.5.2 Study characteristics

We describe included studies in Table A.1. The sample size ranged from 8 to 192 (mean=54.4; Figure A.2), while only six studies (16%) included ≥100 subjects[69], [77], [97], [303], [304], [305]. Years of publication ranged from 2001 to 2022; twelve studies (32%) were published after 2020. Most studies included both children (i.e., aged <=18 years old; n=18; 49%) and adults, whereas 11 studies (30%) only included children[70], [71], [72], [82], [83], [90], [91], [92], [98], [184], [306] and eight (22%) only included adults[69], [77], [85], [86], [96], [97], [182], [303]. Twenty-four studies (65%) included any type of epilepsy, whereas seven studies (19%) only included generalized epilepsy[80], [83], [86], [88], [89], [93], [184] and six (16%) only included focal epilepsy[85], [95], [96], [182], [303], [307]. Type of epilepsy, however, was not available in thirteen studies (35%). Five studies (14%) only included patients with electro-clinical syndromes (absence epilepsy[83], idiopathic generalized epilepsies[80], [88], [93], epileptic encephalopathy with spike-wave activation in sleep[82]).

Thirteen studies (35%) provided a definition for the reference standard (diagnosis of epilepsy)[69], [72], [73], [74], [77], [83], [86], [87], [88], [94], [97], [184], [305]. In seven studies (19%), the diagnosis was based on a history of two or more seizures, or one seizure with abnormal neuroimaging or IED on EEG[69], [72], [77], [86], [98], [184], [305]. Three studies (8%) based the diagnosis of epilepsy on EEG features only[83], [88], [94], and three based the diagnosis on the EEG report mentioning a diagnosis of epilepsy[73], [74], [87]. The index tests are described in the section **Signal processing and machine learning,** and the computational biomarkers that were used are listed in Table A.2.

Three public datasets were used by five of the included studies (14%). Three studies used the Temple University Hospital (TUH) EEG dataset ("Epilepsy corpus"), with different sets of inclusion and exclusion criteria, resulting in sample sizes of 40–60 patients (for one study, the final sample size was not available)[73], [74], [87]. One study used the Emotiv dataset, a case-control dataset with 97 subjects recorded with an Emotiv low-cost scalp EEG helmet[81]. One study used the LEMON EEG dataset for the control group only[303].

A.5.3 Risk of bias and applicability

Risk of bias was high or unclear in at least two domains for all studies (Figure A.3Figur). The final consensus for each study and the description of the assessments are provided as supplementary materials (**Figure S1 and Table S2**). For patient selection, no study had a low risk of bias. The most common reason for a high risk of bias in this domain was the use of distinct inclusion and exclusion criteria for subjects with and without epilepsy (e.g., patients with a diagnosis of epilepsy undergoing presurgical evaluation for cases, and healthy individuals for controls). Other reasons were the exclusion of patients without proper justification, and a study population that was not representative of clinical practice. For the index test, two studies had a low risk of bias[74], [98]. High risk of bias in this domain was frequently attributed to failure to validate the index test on an independent sample of patients. In four cases (11%), the EEG recording protocol or setting was different for cases and controls[86], [88], [94], [303]. For the reference standard domain, nine studies (24%) had a low risk of bias[69], [72], [77], [83], [86], [88], [97, p. 202], [184], [305]. A common reason for a high risk of bias included failure to provide a definition for the reference standard. Finally, for the flow and timing domain, two studies had a low risk of bias[74], [94]. For most studies, the risk of bias was unclear because of an unspecified reference standard. Eight

studies (22%) had a high risk of bias in this last domain because they used a different reference standard for cases and controls.

A.5.4 Results of individual studies

Reports of performances for individual studies must be interpreted in the context of high risk of bias in several domains. Diagnostic performances are reported in Table A.3. The diagnostic accuracy ranged from 64% to 100%. Three studies (8%) provided a measure of statistical precision on their diagnostic performance metrics[92], [96], [305]. In the absence of pooled estimates, we assessed applicable GRADE criteria. The evidence quality was judged very low, starting at "low" for the study design and downgraded for high risk of bias, inconsistency (high variability in reported accuracy), and indirectness of evidence (differences between the studied and target populations). Publication bias and imprecision were omitted, as only three studies reported statistical precision.

We analysed how performance was impacted by study size and risk of bias (**Figure S2**). Sample size did not correlate with diagnostic performance. There was no clear trend towards inflated performances for studies at high risk of bias in any of the QUADAS-2 domains although no study had low risk of bias for the Patient selection domain. The inter-test variability was smaller for AUROC than for accuracy. There was a visible trend towards reduced inter-test variability among studies with low risk-of-bias in the Index test (Accuracy and AUROC), Reference standard (AUROC only), and Flow and timing (AUROC only) domains.

A.6 EEG processing and machine learning methods

EEG processing methods for each study are described in Table A.2. Some technical terms related to EEG processing and machine learning are further defined in Table A.4Table.

A.6.1 EEG recording

The range of EEG recording times was 12 seconds to 3 hours (median: 20 minutes, interquartile range [IQR]: 5–25 minutes). The median number of electrodes was 19 (IQR: 19–20.5). In studies reporting EEG montage, 21 (58%) used a referential, and four (11%) used a bipolar montage. Sampling frequency ranged from 114 Hz to 512 Hz, with two studies using frequencies above 1000 Hz (2500[303] and 5000 Hz[182]). The most common sampling frequencies were either 256 Hz or 250 Hz (n=21, 58%).

A.6.2 Segmentation and handling of artifact

Thirty-six of the 37 studies (97%) segmented EEG recordings before analysis. Twenty-three studies (62%) performed manual selection of the EEG segments, most according to pre-specified criteria such as absence of artifacts or absence of ictal activity. The duration of individual EEG segments ranged between 1 and 240 seconds (median=11, IQR: 8–32). One study used the whole, non-segmented EEG for classification[77].

Ten studies (27%) performed artifact detection and rejection, most of which used independent component analysis (ICA, Table A.3)[87], [89], [91], [97], [182], [303], [304]. Another approach was to remove outlier segments based on amplitude[80], [182]. Twenty studies (54%) identified artefactual segments visually from the recordings. No study evaluated the inter-rater reliability of manual selection nor its effect on diagnostic performances.

A.6.3 Computational biomarkers of epilepsy

The computational biomarkers extracted from the EEG signal can be broadly categorized into the following categories: linear, non-linear, connectivity, and deep learning (Table A.2 and Table A.3). Here, we describe in more detail which features were used in the individual studies. Estimation of the diagnostic accuracy of each individual feature, along with comparison between features, was deemed uninformative due to high risk of bias.

Linear

The relative spectral powers of delta (≤4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–40 Hz), and gamma (≥ 40 Hz) bands were used in seven studies[69], [70], [71], [72], [73], [74], [75]. Two studies compared alpha sub-bands (6–9 Hz vs. 8–13 Hz and 7.5–10.5 Hz vs. 10.5–13.5 Hz)[88], [303]. These studies used several methods to extract the power spectral density, including Fast-Fourier transform[70], [71], [75], [303] and an autoregressive model[72]. In all but two studies[78], [88], relative band power was a useful discriminant between groups. Besides estimating power spectral density, autoregressive models can be used to quantify the stationarity of a signal by computing its prediction errors[308], and autocorrelation functions provide a similar information. The linear methods for quantifying stationarity did not show consistent results across studies[79], [184], [308]. Hjorth parameters quantify higher-order statistical moments of the signal in both the

time- and frequency-domains[76]. They were extracted in two studies and seemed discriminant[71], [75].

Zelig et al. (2022) extracted Paroxysmal Slow Wave Events (PSWE), defined as 2-second EEG windows with a median peak frequency of < 6 Hz. In a cohort of 70 patients presenting after a first seizure, the rate of PSWE in the first routine EEG could predict the diagnosis of epilepsy at 18-month with an AUROC of 0.72, regardless of ASM.

Non-linear

Entropy was the most common feature explored for the automated diagnosis of epilepsy. Several algorithms have been developed to estimate entropy from finite physiological time-series. In the selected studies, Shannon[78], [79], Spectral[78], [80], Approximate[79], [81], Permutation[82], Sample (multiscale)[82], [83], Fuzzy[84], and Renyi entropy[78] were used. In some cases, entropy was computed after processing the signal in different frequency bands, either with wavelet decomposition[78] or using a coarse-graining procedure[83], allowing to estimate its value across different timescales.

Other nonlinear features included fractal dimensions (using Higuchi's, Katz', and Petrosian's algorithms)[75], [78], Hurst index (or exponent)[79], zero-crossing interval analysis[85], recurrence quantitative analysis[83], characteristic response analysis (a model of the dynamics of the covariance matrix through time)[305], the bispectrum magnitude (variance and average)[84], periodicity[79], and Kolmogorov complexity[81].

Connectivity and topographical markers

All but one[96] of the 14 connectivity studies used a sensor-based connectivity analysis[69], [78], [81], [86], [87], [88], [89], [90], [91], [92], [93], [94], [95]. The connectivity measure varied widely across studies (Table A.2). A challenge of connectivity estimation is that some sensors may be spuriously connected due to a common underlying source or because of scalp conduction. When these spurious connections occur, the two sensors are phase-aligned (zero-lag), while a "true" communication between brain regions has a small time lag[309]. Therefore, one technique is to use a connectivity measure that accounts for this time lag, which four studies used: lagged correlation[93], lagged coherence[96], Granger's causality[87], and transfer entropy[97]. Another approach reported in two studies was a model of interactions between brain regions based on the

Kuramoto oscillator to calculate parameters that could embody the seizure-generating capacity of the network[88], [93]. Each study analysed the connectivity across several frequency bands.

Once the connectivity matrix is estimated for each frequency band, the studies either directly used the matrix as input into a classification scheme[87], [94], [95] or calculated higher-order features that describe the topology of the underlying network (Table A.2). The discriminative power of each feature was not consistent across studies. Only network efficiency (the average of the shortest path between pairs of nodes) was higher in people without epilepsy in the three studies in which it was analyzed[89], [96], [97]. Overall, the discriminative power of the network features was highly dependent on hyperparameters[91], [97], frequency band[69], [86], [87], [96], and localization[86], [87], with conflicting results between studies. None of the studies performed statistical testing to test the robustness of the estimated network or check it against a random network[310].

Microstates analysis was reported in two studies. Although this analysis can be applied to different frequency bands independently, one study found that microstates features were only discriminant in the beta band[78].

Deep learning

Four studies used deep learning (DL) models, specifically convolutional neural networks (CNN)[73], [74], [87], [98]**.** Two studies performed significant preprocessing on the input signal: one pre-transformed the EEG into connectivity matrices based on Granger causality (6x6–24x24 images)[87] and the other into power spectral density plots (32x32 images)[73]. The other two studies input the raw EEG data (18 channels x 2s and 19 channels x 10s, both 256 Hz), with minimal processing (band pass and notch filtering)[74], [98]. The number of layers in the CNNs ranged from one convolution layer to three blocks of two convolution layers. The number of parameters was not available, but was estimated from figures to range from ~2 960[87] to ~92 000.[98]

The number of recordings used for optimization in those four studies was 48, 32, < 252, and < 1 648 (estimated from figures for the last two studies). When training curves were provided, they revealed overfitting on the training data (i.e., no decrease in loss on the validation set). No study used pre-training nor data augmentation.

Optimization algorithms included Stochastic gradient descent, Adaptive moment estimation (ADAM), and Root Mean Squared Propagation (RMSProp). Only one study used regularization (L2-regularization with dropout) [74].

Comparison between feature extraction approaches

Figure A.4 depicts AUROC and accuracy for the eight studies that did not show data leakage (sharing of information between training and testing set; see section 4.4.3). Tests based on connectivity markers showed high variability in AUROC and accuracy compared to univariate features with no feature extraction. This finding could reflect the heterogeneous data processing related to connectivity analyses. Among these eight studies, only one investigated connectivity and non-linear features across various frequency bands.[78] This study indicated a tendency for improved accuracy when using features extracted from the beta band (Katz's fractal dimension, Shannon entropy, Spectral entropy, Renyi entropy, and microstates features). When assessing all 37 studies, the most performant band varied between the delta,[77] theta,[85], [90] alpha,[72], [90], [93] and beta[78], [86] bands.

A.6.4 Machine learning methods

Thirty of 37 studies (81%) used machine learning to map the extracted features to epilepsy diagnosis. The remaining studies used a receiver operating characteristic (ROC) curve or simple thresholding based on a single, continuous biomarker value[77], [80], [85], [88], [93], [96], [305]. **Supplementary Table S1** summarizes machine learning approaches in included studies.

Algorithms

The support vector machine (SVM) was the most popular across all studies (n=10, 27%)[71], [73], [74], [78], [83], [84], [94], [95], [97], [303]. Studies mainly used radial basis function kernels and polynomial kernels. In some cases, the SVM was directly applied to the pairwise connectivity measures[94], [95].

Multilayer perceptrons were also widely used (n=7, 19%)[71], [75], [78], [79], [81], [82], [304]. Four studies (11%) used convolutional neural networks (discussed in the previous section)[73], [74], [87], [98]. Regression algorithm included logistic regression (n=6, 16%)[69], [73], [89], [182], [184], [303], and linear discriminant analysis (n=3, 8%)[72], [182], [184], often combined with regularization to put a constraint on the value of the parameters and reduce overfitting. Other

classifiers included K-nearest-neighbors (n=5, 14%)[78], [86], [91], [92], [306], gaussian mixture models or naïve bayes with gaussian kernel[90], [303], random forest or other decision trees[74], [78], and gradient boosting[78], [184].

Six studies (16%) compared classifiers to one-another[71], [73], [74], [78], [184], [303]. In Ahmadi et al. (2020), SVM (linear and radial basis function [RBF] kernels) seemed superior to gradient boosting, decision trees, and random forest across experiments. In Varatharajah et al. (2020), both regularized logistic regression and naïve bayes had superior performances over SVM (RBF kernel). In these two studies, classifiers were trained on extracted features and not on the raw, EEG time series. Uyttenhove et al. (2020) compared CNNs trained on the preprocessed windowed EEG signal to an SVM and a random forest trained on the band powers of delta and alpha sub-bands (1.5–2Hz, 10.5–11Hz, 11–11.5Hz, and 11.5–12Hz). They showed that CNNs had higher performance when tested on the TUH Epilepsy Corpus. For each of these studies, there were few details on the hyperparameter optimization of each model, which could have significantly affected the final performances.

Performance evaluation

The most common method for evaluating classification performances was K-fold cross-validation (CV, with K = 5 or 10), used in 10 studies (27%)[78], [81], [82], [83], [86], [87], [89], [95], [97], [98], [182]. A common variation was leave-one-out (or leave-one-pair-out) CV (n=8, 22%)[72], [75], [78], [85], [88], [94], [182], [303]. Repeated or nested-CV was used in five studies (14%)[71], [86], [98], [184], [303]. A potential advantage of CV or repeated testing is that they evaluate the variance of the performances across different partitions of the data. However, none of the studies that performed CV or repeated testing reported the variance of the estimated performances[74], [92], [96], [184].

One common culprit for data leakage was to train the classification algorithm on epochs from one EEG recording, and then evaluate it on different epochs from the same EEG. This could be prevented by grouping together epochs from a single subject into the same data subset. This was done in eight studies (22%)[72], [74], [78], [86], [88], [94], [98], [303].

In five studies (14%), the authors evaluated performances in a dedicated testing set[71], [74], [84], [184], [307]. However, this prevented data leakage in only two of these studies (see next

section)[74], [307]. For the remaining studies, performances were either tested directly on the training data or were not detailed.

Data leakage and train-test loops

Eight studies (22%) did not present data leakage for at least one classification pipeline[74], [78], [85], [88], [94], [98], [303], [307]. In machine learning, data leakage refers to the unintentional sharing of information from the testing set to the training set, resulting in over-optimistic validation performances. Data leakage occurred at different stages of the processing pipeline: feature extraction[72], [78], [84], [97], [182], [305], [308], feature selection[69], [72], [73], [79], [80], [83], [84], [85], [86], [88], [89], [96], [182], [304], [306], and model training and evaluation[69], [72], [74], [77], [78], [86], [88], [90], [92], [93], [94], [96], [98], [303], [304], [305], [306], [308]. Figure A.5 illustrates the most common examples of data leakage. For feature extraction, data leakage occurred when the computation of features required a model to be fitted to the whole dataset, which, for these studies, included samples from the testing set (Figure A.5Figure**B**). Feature selection caused data leakage in all studies that performed it (Figure A.5Figure**C**)[69], [72], [73], [79], [80], [83], [84], [85], [86], [88], [89], [96], [182], [304], [306]. Eight studies (22%) reported grouping samples from the same patients in the same set (training or evaluation), avoiding data leakage that would have occurred by training on epochs from one EEG and testing on different epochs from the same EEG (Figure A.5Figure**E**)[72], [74], [78], [86], [88], [94], [98], [303]. Ten studies (27%) did not use any external validation method when assessing diagnostic performance[69], [77], [90], [92], [93], [96], [304], [305], [306], [308].

Study reproducibility

Six studies (16%) were judged reproducible[80], [83], [86], [93], [98], [184]. The following elements were the most frequently unspecified or poorly specified in studies judged as not reproducible: hyperparameter tuning (n=16, 43%), EEG segmentation (n=16, 43%), model evaluation (n=9, 24%), feature extraction (n=9, 24%), and handling of artifacts (n=9, 24%).

In addition, only three studies (8%) did not involve manual selection of EEG segments[77], [94], [95]. Two studies (5%) provided a certain access to parts of the computer code used for the analysis[87], [93]. Four studies (10%) used publicly available data[73], [74], [81], [87].

Comparison between machine learning approaches

A comparison of the different machine learning models for the eight studies with no data leakage is shown in Figure A.4B. When looking at individual studies, we observed a trend towards higher performances for simpler models in two studies (logistic regression, decision trees),[78], [303] although the magnitude of this difference in accuracies was not reported.

Across all eight studies, deep learning did not clearly show higher performances. However, a direct comparison between deep learning and traditional ML was done in only one study.[74] This study used two different CNN architectures: EEGNet[118], with one split convolution layer (~1 000 parameters) and tiny-VGG (t-VGG)[311], a compact version of the Visual Geometry Group (VGG) architecture with 3 blocks of 2 convolution layers (~21 000 parameters)[74]. They showed that the t-VGG had superior performance for the diagnosis of epilepsy. Few details, however, were provided regarding the training hyperparameters of EEGNet in their study, while they used heavy regularization during the training of t-VGG. In another study, increasing the overlap percentage during segmentation improved performances of CNN, which may be related to the increased size of the training sample with larger overlap (6,000 vs. 11,960 samples).[98] A rule-of-thumb for determining the sample size requirement of a deep neural network is to use 50 training data points per parameter.[312] In the four deep learning studies, the number of parameters were approximately 33,100,[74] 92,000,[98] 2,900,[87] and 19,700[73] (estimations based on study texts). Thus, we estimate that the number of data points represented 7.2%,[74] 0.3%,[98] 0.04%[87], and 0.004%[73] of the sample recommended sample size.[312]

A.7 Discussion

We performed a systematic review of studies reporting computational biomarkers of routine EEG to assess their diagnostic performance for epilepsy. We screened 10 166 studies and included 37 studies, the largest of which had 192 subjects. The included studies reported biomarkers used to classify epilepsy based on linear (43%), non-linear (27%), connectivity (38%), and convolutional neural network (10%) models. Although reported accuracy measures were often high (up to 100%), methodological issues such as spectrum effects and data leakage were ubiquitous and limit the interpretation of these estimates. Therefore, despite several studies published in the last 20 years, the diagnostic performance of computational analysis of routine EEG remains unclear.

The discovery of new reliable interictal markers of epilepsy from routine EEG would significantly impact the approach to the diagnosis of epilepsy[157]. While routine EEG plays an important part

in the classification of epilepsy types and identification of epilepsy syndromes, its role in the diagnosis of epilepsy is mostly restricted to capturing IEDs in patients presenting after a first unprovoked seizure[31], [173]. Because of the sporadic nature of IEDs, their absence cannot rule out a diagnosis of epilepsy (sensitivity), and thus their use as diagnostic biomarkers is limited[30], [31]. In addition, because of their resemblance with other physiological sharply contoured waveforms, overreliance on IEDs can lead to the misdiagnosis of epilepsy (specificity)[37], [253]. The rate of misdiagnosis in epilepsy in the community is estimated to be around 20%[190], [191]. Erroneous diagnoses carry unnecessary and harmful consequences such as stigma, adverse effects from medication, and lifestyle or employment restrictions[140]. Alternative biomarkers could counterweight the limitations of traditional EEG interpretation, potentially accelerating the diagnosis of epilepsy while reducing the burden of over-diagnosis[139]. Several modalities have been proposed as a source of diagnostic and prognostic biomarkers for epilepsy, including neuroimaging, body fluids (blood, cerebrospinal fluid), and metabolic imaging[157]. Compared to these modalities, EEG is inexpensive, technically easy to acquire, and confers functional information with high temporal resolution[64], [65]. Moreover, great effort was put in recent years to standardize the acquisition and storage of routine EEG data[193], [227]. For these reasons, EEG is an invaluable candidate in the search of new interictal markers of seizure risk[157].

We observed a high risk of bias in all included studies. Patient selection might have inflated diagnostic performances reported in most studies especially owing to adopting a “case-control” type of study design.[170], [171] In case-control diagnostic studies, the diagnostic test aims to identify cases (patients with epilepsy) and controls (patients without epilepsy), where both groups are drawn from separate populations (e.g., patients undergoing presurgical evaluation vs. patients evaluated for headaches). Many clinical conditions affect the EEG signal, such as psychiatric diseases, brain lesions, cognitive disorders, medication, and age[31], [178], [179], [313], [314], [315]; failure to account for systematic differences in these comorbidities between cases and controls can result in spectrum effects. This can largely inflate performances of diagnostic test accuracy studies. In this review, the impact of patient selection could not be measured because no studies showed low risk of bias in this domain. The better way to perform patient selection in diagnostic test accuracy studies is to use a consecutive sample of participants respecting common selection criteria (e.g., consecutive patients presenting to the emergency department after a first seizure)[316]. This second option tends to better replicate the scenario where the test will be applied

when deployed in real-life[172]. The need for more robust patient selection methodology is echoed in other recent systematic reviews on the use of machine learning in healthcare[317], [318], [319].

Validation of the biomarkers' performances was another important issue in the evaluation of the risk of bias. Only 22% of the studies did not exhibit data leakage during training and classification. Data leakage occurs when a sample in the evaluation set is used to optimize the classification method[320]. This can happen when the features are computed (feature extraction), when the most discriminative features are selected (feature selection), during the selection of hyperparameters (model tuning), or during the optimization of the classification algorithm (model training) (Error! Reference source not found.)[321]. Classification algorithms frequently require setting specific hyperparameters that control the flexibility of the model and its capacity to fit a particular dataset; the selection of these hyperparameters was largely unreported and can bias accuracy measures upwards [163]. Robust model selection and hyper-parameter tuning do not involve the testing data, an important principle when evaluating clinical predictive algorithms[163], [322]. The studies with low risk of bias in the Index test domain demonstrated smaller inter-test variability. This may highlight the impact of avoiding data leakage on a more precise estimation of diagnostic performance for a given population.[172] However, this estimate may not be generalizable to real-world scenarios depending on the selection criteria used for the study population.

We reported the methods used for processing the EEG signal and predicting the diagnosis, including pre-processing techniques, algorithms for feature extraction, and classification models. A widespread limitation of the EEG processing was the manual selection of artifact-free segments in 54% of studies, without quantifying the effect of this operation on downstream performances, introducing a potential source of bias. Ideally, the processing pipeline should be fully automated and identical for all patients, including artifact detection and segmentation (for example, see [159], [323]). Because of its relatively low signal-to-noise ratio, EEG data is subject to high variability induced by the recording setting, apparel, and even patient-related characteristics (*e.g.*, hair, muscle activation, eye movements).[324], [325], [326] In future studies, large-scale initiatives integrating rEEG recordings from multiple centers along with a more widespread use of ambulatory EEG as a diagnostic tool in patients with first unprovoked seizures[327] will likely amplify this challenge. Automated methods for artifact detection and rejection based on deep neural networks are promising alternatives to manual identification,[328], [329], [330] but their capacity to increase downstream performances remains unclear.[331]

EEGs were segmented into short epochs (typically ≤1 minute) in almost all studies. As a result, the longer-term dynamics of the computational markers were unexplored. The diagnosis of epilepsy relates to a chronically higher propensity to seizure, yet the markers that are evaluated operate on the millisecond-second timescale. Some models of interictal-ictal transition derived from intracranial EEG suggest that there may exist a slowly fluctuating state that embodies the seizure threshold[332], an observation replicated in studies of chronic EEG[47]. Taking these slower dynamics into account could improve the accuracy of seizure propensity assessment on routine EEG.

We could not perform a reliable comparison of the wide range of potential computational biomarkers explored in included studies. It is uncertain whether the studied biomarkers truly represent seizure propensity or are instead a proxy of other conditions that are more prevalent in people with epilepsy, such as ASM therapy and brain lesions. Several markers such as band power were highly discriminant in some studies[70], [72], [303], but not better than chance in others[78], [88]. Most studies evaluated a wide range of features over several frequency bands on a small group of patients, without assessing the variance of the results or using robust model evaluation techniques. In particular, connectivity features were impacted by a low robustness to hyperparameters, which was directly demonstrated in two of the included studies[91], [97]. Statistical validation of network models could help characterize the usefulness of connectivity analysis in future studies[333], [334]. As shown in Figure A.4Figure, methods that take the raw EEG data as input and do not rely on feature extraction may be more robust to the variability introduced by processing parameters and potentially generalize better to external data.

The SVM was the most popular classification algorithm. In a study on the performance of several model architectures for tabular data, ensembles of decision trees (XGBoost, LightGBM, and CatBoost) significantly outperformed deep neural networks and other architectures[335]. This category of machine learning models (initially published in 2016)[336] was used in only two studies (outperforming other models in only one)[78], [184]. An ensemble of decision trees have a high complexity and, without proper hyperparameter tuning and regularization, can easily overfit small datasets, which could explain this discrepancy[336]. For smaller datasets, regularized logistic regression and SVM, which have very few hyperparameters, might be preferable. For complex input such as raw EEG signal, deep neural networks have shown promising performances for the identification and prediction of seizures[337], flagging of abnormal recordings[125], and detection

of interictal discharges[37]. Only two studies used a deep convolutional neural network on the raw EEG data[74], [98]. The sample sizes of the deep learning studies were orders of magnitude smaller (between 0.004% and 7% of suggested sample size) than what is generally suggested.[312] Combined with the complexity and noise of the scalp EEG data, the sample sizes may not have been sufficient to harness the full capacity of deep neural networks. Several questions regarding deep learning remain unanswered, including the minimal quantity of EEGs required, the impact of architecture and optimizer, and the potential benefits of pretraining, self-supervised training, data augmentation, and transfer learning, all of which improved performances in other EEG-related classification tasks[105]. For seizure prediction, where the task consists in predicting (usually from long-term scalp or intracranial EEG data) when a seizure will start minutes or hours in advance, transformer models are becoming the state-of-the-art on benchmark datasets.[338], [339], [340] Transformers are typically larger and more data-hungry than CNN, but might scale better to large datasets.[341]

Understanding the predictions of a machine learning model can provide insights into the neurophysiological manifestations of epilepsy, monitor biases and flaws in the data, and improve acceptability from patients and physicians[342]. This concept is referred to as interpretability, and can take many forms. In one study, the authors used a Kuramoto model to estimate local and global seizure susceptibility from the patients' EEGs[93]. The Kuramoto model is an abstract model of the synchronization between weakly coupled oscillators. As such, their experiment led to the hypothesis that there is a higher coupling strength in patients with generalized epilepsy compared to controls. In another study, the authors investigated the gradient flow through the fitted CNN to identify the regions in the input data that had the highest impact on the CNN's prediction[74]. They found that the EEG regions with highest impact had highly epileptiform anomalies; this would however indicate a limited utility of this approach in the absence of IEDs. In general, interpretability is improved by imposing constraints and sparsity to a machine-learning model[343]. Constraints include imposition of structure and abstraction of unimportant features. Sparsity means that the model is described by a small number of critical parameters. For predicting the diagnosis of epilepsy, an ideal model would provide: 1) a quantification of seizure recurrence risk, 2) actionable parameters (e.g., parameters that can be modified by medication), and 3) parameters that are related to the dynamics of the cortical activity (susceptibility to bifurcations,

altered connectivity, shifts in frequency). Such a model would have the potential to extrapolate to other use cases (e.g., intensive care unit, predict epileptogenicity, post-operatory outcome).

How automated analysis of EEG will integrate into the current diagnostic pathway is yet to be determined. The exact role will likely depend on whether these algorithms prove more sensitive or specific to epilepsy than the current diagnostic approach. If these algorithms were sensitive (*i.e.*, low false negative rate), they could be used as a screening test to exclude epilepsy in patients with low clinical suspicion, reducing the burden of repeat EEGs or accelerating the investigation for alternative conditions. If specific (*i.e.*, low false positive rate), they could be considered as add-ons to IEDs in patients with high pre-test probability, either to individualize the estimation of seizure recurrence risk for a single patient or to provide electrophysiological evidence of epilepsy in patients who do not show IEDs on repeat EEGs. The overhead of the automated analysis of EEG is small and these algorithms could easily be integrated into EEG interpretation software. Even large deep learning models require little computational capabilities to provide inference.[344] Although inference is cheap, training modern and robust ML models requires important computational resources and large, multicenter datasets, both of which come at a potentially very high cost. Another and even more important caveat is the risk of increasing social and racial disparities that are well documented in epilepsy.[285], [286], [287] By training on data that contain these bias, researchers must take active steps to identify and correct for these inequities.[289], [290] Simulation studies could help quantify the net clinical benefits and provide an accurate cost-benefits estimate,[345] which will ultimately hinge on the diagnostic performances of the algorithms.

The strengths of our study include the pre-registration and publication of our study protocol in a peer-reviewed journal, the inclusion of all computational methods, and rigorous study selection and data extraction processes conducted by two independent and mutually blinded reviewers. Our study, however, has limitations. We excluded studies that only used automated IEDs and seizure detection. Although such methods are reported[103], [104], any increment in accuracy from computational identification of IEDs and seizure for the diagnosis of epilepsy is intrinsically limited by their low prevalence in routine EEGs[346]. We considered reports using both IEDs/seizures and other biomarkers of epilepsy on routine EEG, but did not identify such studies. Our goal was to study biomarkers that may help circumvent known drawbacks of human expert assessment and reduce the current reliance on epileptiform discharges. Another limitation is the

high methodological heterogeneity in the studies which prevented any meta-analyses to be performed, although this limitation reflects the state of the existing literature on the topic of interest.

A.7.1 Recommendations

Considering these findings, we propose the following recommendations to guide future studies of computational analysis of EEG for the diagnosis of epilepsy.

Patient selection, reference standard, and study design. Patient selection should be carefully planned to minimize spectrum effect when assessing diagnostic performances. The test should be validated on a consecutive sample of patients that represent the population in which the index test is intended to be used. The reference standard—the diagnosis of epilepsy—should be clearly defined, applied to all patients, and be based on the ILAE's practical definition of epilepsy[2]. Enough details should be provided in the reporting of the study to adequately assess the risk of bias of the methodology, including the start and end of the recruitment period, the number of patients screened for inclusion, the number excluded and reasons for their exclusion. Contemporary reporting standards are available to improve the planification and reporting of diagnostic accuracy studies[166]. Although great effort has been made to publicly share EEG data, current available databases do not yet satisfy these criteria.

Validation of performances. The presence of data leakage must be evaluated at every step of the processing pipeline, from the pre-processing of the EEG signal (using methods that rely on multiple EEGs) to the selection of optimal features and the optimization of the classification algorithm, regardless of the method used for validating performances. Ideally, external validation should also be assessed on independent data, both in terms of location (e.g., different hospital) and time (non-overlapping time periods). Reporting of diagnostic accuracy should be accompanied by a measure of statistical precision, such as a 95% confidence interval.

Code and algorithms. Code should be publicly available to ensure reproducibility of all analyses. Automated segmentation of EEG should be preferred to manual selection of EEG segments. In the case of connectivity analyses, there should be rigorous statistical validation of the network model to increase confidence in the model's prediction. Interpretability should be at the forefront of the design of the machine learning model to increase acceptability and monitor for biases during learning. Transformers, deep CNNs, and graph neural network have revolutionized

our capacity to model complex data and potentially remove the dependency on data pre-processing; they should be considered important candidates for the analysis of clinical EEG.

Clinical translation and applicability. Future studies should provide clear paths towards clinical translation. They should more intentionally target specific clinical populations (e.g., patients evaluated after a first unprovoked seizure, patients with unexplained neurological episodes suspicious of epilepsy) and directly measure the clinical impact compared to current approaches. Small, proof-of-concept studies should make way for larger, multicenter evaluations of diagnostic performances. Integration into clinical workflow, including ease of use, time saved/lost, integration with available tools, computational requirements, and challenges in applicability, should be provided.

A.8 Conclusion

After two decades of research, the current literature provides insufficient evidence to assess the utility of computational analysis of routine EEG to diagnose epilepsy. Studies in this field are at high risk of bias, specifically for patient selection, the definition of the reference standard, and the methodology used to validate diagnostic accuracy. Because of its accessibility and information content, the routine EEG remains an important contender in the search for quantitative markers of seizure risk. We provide recommendations that could guide the design of future studies to maximize the potential for clinical translation of this technology.

A.9 Declarations

Declaration of interest

ÉL is supported by a scholarship from the Canadian Institute of Health Research (CIHR). BR wishes to acknowledge financial support from the Centre for Clinical Brain Sciences of the University of Edinburgh, the CIHR, the Fonds de recherche du Québec—Santé (FRQS) and the Ministère de la Santé et des Services sociaux du Québec, and the Power Corporation of Canada Chair in Neurosciences of the University of Montreal. MRK and DKN report unrestricted educational grants from UCB and Eisai, and research grants for investigator-initiated studies from UCB and Eisai. DKN and FL are supported by the Canada Research Chairs Program, the Canadian Institutes of Health Research, and Natural Sciences and Engineering Research Council of Canada. OG is supported by the Institute for Data Valorization (IVADO). EBA is supported by IVADO

(51628), the CHUM research center (51616), and the Brain Canada Foundation (76097). Funding sources had no role in the design or conduct of the study.

Data availability

Data collected for this study will be available upon reasonable request.

Authors' contributions

Each author contributed to this systematic review. EL planned the study, reviewed the search strategy, participated in the data collection and analysis, drafted the initial manuscript, and is the guarantor of the review. DT, FL, DKN, and EBA participated in the conception of the study. BN and RP designed the search strategy. JNB, BR, and OG participated in data collection. JNB, BR, OG, DT, MRK, FL, DKN, and EBA provided content expertise and critically reviewed the manuscript. All authors reviewed and approved the final manuscript.

Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.12.006.

A.10 Tables

Table A.1 : Characteristics of included studies

Study	Country	Epilepsy type	Total sample size	Group	Description	Age range	Sex (F/M)	Comorbidities	Number of ASM	Framework
Cao, 2021	UK	Generalized	39	Epilepsy	15 PWE	33 ± 12	10/5	None	0-2	Connectivity
				No epilepsy	10 HC and 14 with NEAD	HC: 37 ± 15 NEAD: 33 ± 13	HC: 6/4 NEAD: 10/4	NEAD (14)	HC: 0 NEAD: 0-4	
Guerrero, 2021	Colombia	NA	40	Epilepsy	20 PWE (TUH Epilepsy corpus)	NA	NA	NA	NA	DL, Linear
				No epilepsy	20 w/o epilepsy who underwent a rEEG (TUH Epilepsy corpus)	NA	NA	NA	NA	
Rijnders, 2021	United States	NA	60	Epilepsy	30 PWE (TUH Epilepsy corpus)	52.5 (mean)	19/11	Stroke (3), DM (2), dementia, HBV/HCV (NA)	NA	DL, connectivity
				No epilepsy	20 w/o epilepsy who underwent a rEEG (TUH Epilepsy corpus)	53.7 (mean)	17/13	Stroke (8), DM (3), dementia (2), HBV/HCV (2)	NA	
Zelig, 2021	Israel	Focal, generalized, unknown	100	Epilepsy	28 admitted to the ED after first seizure who developed epilepsy	51.4 ± 20.9	12/16	Headache, brain tumors, IC hemorrhage, MG, depression, AD/HD, autism, schizophrenia, anxiety, substance abuse.	Unclear	Linear
				No epilepsy	42 admitted to ED after fst sz who remained seizure-free & 30 patients undergoing rEEG for neuropsychiatric diseases	Fst sz: 48.5 ± 17.8 Others: 55.1 ± 3.1	Fst sz: 15/27 Others: NA	Similar to cases for fst sz patients; NA for others	Unclear	
Ahmadi, 2020	Belgium	NA	10	Epilepsy	5 PWE	NA	NA	NA	NA	Connectivity, nonlinear
				No epilepsy	5 with PNES	NA	NA	PNES	NA	
Lin, 2020	Taiwan	Focal and generalized	50	Epilepsy	25 PWE	4-17	9/16	NA	NA	DL
				No epilepsy	25 with Tourette's syndrome or syncope	4-15	NA	Tourette's syndrome (92%), syncope (8%)	NA	
Ouyang, 2020	Taiwan	Generalized epilepsy	63	Epilepsy	23 with GE	5-18	10/13	0	0-1	Linear
				No epilepsy	23 age-matched HC	5-18	21/19	NA	0	
Prahbu, 2020	Guinea-Bissau	NA	97	Epilepsy	51 PWE	12-38	21/30	NA	NA	Connectivity
				No epilepsy	46 HC	17-33	5/41	NA	NA	
Song, 2020	China	NA	100	Epilepsy	50 PWE	29.59 ± 4.34	25/25	NA	NA	Nonlinear
				No epilepsy	50 age-matched HC	26.86 ± 3.69	25/25	NA	NA	
Uyttenhove, 2020	Belgium	NA	NA	Epilepsy	PWE (TUH Epilepsy corpus)	NA	NA	NA	NA	DL
				No epilepsy	Patients w/o epilepsy who underwent a rEEG (TUH Epilepsy corpus)	NA	NA	NA	NA	

Varatharajah, 2020	United States	Focal	192	Epilepsy	48 with DRFE	18-66	25/23	NA	NA	Nonlinear
				No epilepsy	144 HC	20-77 (before exclusion)	82/121 (before exclusion)	NA	NA	
Yağmur, 2020	Turkey	NA	108	Epilepsy	88 PWE	NA	NA	NA	NA	Linear
				No epilepsy	20 HC	NA	NA	NA	NA	
Panwar, 2019	India	Focal, generalized, focal and generalized	100	Epilepsy	50 PWE (gen., focal, and LGS)	6-69	16/34	NA	NA	Nonlinear
				No epilepsy	50 HC	6-79	20/30	NA	NA	
Tripathi, 2018	India	NA	20	Epilepsy	10 PWE	3-5	3/7	NA	NA	Linear
				No epilepsy	10 HC	3-5	3/7	NA	NA	
V, 2018	India	Focal	42	Epilepsy	21 with TLE	19-31	0/21	NA	2.66 (mean)	Linear
				No epilepsy	21 HC from existing imaging data bank	24-32	0/21	NA	NA	
Bosl, 2017	United States	Generalized	73	Epilepsy	26 with absence seizures	8.6 (1.7)	13/13	NA	NA	Nonlinear
				No epilepsy	47 undergoing rEEG w/o epilepsy	7.74 (4.3)	15/9	ASD	NA	
Mazzucchi, 2017	Italy	Focal	44	Epilepsy	22 with cryptogenic FE	18-76	13/9	NA	0-4	Connectivity
				No epilepsy	22 age-matched HC	20-73	6/16	NA	NA	
Tibdewal, 2017	India	Focal, generalized	60	Epilepsy	30 with DRFE undergoing pre-surgical evaluation	NA	NA	NA	NA	Nonlinear
				No epilepsy	30 HC	NA	NA	NA	NA	
Uriguen, 2017	Spain	Generalized	30	Epilepsy	20 with IGE	11-70	14/6	NA	0-3	Linear, nonlinear
				No epilepsy	10 HC	23-60	3/7	NA	NA	
Schmidt, 2016	UK	Generalized	68	Epilepsy	30 patients with IGE w/o ASM	NA	NA	NA	0	Connectivity
				No epilepsy	38 HC	NA	NA	NA	NA	
Dasgupta, 2015	India	Generalized	81	Epilepsy	51 with GE	F: 15.21 (mean), M: 13.46 (mean)	26/25	NA	NA	Connectivity
				No epilepsy	30 HC	F: 16.87 (mean), M: 17.67 (mean)	15/15	NA	NA	
Pyrzowski, 2015	Poland	Focal	78	Epilepsy	51 with TLE or FLE, mostly hospitalized for ASM resistance	18-68	36/15	Mood disorder (4), cardiac disease (6), neurosis (2), stroke (2), cerebral palsy, cognitive impairment (2), brain tumor	0 (4), 1 (12), 2 (20), 3 (15)	Nonlinear
				No epilepsy	13 with vEEG confirmed PNES & 14 admitted for headaches	19-57	22/5	Mood disorder (2), migraine (2), meningitis, opioid usage disorder	0 (14), 1 (12), 2 (1)	
Rajaei, 2015*	United States	Focal, generalized	14	Epilepsy	7 PWE	2-14	¾	NA	NA	Nonlinear
				No epilepsy	7 HC	8-18	¾	NA	NA	
			16	Epilepsy	9 PWE	4-15	4/5	NA	0	Connectivity

Sargolzai, 2015 (1)*	United States	Focal, generalized		No epilepsy	7 HC	8-18	¾	NA	0	
Sargolzai, 2015 (2)	United States	Focal, generalized	18	Epilepsy	11 PWE	8-18	5/6	NA	NA	Connectivity
				No epilepsy	7 HC	2-15	3/4	NA	NA	
Schmidt, 2014***	UK	Generalized	75	Epilepsy	35 with IGE	18-59	21/14	NA	0-4	Connectivity, linear
				No epilepsy	40 HC	30.7 (mean)	20/20	NA	NA	
Yang, 2014	China	NA	20	Epilepsy	10 with ESES	3-9	6/4	NA	NA	Nonlinear
				No epilepsy	10 HC	3-9	6/4	NA	NA	
Sargolzaei, 2013*	United States	Focal, generalized	8	Epilepsy	4 PWE	NA	2/2	NA	NA	connectivity
				No epilepsy	4 HC	NA	2/2	NA	NA	
Cabrerizo, 2012*	United States	Focal, generalized	17	Epilepsy	9 PWE undergoing rEEG	1-15	3/6	NA	0	Linear
				No epilepsy	8 patients w/o epilepsy undergoing rEEG	8-18	3/5	NA	0	
Chaovalitwongse, 2011**	United States	NA	15	Epilepsy	10 PWE undergoing rEEG	NA	NA	NA	NA	Linear, connectivity
				No epilepsy	5 patients undergoing rEEG	NA	NA	NA	NA	
Douw, 2010	Netherlands	Focal, generalized	114	Epilepsy	57 PWE who underwent routine EEG after a first seizure	50 (SD: 18)	29/28	White matter abnormalities, brain tumor, cortical atrophy, arachnoid cyst	0-1	Connectivity, linear
				No epilepsy	57 age-matched patients w/o epilepsy who underwent routine EEG after a fst sz	54 (17)	29/28	Stress, syncope, TIA, brain contusion, neuropathy, sleeping disorders, hypoglycemia, migraine, drug abuse, motor neuron disease, orthostatic hypotension, white matter abnormalities, brain tumor, cortical atrophy.	0	
Luo, 2010	China	NA	34	Epilepsy	21 PWE	NA	NA	NA	NA	Linear, nonlinear
				No epilepsy	13 HC	NA	NA	NA	NA	
Bao, 2009	China	NA	12	Epilepsy	6 PWE	NA	NA	NA	NA	Linear, nonlinear
				No epilepsy	6 HC	NA	NA	NA	NA	
Fan, 2009**	United States	Focal	10	Epilepsy	5 DRTLE	NA	NA	NA	NA	Connectivity
				No epilepsy	5 HC	NA	NA	NA	NA	
Cassar, 2008	Greece	Focal, generalized	40	Epilepsy	20 PWE	9-13	11/9	None	NA	Linear
				No epilepsy	20 age- and sex-matched HC	9-13	11/9	None	NA	
Poulos, 2003	Greece	NA	86	Epilepsy	42 PWE	NA	NA	NA	NA	Linear
				No epilepsy	44 with non-epileptic loss of consciousness	NA	NA	NA	NA	
Ruseckaite, 2001	Lithuania	Focal	40	Epilepsy	PWE	NA	NA	NA	NA	Linear
				No epilepsy	HC and head trauma patients	NA	NA	NA	NA	

*Same patients as ref.[90] **Same patients as ref.[94] ***Same patients as ref.[88] ASM: antiseizure medication; Db: diabetes; DL: deep learning; DR: drug-resistant; ESES: electrical status epilepticus during slow-wave sleep; FE: focal epilepsy, FLE: frontal lobe epilepsy; GE: generalized epilepsy; HBV: hepatitis B virus; HC: healthy controls; HCV: hepatitis C virus; IED: interictal epileptiform discharge; IGE: idiopathic generalized epilepsy; NA: Not available; NEAD: non-epileptic attacks disorder; PNES: psychogenic non-epileptic seizures; PWE: patients with epilepsy; rEEG: routine electroencephalography; TLE: temporal lobe epilepsy; vEEG: video-electroencephalography.

Table A.2 : EEG recording and pre-processing details for each study

Study	EEG duration	Electrodes (N)	Sampling freq. (Hz)	Automated artifact detection	Frequency bands	Manual segment selection	Criteria for segment selection	Segment duration (s)	Overlapping segments	Montage	Channel selection	Criteria for channel selection
Cao, 2021	72s for HC, 48s for EG and NEAD	21	500	None	0.79-4, 4-8, 8-15,15-32, >32	Yes	No interictal abnormalities and relatively artifact-free	4	No	Bipolar	Manual	Removed Fp1 and Fp2 due to high levels of eye blink artifacts
Guerrero, 2021	20-30 min	21	250, 256, 400, 512 Hz	NA	NA	NA	NA	NA	NA	Bipolar (longitudinal)	None	-
Rijnders, 2021	20 min	21	250	ICA (removed component with highest correlation with Fp1) and trend line removal	1-4, 5-7, 8-13, 14-29, 30-55	Yes	Calmest segment	50	No	Referential (avg)	None	-
Zelig, 2021	20 min	19	512	NA	1-4, 4-8, 8-12, 12-20, 20-30, 30-40	No		Entire recording	-	NA	None	-
Ahmadi, 2020	3h	27	256	None	1-4, 4-8, 9-13, 13-30, 30-40	Yes	IED-free, least amount of noise or artifacts	16	No	Referential (G2)	None	-
Lin, 2020	20	19	256	None	0.5-60	Yes	No eye movement or muscle artifacts, no segments from IPS nor HV	2	0%, 50%, 90%, 95%,	Referential (Cz)	None	-
Ouyang, 2020	20 min	19	256	None	0.5-60	Yes	Artifact-free	5	No	Referential (Cz,)	None	-
Prahbu, 2020	5	14	128	NA	NA	NA	NA	NA	NA	NA	Automated	Best performing subset in classification task
Song, 2020	2 min	16	512	ICA (NA)	1-4, 4-8, 8-13, 13-30	Yes,	No obvious signal loss	20	No	NA	None	-
Uyttenhove, 2020	NA	19	256	None	0.5-128	No		10	No	Referential (avg)	None	-
Varatharajah, 2020	16 min (controls), NA (cases)	62 (controls), 31 (cases)	2500 Hz (controls), 256 Hz (cases)	ICA (manual selection)	7.5-10.5 10.5-13.5	Yes	Controls: segments with artifacts. Cases: segments with eye closure and no	10	No	Bipolar	Manual	Artefactual channels (4)

							epileptiform activity					
Yağmur, 2020	18 min	16	200	ICA (NA)	0.1-500	NA	NA	NA	NA	NA	None	-
Panwar, 2019	5 min	17	250, 256	None	0.5-15	Unclear	Unclear	1	Yes	Referential (avg)	None	-
Tripathi, 2018	30	19	NA	NA	1-4, 4-8, 8-13, 13-30	NA		NA	NA	Referential	Manual	NA
V, 2018	NA	32	5000	Average subtraction method considering R peaks as reference[347] and ICA (maual selection)	2-20	Yes	First 120s artifact-free segment	120	No	Referential (avg)	None	-
Bosl, 2017	30 s (Hospital subjects) 12 s (Laboratory subjects)	19	200 Hz (Hospital subjects) 500 Hz (Laboratory subjects)	None and NetStation software artifact detection tool (ASD group, manual selection of artifactual components)	None, 0.1-100	Yes, Unclear	Visual review to select 30-s samples containing no spikes or evidence of epileptiform activity and with no artifacts, Exclude segments with eyes saccades and blinks. (automatically detected artifacts)	30, 12	No, No	Average, Average	None and manual	19 channels are selected corresponding to the electrode locations for hospital patients
Mazzucchi, 2017	15 min	19	128	None	1-4, 5-7, 8-13, 14-30, 31-60	Yes	Absence of artifacts, absence of IEDs	2 s	NA	NA	None	-
Tibdewal, 2017	12-15 min	19	114	NA	NA	NA	NA	8	NA	NA	Manual	Removed O1-O2 (corrupted data during acquisition)
Uriguen, 2017	NA	32	200	Kurtosis threshold or statistical outliers (threshold: 3SD)	0.5-70	Yes,	No seizure activity, no epileptiform patterns	10.24	No,	Referential (mastoids)	None	-
Schmidt, 2016	NA	19	256	None	6-9, 8-13	Yes	Artifact free and GSW free	20	No	NA	None	-
Dasgupta, 2015	20-30 min	16	NA	ICA + neural network (underspecified) with manual selection of ICs	4-60	Yes	Noise-free	NA	NA	NA	None	-

Pyrzowski, 2015	20 min (19.5 - 22.1)	19	250	None	4-13	NA	NA	1-120	NA	Referential	None	-
Rajaei, 2015	NA	19	200, 512	None	0.5-70	Yes	Free of artifacts and ictal events	10	No	Referential	None	-
Sargolzai, 2015 (1)	NA	19	200, 500, 512	ICA[348] (NA)	NA	Yes	No seizures and no artifacts	9-90	NA	Referential (avg)	None	-
Sargolzai, 2015 (2)	NA	19	200, 500, 512	None,	NA,	Yes	Artifact-free and seizure-free	9	50%,	Referential	None	-
Schmidt, 2014	50	19	256	None	1-3, 3-6, 6-9, 10-14, 15-30, 30-70	Yes	Artifact-free, eyes closed	20	No	Referential	None	-
Yang, 2014	NA	16	500	None	0.5-35	Yes	No artifacts	8	NA	NA	None	-
Sargolzaei, 2013	NA	19	NA	None	0.1-70	NA	NA	NA	NA	Referential	None	-
Cabrerizo, 2012	20-40 min	19	500 and 512	None	<4, 4-8, 8-13, 13-20, 20-36, 36-44	Yes	Free of artifacts, free of seizures, eyes closed	1	No	Referential	No	-
Chaovalitwongse, 2011	13-45 min	14-18	200 and 250	None	NA	No	Random sampling	60, 120, 180, 240	NA	Bipolar,	Manual	Channels that were consistent across EEGs
Douw, 2010	30 min	21	500	None	0.5-4, 4-8, 8-10, 10-13, 13-30, 30-45, 55-80	Yes	Artifact-free segments	8	No	Referential average	Manual	Fp1-2 and A1-2
Luo, 2010	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-
Bao, 2009	NA	22	200	None	2-34 (1Hz incr.), 2-34 in (2Hz incr.), 2-34.5 in (2.5Hz incr.)	No		20.48, 40.96	NA	Referential, NA	None	-
Fan, 2009	20-30 min	19	250	UNICA[349]	NA	No	Random sampling	30	NA	Referential	None	-
Cassar, 2008	NA	30	400	None	0-4, 4-8, 8-13, 13-30, 30-45, 45-90	Yes	Free of technical and biogenic artifacts	10.24	No	Referential (A1 + A2)	None	-
Poulos, 2003	20 s	2	200	None	5-70	Yes	No epileptiform discharges	20	Yes	O2-Cz	Manual	Channel with "best" PDR
Ruseckaite, 2001	45 s	16	NA	NA	NA	NA	NA	3	No	NA	None	-

EG: Epilepsy group; HC: Healthy controls; ICA: Independent component analysis; NEAD: Non-epileptic attack disorder; PDR: Posterior dominant rhythm.

Table A.3 : Biomarkers assessed in included studies by computational framework

Framework	Feature	Studies
Linear	Power spectral density	[70], [72], [73], [74], [75], [303]
	Peak alpha frequency	[88]
	Mode of frequency spectrum	[307]
	Prediction error of autoregressive model	[184]
	Auto-correlation coefficient or standard deviation	[79], [308]
	Hjorth parameters (activity, mobility, complexity)	[71], [75]
	Statistical features (average, variance, standard deviation, skewness, kurtosis, Euclidean distance, T-Statistical distance, interquartile range, mutual information)	[84], [94], [304]
	Paroxysmal slow wave events (rate per min)	[77]
Nonlinear	Shannon entropy	[78], [79]
	Spectral entropy	[78], [80]
	Approximate entropy	[79], [81]
	Permutation entropy	[82]
	Multiscale entropy	[82], [83]
	Fuzzy entropy	[84]
	Renyi entropy	[78]
	Fractal dimension	[75], [78]
	Hurst indices	[79]
	Zero-crossings interval analysis	[85]
	Recurrent quantitative analysis	[83]
	Characteristic response analysis	[305]
	Bispectrum magnitude (average and variance)	[84]
	Periodicity	[79]
	Kolmogorov complexity	[81]
Connectivity	**Connectivity measures**	
	Mutual information	[86]
	Coherence	[86]
	Lagged coherence	[96]
	Phase-locking value	[86], [88]
	Pearson's correlation coefficient	[86], [89]
	Euclidean distance	[94], [95]
	Cosine similarity	[90], [91], [92], [306]
	Horizontal visibility graph	[78]
	Synchronization likelihood	[69]
	Granger causality	[87]
	Phase-space recurrence	[306]
	Tucker decomposition	[97]
	Transfer entropy	[97]
	Connectivity features	
	Statistical (maximum, mean)	[86]
	Average degree	[88], [91], [92], [306]
	Closeness or betweenness centrality	[78], [90], [91], [92]
	Density	[89], [90], [91], [92], [306]
	Energy	[89], [90], [91], [92]
	Clustering coefficient	[78], [89], [90], [91], [92], [96], [97]
	Network efficiency	[89], [97]
	Rich club coefficient	[89], [90], [91], [92], [306]
	Small world index	[89]
	S-metric	[90], [91], [92], [306]
	Characteristic path length	[92], [96]
	Average vertex eccentricity	[92]
	Graph radius	[92], [306]
	Largest eigenvalue	[78]
	Other connectivity-based features	

	Dynamical connectivity analysis (local and critical coupling constant, global order parameter)	[88], [93]
	Microstates analysis (occurrence, duration, time coverage)	[78], [182]
Deep learning	No feature extraction	[74], [98]
	Prior feature extraction	[73], [87]

Table A.4: Performance of computational EEG biomarkers for the diagnosis of epilepsy

Study	Classifier; feature(s)	Sens (%)	Spec (%)	Acc (%)	Prec (%)	Rec (%)	F_1 (%)	AUROC	Data leakage	Statistical testing
Cao, 2021	kNN; CohMean beta, eyes closed (F4C4-FzCz): Epi vs HC			97.22				0.983	Yes	No
	kNN; CohMean beta, eye closed, (F3C3-FzCz): Epi vs HC							0.969	Yes	No
	kNN; CohMean beta, eye closed, (CzPz-C4P4): Epi vs HC							0.888	Yes	No
	kNN; CohMean beta, eye closed, (C3Cz-P3Pz): Epi vs HC							0.929	Yes	No
	kNN; MI delta, eye open, (T4T6-P4Pz): Epi vs NEAD			74.44					Yes	No
	kNN; PLV gamma, eye open (T3C3-CzPz): Epi vs NEAD			74.24					Yes	No
Guerrero, 2021	LR; relative band power (best model)			73.3	73.9	68	70.8	0.71	Yes	No
	ANN; relative band power (best model)			86.1	81	84	82.4	0.95	NA	No
	SVM; relative band power (best model)			77.3	77.5	74.3	75.8	0.78	NA	No
	CNN; relative band power (best model)			61.5	62.2	58.7	60.4	0.60	NA	No
Rijnders, 2021	CNN; scaled GC matrix, one model per electrode combination, voting	83	87	85			0.85		Yes	No
	CNN; scaled GC matrix, FP1, F3 and P3 electrodes	80	77	78			0.79		Yes	No
Zelig, 2021	ROC; rate of PSWE							0.72	Yes	No
	ROC; rate of PSWE (only early (<72h) EEG)							0.82	Yes	No
Ahmadi, 2020*	Gradient Boost; microstates-derived features			75.4	79.2	75.4			Yes	No
	SVM (Radial basis function); linear, nonlinear and connectivity, alpha			59.2	69.06	59.2			No	No
	SVM (Linear); linear, nonlinear and connectivity, beta			63.8	68.25	63.8			No	No
	RandomForest; linear, nonlinear and connectivity, delta			58.8	68.43	58.8			No	No
	SVM (Radial basis function); linear, nonlinear and connectivity, theta			53.4	64.92	53.4			No	No
	SVM (Linear); linear, nonlinear and connectivity, gamma			55.4	70.01	55.4			No	No
Lin, 2020	CNN; raw signal (0% overlap)	48	82	65			57.83	0.6496	No	No
	CNN; raw signal (50% overlap)	56	82	69			64.36	0.7010	No	No
	CNN; raw signal (90% overlap)	62	90	76			72.09	0.7880	No	No
	CNN; raw signal (95% overlap)	70	90	80			77.77	0.8188	No	No
Ouyang, 2020	XGBoost; autoregressive model errors	89.98	81.81	85.17				0.8754	Yes	No
	L1-Reg. LR; autoregressive model errors	90.47	90.47	84.83				0.8632	Yes	No
	RDA; autoregressive model errors	65.41	86.11	76				0.8908	Yes	No
Prahbu, 2020	MLP; KC and ApEn for 14 electrodes	95.0	98.0	96.5	98.1			0.964	Yes	No
	MLP; KC and ApEn for 6 electrodes	99	94.5	97	95.5			0.967	Yes	No
Song, 2020	SVM with medium Gaussian kernel; connectivity features	86.60	90.0	88.3					Yes	No
	SVM with linear kernel; connectivity features	73.3	70.0	71.70					Yes	No
	SVM fine Gaussian kernel; connectivity features	60	93.3	76.7					Yes	No
	SVM with coarse Gaussian kernel; connectivity features	96.7	43.3	70					Yes	No
Uyttenhove, 2020	t-VGG; raw signal	75.89	78.57	76.5		75.89			No	Yes†
	t-VGG GAP; raw signal	81.56	80.95	81.42		81.56			No	Yes†
	SVM; band power	75.18	71.43	74.32		75.18			No	Yes†
	RandomForest; band power	92.91	52.38	83.61		92.91			No	Yes†
	EEGNet; raw signal	75.89	73.81	75.41		75.89			No	Yes†
Varatharajah, 2020	Naive Bayes with Gaussian prior; band power				0.46	0.75	0.57	0.79	No	No
	SVM (radial basis function); band power				0.89	0.57	0.56	0.66	No	No
	LASSO; band power				0.89	0.56	0.55	0.76	No	No
	GNB (FT channels); band power				0.69	0.73	0.7	0.81	No	No
	SVM-RBF (FT channels); band power				0.889	0.55	0.53	0.73	No	No

	LASSO (FT channels); band power				0.38	0.5	0.43	0.82	No	No
Yağmur, 2020	PCA-MLP; statistical features			96	96	97			Yes	No
	LDA-MLP; statistical features			96	98	95			Yes	No
	Forward selection-MLP; statistical features			85	94	88			Yes	No
	Backward selection-MLP; statistical features			94	95	96			Yes	No
Panwar, 2019	ROC classifier; characteristic method analysis							0.87	Yes	Yes
Tripathi, 2018	Normalised band power			90					NA	No
V, 2018	LDA; microstates features	85.7	66.6	76.1	0.69	0.85	0.76	0.7	Yes	No
	Logistic regression; microstates features	80.9	57.1	69.0	0.65	0.8	0.72	0.67	Yes	No
Bosl, 2017	Linear SVM with RFE; nonlinear features (Epi vs HC+ASD)	100	100	100					Yes	No
	SVM; nonlinear features (Epi vs ASD)	72	77	75					Yes	No
Mazzucchi, 2017	ROC classifier, path length pre- vs. per-HV	41	100	70				0.71	Yes	Yes
Tibdewal, 2017	SVM; BMA-BMV	96.96	100	97.05					Yes	No
	SVM; IQR-MI	98.82	100	99.41					Yes	No
	SVM; MD-MI	98.82	100	99.41					Yes	No
	SVM; MD-IQR	97.65	100	98.82					Yes	No
Uriguen, 2017	ROC; spectral entropy, all channels			85				0.84	Yes	No
	ROC; spectral entropy, optimal channels	86	76	81					Yes	No
Schmidt, 2016	Peak alpha frequency	0	100	0††					No	No
	Connectivity based on PLV (mean degree)	3.3	100	10††					No	No
	Seizure-generating capability based on phase oscillator model	56.7	100	61.7††					No	No
Dasgupta, 2015	Ridge regression with mRMR; connectivity features			79.01				0.87	Yes	No
Pyrzowski, 2015	ROC, alpha score from zero-crossings analysis							0.81	Yes	No
	ROC, Shannon entropy from zero-crossings analysis							0.76	No	No
	ROC, min-entropy from zero-crossings analysis							0.71	No	No
Rajaei, 2015	KNN; connectivity features	85.7	100	92.8					Yes	No
Sargolzai, 2015 (1)	KNN; connectivity features	88.8	85.7	87.5					Yes	No
	KNN with feature selection; connectivity features			96.87					Yes	No
Sargolzai, 2015 (2)	GMM with PCA; connectivity features	81.8	100	88.9					Yes	No
Schmidt, 2014	ROC; theta band critical coupling constant	76.9	65.7		69.2			NA	Yes	No
	ROC; low-alpha band global order parameter for Fp1	71.4	74.4		NA			0.78	Yes	No
Yang, 2014	ANFIS; PermEn			89					Yes	No
	ANFIS; PermEn			82					Yes	No
Sargolzaei, 2013	KNN; connectivity features	75	100						Yes	Yes
Cabrerizo, 2012	ANN; linear features	96.42	95.50	96.03					Yes	No
	SVM; linear features	97.06	96.63	96.79					Yes	No
Chaovalitwongse, 2011	NSVM (Quadratic) dataset I; connectivity	96	100	98					No	No
	NSVM (Quadratic) dataset II; connectivity			100					No	No
	A-SFM (Euclidean), dataset I; connectivity			40					No	No
	A-SFM (T-statistics), dataset II; connectivity			0					No	No
	V-SFM (Euclidean), dataset I; connectivity			100					No	No
	V-SFM (T-statistics), dataset II; connectivity			0					No	No
Douw, 2010	LR; theta band SL	53	70	61					Yes	No
	LR; theta band power	58	77	NA					Yes	No
	LR; theta band SL (EEG with no IEDs)	62	76	69					Yes	No
Luo, 2010	ANN, top three features; linear and nonlinear features	92.2	91.7					0.883	Yes	No
	ANN, all features; linear and nonlinear features							0.908	Yes	No
Bao, 2009	Probabilistic NN, voting across channels for each segment, segment length 40.96s, cut-off frequency NA, band-pass filt. NA; linear and nonlinear features	83.33	84.69	84.27					Yes	No
	Probabilistic NN, voting across channels for each segment, segment length 40.96s, cut-off frequency 56Hz, band-pass filt. 2-32 in 1Hz increments; linear and non-linear features			94.07					Yes	No
Fan, 2009	C-SVM with gaussian kernel; connectivity			94.8					Yes	No
	C-SVM with linear kernel; connectivity			50.6					Yes	No
	SVM with gaussian kernel; connectivity			69.4					Yes	No
	SVM with linear kernel; connectivity			53.8					Yes	No
Cassar, 2008	ARMA model with one band (unspecified) and one electrode (unspecified)	100	65	85					Yes	No
Poulos, 2003	Least-squares; auto-correlation coefficient	0.83	0.90						Yes	No
Ruseckaite, 2001	Euclid classifier, mode of frequency spectrum (background segment)			70					No	No

*Reported from Table 3.[78] †Only for between test comparisons. ††Calculated from study. Acc: Accuracy; ANFIS: Adaptative neuro-fuzzy inference system; ANN: Artificial neural network; ARMA: Autoregressive moving average; ASD: Autism spectrum disorder; BM(A/V): Bispectrum magnitude

(average/variance); CNN: Convolutional neural network; CohMean: Mean of coherence; Epi: Epilepsy; F_1: F1-score; GC: Granger causality; GMM: Gaussian mixture model; GNB: Gaussian Naïve Bayes; HC: Healthy controls; HV: Hyperventilation; IQR: Interquartile range; KC: Kolmogorov complexity; kNN: k-nearest-neighbor; LDA: Linear discriminant analysis; LR: Logistic regression; MD: Mahalanobis distance; MI: Mutual information; MLP: Multilayer perceptron; mRMR: Maximum relevance minimum redundancy; NEAD: Non-epileptic attack disorder; PCA: Principal component analysis; PLV: Phase-locking value; Prec: Precision; PSWE: Paroxystic slow wave events; RDA: Regularized discriminant analysis; Rec: Recall; RFE: Recursive feature elimination; ROC: Receiver operating characteristic curve; Sens: Sensitivity; Spec: Specificity; SFM: Support feature machine; SVM: Support vector machine; t-VGG: tiny-VGG.

Table A.5 : Glossary for technical terms related to EEG processing and machine learning

Terms	Definitions
Linear markers	Markers derived from linear analysis, usually extracted with time-frequency decompositions like the Fourier or wavelet transform. These methods assume independent and stationary oscillating processes. Even though the EEG signal is highly non-linear and non-stationary[22], [350], this simple representation is closely tied to the way neurologists visually inspect EEG recordings.
Non-linear markers	Markers derived from the analysis of non-linear dynamics, either summarized using higher-order features such as entropy and fractal dimensions or analyzed with dynamical models like in recurrent quantitative analysis[351].
Connectivity markers	Markers derived from the analysis of the connectivity between channels (sensor-based) or brain sources (source-based) based on a connectivity measure that represents the strength of pairwise connections between sensors or sources, respectively. Connectivity markers are higher-order features that characterize the network model.
Microstates analysis	In this approach, maps of global field power are extracted at distinct timepoints in the EEG[352]. Using a clustering algorithm, the most characteristic maps for each group are identified—the EEG microstates—on which new EEGs are back-fitted. Features are extracted from time series of microstates, including the duration and coverage (fraction of time that the microstate is active).
Independent component analysis	Blind source separation algorithm that attempts to separate the signal into statistically independent components[348]. The estimated sources are visually inspected to identify those that correspond to artifacts (e.g., blinking, heart rhythms), which are removed before reconstructing the signal with the remaining components. A machine-learning model can also be trained to automatically identify artifactual components[323].
Deep learning	Type of machine learning where models are composed of layers of nonlinear functions that progressively abstract the representation of the raw input data, enabling to capture arbitrarily complex functions[353]**.** For EEG, the main advantage of deep learning is that the model learns its own representation of the input data, without the need of preprocessing and feature extraction.
Support vector machine (SVM)	Soft margin classifier that finds the hyperplane which maximizes the distance between it and the closest observation of each class (called the support vectors). With kernels, the SVM can be optimized on non-linear feature space in a computationally efficient way.
Cross-validation (CV)	Method for validation of predictive performances of a machine-learning model. K-fold CV: in this approach, the dataset is split into K-folds. For K iterations, the machine learning algorithm is optimised on all but one folds, and its predictions are evaluated on the remaining fold. Repeated or nested-CV: the CV is either repeated with different partitions of the data or nested into a second CV loop, both leading to more robust performance estimates[163].

A.11 Figures

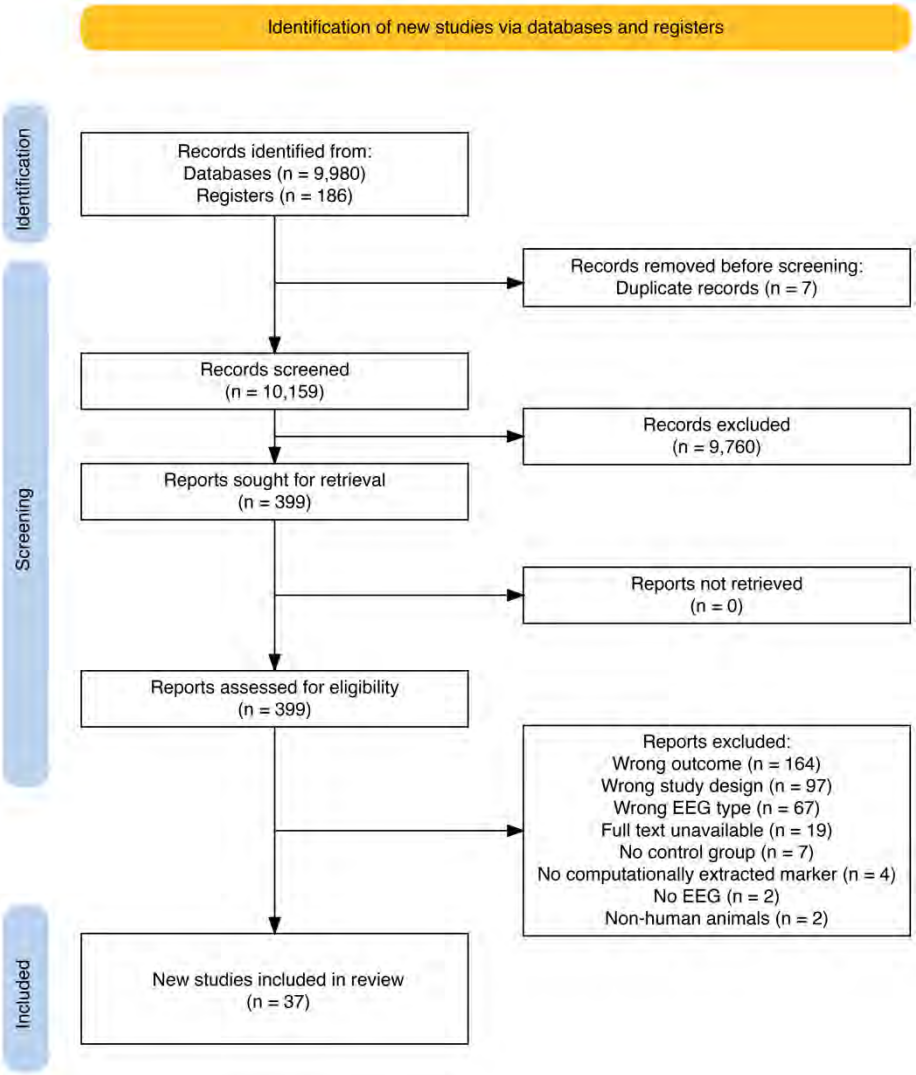


Figure A.1 : PRISMA flowchart of the study selection, screening, and assessment.

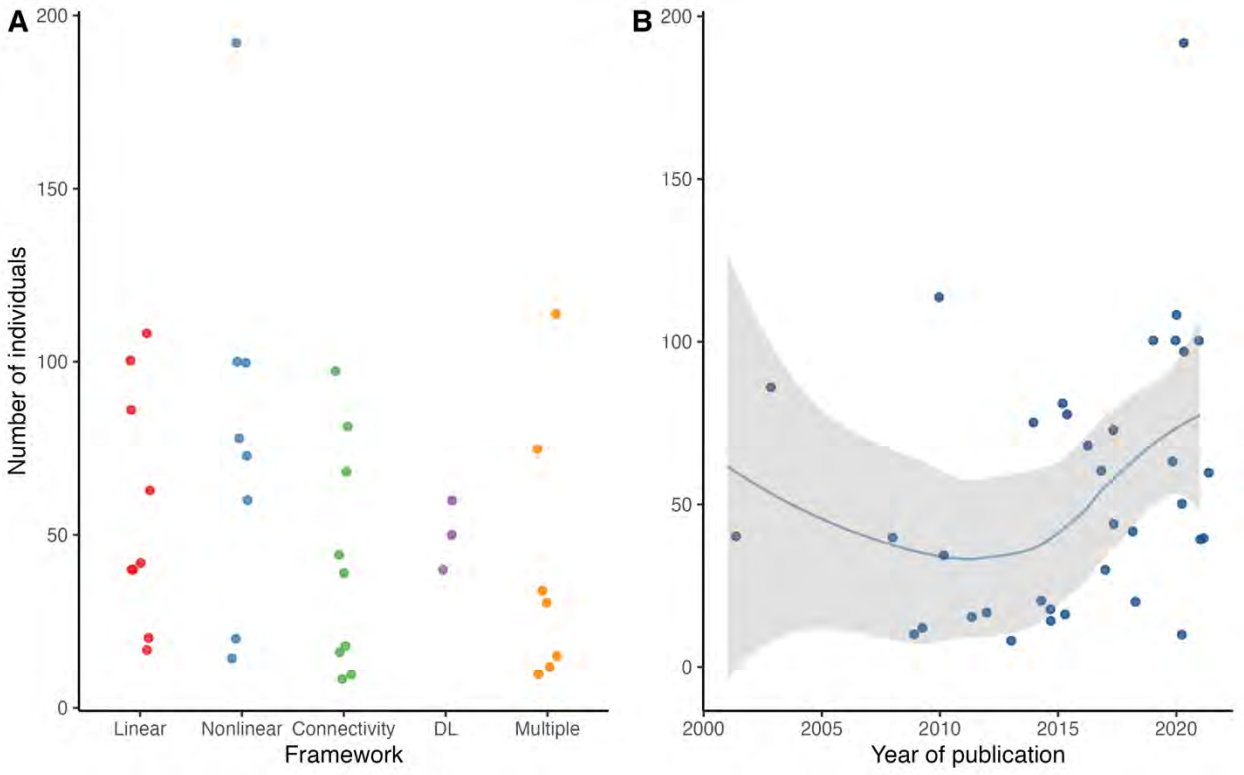


Figure A.2 : Sample size of included studies. A: Number of individuals included in the assessment of computational biomarkers per study. B: Sample size of included studies by year of publication, with a moving average and 95% standard error overlay. Studies with unclear number of participants are not shown.

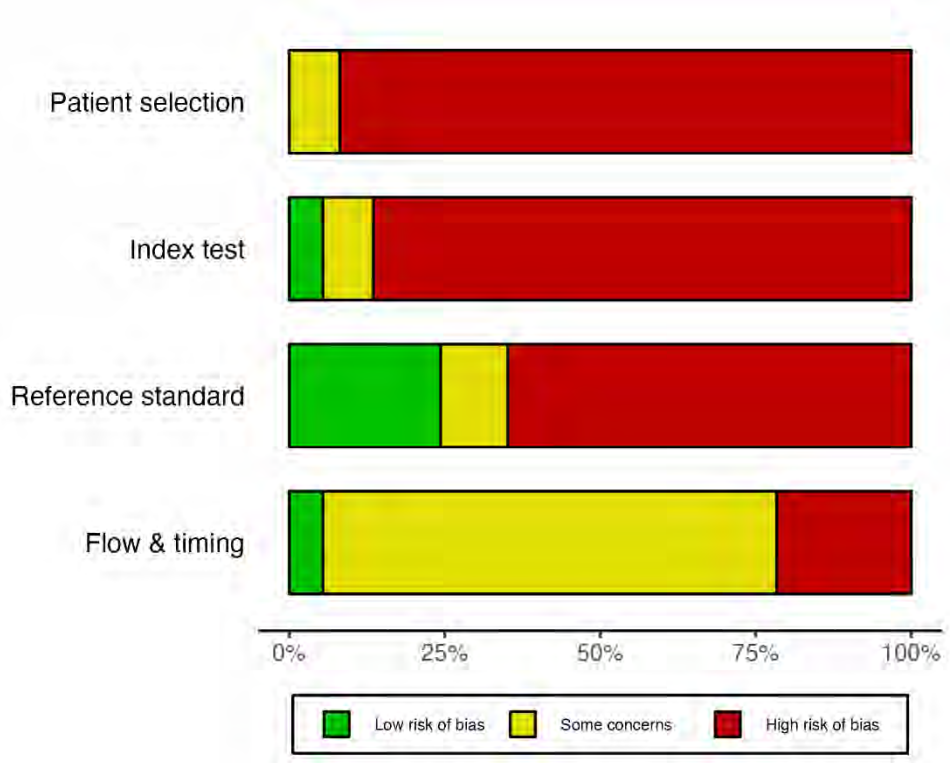


Figure A.3: Summary of the risk of bias for each of the PRISMA domains.

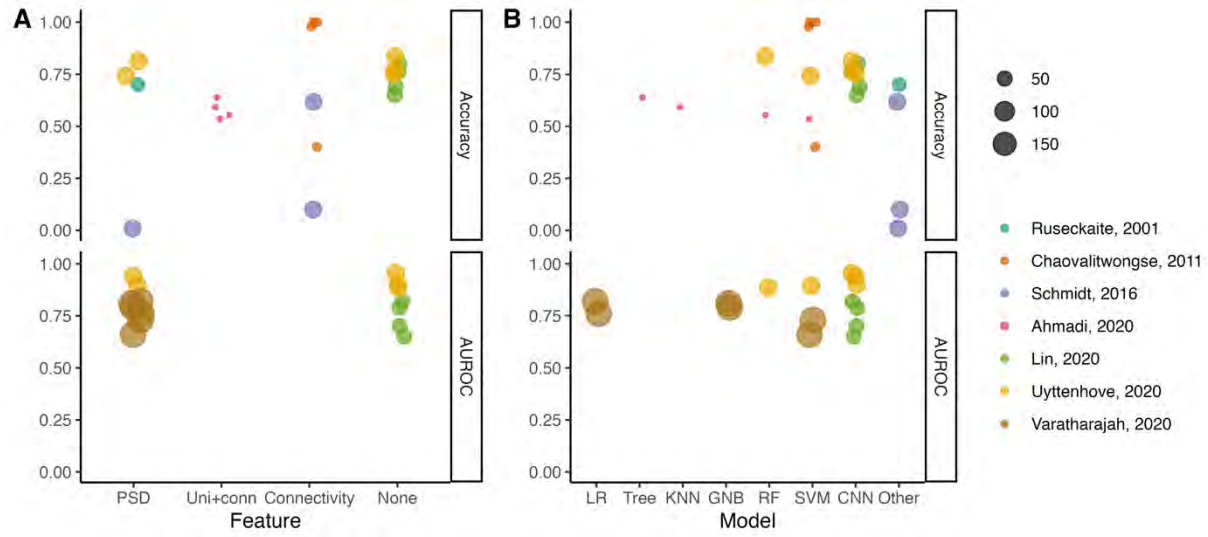


Figure A.4 : Diagnostic performance of studies with no data leakage; all studies reported either Accuracy, AUROC, or both. Each point denotes an individual test reported in the studies (some studies reporting more than one test). A: Performance as a function of the class of feature extracted from the EEG signal. B: Performance as a function of the machine learning model. The size of the points represents sample size. AUROC: Area under the receiver-operating-characteristic curve; CNN: Convolutional neural network; GNB: Gaussian Naïve Bayes; KNN: K-Nearest-neighbor; LR: Logistic regression; PSD: Power spectral density; RF: Random Forest; Uni+conn: Combination of univariate and connectivity features.

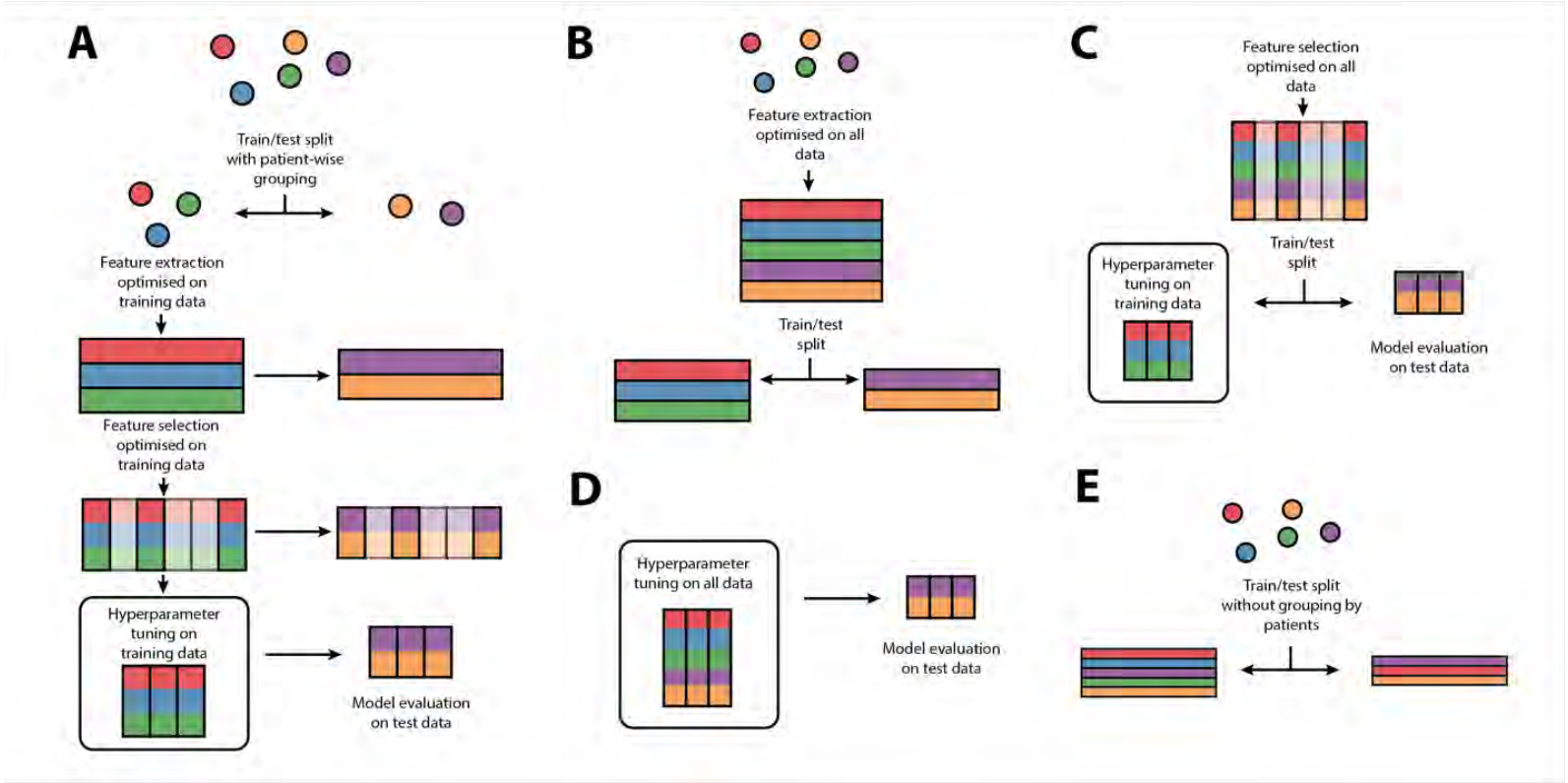


Figure A.5: Examples of common sources of data leakage in the included studies. The circles represent individual observations (e.g., a single EEG recording) and rectangles are the feature vectors for that single observation. Elements in red, blue, and green are in the training set, and elements in purple and orange, the testing set. A: Typical machine learning pipeline without data leakage. First, the individuals (circles) are split into a training and a testing set. Then, features are extracted from the training set; the optimized feature extraction algorithm is then applied to the testing set. Third, a feature selection algorithm is applied to the training data, and the optimal features are selected on the testing data. Fourth, the machine learning hyperparameters are tuned on the training data, and the best model is evaluated on the testing set. B: Data leakage during feature extraction, where the feature extraction algorithm is optimized on both training and testing data (before the train/test split). C: Data leakage during feature selection, where the optimal features are selected on both training and testing data. D: Data leakage during model evaluation, where the hyperparameters are tuned on both training and testing data. E: Data leakage during train/test split, where samples from the same individuals (e.g., different epochs of the same EEG) are present in both training and testing data.