UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES ET INFORMATIQUE
APPLIQUÉES

PAR
CYRUS KALANTARPOUR

UTILISATION DES ALGORITHMES D'APPRENTISSAGE PROFOND POUR DÉTECTER LE SUCCÈS OU L'ÉCHEC DES SÉANCES D'ÉLECTROCONVULSIVOTHÉRAPIE (ECT)

Juin 2023

Mes remerciements vont à :

- Tout d'abord, à mon cher superviseur et directeur, le professeur Usef Faghihi, pour son soutien et sa patience toujours aimables.
- Professeur Meunier pour avoir facilité de nombreuses démarches pédagogiques et pour m'avoir aidé durant mes études.
- Dr François-Xavier Roucaut pour ses conseils médicaux scientifiques et la collecte de fonds pour le projet.
- L'équipe de MITACS pour leurs supports financière.

LISTE DES ABRÉVIATIONS

DL : apprentissage profond (Deep Learning)

ECT : Électroconvulsivothérapie

EEG : l'électroencéphalogramme

CNN : Réseau neuronal convolutif

LSTM : Les réseaux de mémoire à long terme

FFT : Transforme de Fourier rapide

Inférence causale : En général, l'inférence causale est le processus d'inférence de la relation causale entre différentes variables ou événements [1]. Cela implique d'identifier l'effet d'une intervention ou d'un traitement [1] sur le résultat d'intérêt tout en contrôlant les variables confondantes[1] qui peuvent influencer le résultat. Par exemple, le tabagisme cause-t-il le cancer ? La plupart des chercheurs dans le domaine de l'intelligence artificielle utilisent le modèle causal de Pearl et/ou de Rubin [2]. En bref, dans la théorie de Pearl pour trouver les causes des événements, le modèle coupe la relation entre le traitement et ses parents confondants. Le modèle fixe ensuite la valeur du traitement à zéro et un. Enfin, le modèle utilise des théories de probabilité pour calculer la ou les causes des événements.

Logique floue : La logique floue est une logique non classique qui peut gérer le flou et l'incertitude des événements (c'est-à-dire un peu). Pour ce faire, la logique floue traduit les subtilités et les nuances en langage humain tels qu'un peu aux valeurs entre zéros et un afin que la machine puisse les comprendre. Par exemple, un peu peut traduire à une valeur de 0,2. Faghihi et al. [3] ont récemment créé un modèle causal en utilisant la logique floue. Dans la théorie de Faghihi, au lieu de couper les relations entre le traitement et ses parents confondants, le modèle attribue des valeurs floues telles que très faible, faible, moyen et élevé au traitement. Ensuite, en utilisant des théories de logique floue probabiliste comme [3], le modèle calcule les causes des événements. Dans notre deuxième publication (voir ci-dessous), nous avons montré que le modèle de Faghihi est plus concis que celui de Pearl.

Règles causales floues : Nous expliquons les règles causales floues avec l'exemple suivant : par exemple, $max(a, 1 - b)$. Soient a et b deux variables. La règle trouve la valeur maximale de la variable a et de $1 - b$ [4].

Autoencodeur : Un autoencodeur (AE) est un type de réseau de neurones artificiels génératif qui apprend des représentations pour un ensemble de données de manière non supervisée [5].

Autoencodeur variationnel : Un autoencodeur variationnel (VAE) est un type de réseau de neurones génératif qui apprend à encoder et à décoder des données en les représentant dans un espace latent de dimension inférieure. La partie encodeur du VAE mappe les données d'entrée à une distribution de probabilité dans l'espace latent, et la partie décodeur génère des données à partir des échantillons tirés de la distribution de l'espace latent. Le VAE est utilisé pour des tâches telles que la génération d'images, la détection d'anomalies et la compression de données [5].

Autoencodeur variationnel doté à des règles causales (CEVAE) : CEVAE est un type de VAE qui intègre des règles causales telles que celles de Pearl [1] et/ou Faghihi [3]. Dans notre deuxième publication, nous proposons un CEVAE équipé à des règles de la logique floue. L'architecture proposée démontre des résultats prometteurs dans l'extraction de relations causales à partir de données d'observationnels [6].

Résumé et État de l'Art

Publication 1. Le trouble dépressif majeur (TDM) [7] est un gros problème dans notre société. Le TDM peut provoquer des suicides et briser des familles. Plus de 51 milliards[1] de dollars par an sont dépensés dans le secteur de la santé mentale aux États-Unis. Lorsque les traitements médicamenteux échouent, les professionnels de la santé mentale ont recours à l'électroconvulsivothérapie (ECT) [8] pour traiter les patients atteints du TDM. Au cours d'une séance d'ECT, les signaux de l'électroencéphalogramme (EEG) sont enregistrés à partir des activités cérébrales des patients, ce qui permet de décider si le traitement a réussi. Cependant, il n'existe pas de méthode standard pour savoir comment et avec quelle intensité un professionnel de la santé mentale doit appliquer les électrochocs pour traiter les patients souffrant de TDM [8].

Les chercheurs ont utilisé des techniques d'imagerie par résonance magnétique (IRM) multiparamétrique combinées à des méthodes statistiques et/ou à des algorithmes d'apprentissage automatique linéaires pour découvrir la ou les raisons du succès et de l'échec des séances d'électrochocs. Cependant, ces méthodes sont très coûteuses et prennent beaucoup de temps sans produire de bons résultats [9-14].

Au cours de mon mémoire de maîtrise et comme première étape de ce projet, nous nous sommes intéressés à trouver des moyens de classifier les succès ou les échecs des séances d'ECT.

Afin de trouver des caractéristiques et des modèles prédictifs possibles pour la réussite et l'échec des séances d'électrochocs, pour la première fois, nous avons utilisé les EEG recueillis sur le cuir chevelu de patients souffrant de dépression majeure avec des algorithmes hybrides d'apprentissage profond tels que le réseau neuronal convolutif (CNN) et les réseaux de mémoire à long terme (LSTM) (voir première publication ci-dessous). Cependant, les EEG sont complexes et les algorithmes d'apprentissage profond ne peuvent pas les utiliser directement [15]. Par conséquent, des techniques de prétraitement des données sont utilisées pour permettre aux algorithmes d'apprentissage profond de les traiter. Pour ce faire, nous avons utilisé l'algorithme de Transforme de Fourier rapide (FFT) qui calcule la transformée de Fourier discrète (DFT) d'une séquence. La DFT convertit un signal de son domaine d'origine (souvent temporel ou spatial) en une représentation dans le domaine fréquentiel et vice versa.

D'autres techniques que nous avons utilisées c'était la suppression des artefacts [8]. Par exemple, les mouvements oculaires des patients produisent une fréquence indésirable. Une fois les données prêtes nous avons utilisé les algorithmes d'apprentissage profond hybrides tels que les réseaux de neurones convolutifs (CNN) et les réseaux de mémoire à long terme (LSTM). Le réseau CNN-LSTM a obtenu une précision de 83 % pour la classification de la réussite et l'échec des séances d'électrochocs.

[1] https://link.springer.com/article/10.1007/s40273-021-01019-4

Publication 2. L’architecture hybride d’apprentissage profond CNN-LSTM [16] (DL) que nous avons utilisé dans notre étude précédente n’a pas réussi à prédire avec quelle intensité les charges électriques (le dosage de l’ECT) doivent être appliquées au crâne du patient.

Le dosage des ECT est crucial pour rendre le traitement individualisé. L'une des raisons de cet échec est que l'architecture CNN-LSTM ne peut pas faire du raisonnement. Selon Faghihi et ses collègues [3], le raisonnement est un outil puissant pour doter les algorithmes d'apprentissage profond afin qu’ils puissent faire la généralisation.

Dans cette étude, pour la première fois, nous avons intégré des règles causales floues [3] à l'architecture CEVAE (Causal Effect Variational Autoencoder) [10] afin d'extraire les relations causales entre les caractéristiques de l'ensemble de données.

Pour ce faire nous avons ajouté des règles causales floues à la fonction de perte de CEVAE. En comparaison, notre architecture fuzzy CEVAE ou FCEVAE surpasse le CEVAE original et l’outil de raisonnement Dowhy [17] développé par Microsoft en termes d'inférence causale. FCEVAE est capables de prédire l'intensité des impulsions électriques des ECTs (le dosage des ECT).

Mes contributions dans cette thèse : J'ai travaillé sur l'optimisation des séances de thérapie l'électroconvulsivothérapie (ECT) à l'aide des algorithmes d'apprentissage profond doté à des règles causales [3]. Pour ce faire, nous avons collaboré avec le Dr François-Xavier Roucaut, psychiatre à l'hôpital St-Marie de Trois Rivières. Voici mes contributions lors de mon mémoire de maîtrise:

1- J'ai créé un ensemble de données à partir des données recueillies pendant des séances d’ECT par le Dr François-Xavier Roucaut, psychiatre à l'hôpital St-Marie de Trois-Rivières en collaboration avec l'Institut universitaire en santé mentale du Québec et l'hôpital Saint-Jérôme. Nous avons présenté ci-dessous les étapes que nous avons suivies jusqu'à présent.

2- Pour chaque patient, chaque séance d’ECT a un fichier EEG et un fichier contenant des informations personnelles sur les patients telles que l'âge, le sexe et les réponses rétrospectives à la séance d’ECT qui sont remplies manuellement. J’ai écrit un code de Python qui lit automatiquement les fichiers, et les nettoie, et les anonymise (GitHub).

3- Étant donné que les données EEG contiennent des artefacts et des fonctionnalités que les algorithmes d'apprentissage profond ne peuvent pas traiter directement [4, 5], j'ai utilisé différentes techniques telle que Fast Fourrier Transform (voir ci-dessus pour une brève description).

4- Pour détecter les réussis ou les échecs des séances d’ECT, j'ai utilisé une architecture d’apprentissage profond hybride contentant les réseaux de neurones convolutionnels plus un réseau de neurones à mémoire à long terme (CNN-LSTM). Le modèle CNN-LSTM a atteint une précision de 83% pour la classification des EGGS réussis et non réussis.

5- Cependant, le CNN-LSTM ne peut pas trouver pourquoi certaines séances de ECT réussissent tandis que d'autres non. Néanmoins, trouver les causes profondes des tracées d’ECT réussie et non- réussies est essentiel car cela permettra de créer des traitements individualisés pour les TDM. Par conséquent, j’ai doté les algorithmes d'apprentissage profonds à des capacités de raisonnement, en ajoutant les règles floues causales de Faghihi et al. [3] à l'Autoencodeur Variationnel à Effet Causal (CEVAE) ce que nous avons appelé FCEVAE [4]. En ajoutant le raisonnement à CEVAE, nous avons obtenu : 1) les causes sous-jacentes du succès et de l'échec des EEGs, 2-) la prédiction de dose de ECT individualisée avec une précision de 91 % uniquement en utilisant les données pré-choc.

6- A notre connaissance, c'est la première méthode permettant de prédire la charge thérapeutique à appliquer au crâne du patient (GitHub).

Abstract

Publication 1. Major Depression Disorder (MDD) is a big problem in our society. MDD can cause suicide and take families apart. More than \$51 billion[1] a year is spent annually in the mental health sector in US.

So far, researchers have used costly and time-consuming multi-parametric Magnetic Resonance Imaging (MRI) techniques with statistical methods and linear machine learning algorithms to predict ECT outcomes with limited accuracy [11].

In this paper, we utilize a deep learning (DL) method to classify electroconvulsive therapy (ECT) outcomes as the first step toward digitalizing and then optimizing ECT. While there is no standard method for applying ECT to treat MDD, we classify the success or failure of ECT sessions using deep learning algorithms such as convolutional neural networks (CNN) and long short-term memory networks (LSTM). To make this possible, the researchers used electroencephalogram (EEG) data collected from MDD patients and preprocessed it with techniques such as fast Fourier transform (FFT) and artifact removal. The employed CNN-LSTM model achieved an accuracy of 83% in classifying the success and failure of ECT sessions.

By using EEG data, our approach is less expensive and more time-efficient than using MRI family techniques, and we believe it can help mental healthcare professionals achieve better treatment outcomes for patients suffering from MDD.

Our second paper introduces several key terms and concepts that the reader needs to understand them before going through details. In this short introduction, we provided a very brief explanation of the key concepts that we used to explain our work.

Causal inference: Generally speaking, causal inference is the process of inferring the causal relationship between different variables or events [1]. It involves identifying the effect of an intervention or treatment on the outcome of interest while controlling for confounding variables [1] that can influence the outcome. For instance, would smoking causes cancer? Most of the researchers in the domain of Artificial intelligence use Pearl's and/or Rubin's causal model [2]. In short, in Pearl's theory to find the causes of the events, the model cuts the relationship between the Treatment and its confounding parents. The model then, fixes the Treatment value to zero and one. Finally, the model using probability theories calculates the cause(s) of the events.

Fuzzy logic: Fuzzy logic is a non-classical logic that can handle the vagueness and uncertainty (i.e., a little) of events. Faghihi et al [3] recently created a causal model using fuzzy logic. In Faghihi's theory, instead of cutting the relationships between the treatment and its confounding parents, the model assigns fuzzy values such as very low, low, medium and high to the Treatment. Then, using probabilistic fuzzy logic theories such as [1] the model calculates the causes of the events. In our second publication, we showed that Faghihi's model is more concise that Pearl's.

Autoencoder: Autoencoder (AE) is a type of generative artificial neural network that learns representations for a set of data in an unsupervised manner [5].

Variational Autoencoder: A variational autoencoder (VAE) is a type of generative neural network that learns to encode and decode data by representing it in a lower-dimensional latent space. The encoder part of the VAE maps the input data to a probability distribution in the latent space, and the decoder part generates data from a sample drawn from that distribution. The key innovation of the VAE is the introduction of a probabilistic component that allows the model to learn a more robust and interpretable latent representation. During training, the model optimizes the trade-off between minimizing the reconstruction error of the decoded data and maximizing the similarity between the learned distribution in the latent space and a known prior distribution. The resulting VAE can be used for tasks such as image generation, anomaly detection, and data compression [5].

Causal Effect Variational Autoencoder: Causal Effect Variational Autoencoder (CEVAE) is a type of VAE that integrates causal rules such as Pearl or Rubin [2] to estimates individual and average treatment effects for unobserved confounders [6].

In our second publications, we propose CEVAE equipped with fuzzy logic rules. The proposed architecture demonstrates promising results in extracting causal relationships from observational data.

Fuzzy causal rules: We explain fuzzy causal rules with the following example. For instance, max (a, 1-b). Let a and b be two variables. The rule finds the maximum value of variable a and 1-b [3].

Fuzzy rules are a set of if-then statements to describe causal relationships between input and output variables. Each statement consists of a condition (the "if" part) and a conclusion (the "then" part).

The condition part of the rule specifies the input variables and their degree of membership to the fuzzy sets. The conclusion part of the rule specifies the output variable and its degree of membership to a fuzzy set.

Fuzzy causal rules are used in various fields such as control systems, expert systems, and decision-making systems. They are particularly useful when dealing with complex and uncertain systems where precise information is not available.

My Contributions. During my master’s program, I have been working on the optimization of Electroconvulsive therapy (ECT) sessions using Deep Learning algorithms [3]. To do so, we have been collaborating with Dr. François- Xavier Roucaut who is a psychiatrist at the St-Marie Hospital at Trois-Rivieres in collaboration with the University Institute in Mental Health of Québec and St-Jerome Hospital.

My contributions are the followings:

I created a dataset by gathering anonymized ECT sessions’ data from the above three hospitals.

For every patient, each ECT session has an EEG file and a file having patients' personal information such as age, sex and retrospective ECT responses which is filled manually. I wrote a Python code that automatically reads, cleans, matches EEG files with their corresponding above patient's information file, evaluates the information correctness in files (e.g., whether the file is corrupted), and builds a custom train-test set from the raw data (GitHub).

Since, EEG data contains artefacts and features and information that Deep Learning algorithms cannot process directly [4, 5], I tried different techniques (Moving average [6], Fast Fourier Transform (FFT) [6], Cross-frequency analysis [6] and wavelet transforms [6], and among the mentioned techniques, FFT [6] gave us the best result.

To detect successful and unsuccessful ECT sessions, I used a Convolutional Neural Network and Long Short-term neural network (CNN-LSTM). The CNN-LSTM model achieved 83% accuracy for classification of good and bad ECT sessions.

However, the CNN-LSTM cannot explain why some ECT sessions are successful while others no. However, finding the root causes of successful and unsuccessful ECTs are essential as it will result in creating individualized treatments for MDDs. The

ore, to equip DLs with reasoning capabilities, I added causal fuzzy rules from Faghihi et al [1] to the Causal Effect Variational Autoencoder (CEVAE) [10] and created the FCEVAE architecture [4]. Adding reasoning to CEVAE, we achieved:

1) Possible causal patterns recognition in pre – shock EEGs,
2) Individualized ECT dose prediction with 91% accuracy which is not included here but can be found in in the following publications. To see the result the reader is referred to (GitHub)

The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-19)

Using Deep Learning algorithms to detect the success or failure of the Electroconvulsive Therapy (ECT) sessions

[1]**Cyrus Kalantarpour**, [1]**Usef Faghihi, [2]François-Xavier Roucaut**,
[1]Université du Québec à Trois-Rivière
[2]Université du Québec à Montréal
cyrus.kalantarpour@uqtr.ca, usef.faghih@uqtr.ca,francois-xavier_roucaut@ssss.gouv.qc.ca

Abstract

Major Depression Disorder (MDD) is a big problem in our society. MDD can cause suicide and take families apart. When treatment with medications fail, mental healthcare professionals, use Electroconvulsive Therapy (ECT) to treat patients with MDD. During an ECT session, electroencephalogram (EEG) signals let the mental healthcare professionals record patients' brain activities which are helpful to decide whether the treatment was successful. However, there is no standard way to know how and with what intensity a healthcare professional needs to apply electroshock to treat patients with MDD. So far, to our knowledge, researchers have used multi-parametric magnetic resonance imaging (MRI) techniques combined with statistical methods and/or linear machine learning algorithms to predict patients' responses to ECT. However, the aforementioned methods are very expensive and time-consuming. In this study, we will be using Deep learning algorithms to detect the effectiveness of ECT sessions based on the EEG.

1 Introduction

Major Depression Disorder (MDD) is the cause of more than one million suicide per year (Sun et al. 2020). Electroconvulsive (ECT) therapy has been used by mental healthcare professionals since 1930 to treat patients with MDD (Tsuchiyama et al. 2005). Yet, there is no methodological technique to individualize ECT in order to obtain successful results. So far, most of the researchers used multi-parametric magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), and/or resting state-fMRI techniques combined with statistical methods and/or support vector regression model (Gong et al. 2020, Gärtner et al. 2020, Van Waarde et al. 2015), to categorize and predict the success or failure of the ECT method with 67-70% accuracy (Sun et al. 2020). Furthermore, most of the aforementioned studies suggest frontal and temporal networks of the brain as good predictors for the success or failure of ECT treatments. In another study, (Min et al. 2020), limited their experience to MDD patients suffering from schizophrenia. They used a random-forest algorithm with fMRI for their predictions. Their results suggest higher connectivity in the patient's frontal area with 82% accuracy of predicting successful and unsuccessful ECTs. Another limitation with this study is that the authors only used patients with schizophrenia.
Furthermore, using MRI/ fMRI/rs-fMRI with ECT is very expensive and time consuming. In this study, we suggest the use of Deep learning algorithms to predict the result of the ECT from patient pre-during-post shock EEGs. Comparing to MRI family techniques mentioned above our technique uses patients' brain EEGs pre-shock to predict the success or failure of the ECT techniques. This makes our approach less expensive and timely efficient comparing to the above techniques. Deep learning algorithms are capable of learning unseen patterns (Faghihi et al. 2020, Robert et al. 2020). We believe that DLs can help us to individualize ECT techniques for patients suffering from MDDs.

Among others, they are used for detecting different types of cancers (Cruz-Roa et al. 2013), sentiment analysis (Baziotis, Pelekis and Doulkeridis 2017), detecting cataract (Yu et al. 2019).
However, to our knowledge, so far, there is no study used deep learning algorithms for predicting the success or failure of the ECT technique only based on EEG data. One reason is that EEGs are very complex and DLs cannot use EEGs directly (Hu and Zhang 2019). Furthermore, while some of the mental health professionals may use lateral regions of the brain for the ECT shocks, others may use lateral and frontal regions. This will result in DLs behave differently when the nature of data changes (Chen et al. 2020).

Thus, before applying DLs to data, one needs to do preprocessing and adapting the data in a way that DLs can process them.

In what follows, we will first very briefly explain EEGs preprocessing techniques. We will also explain how one can distinguish between a good and bad ECT using EEG traces. We then, very briefly explain DL architecture we used in this study. Finally, we will compare different DLs performance for ECT outcome prediction.

2 Pre-Processing EEGs

This section is divided into two subsections 1) using correlation technique to find whether there are correlations between Hyperpolarization, Depolarization, and Repolarization phases and post-shock phases; 2) Using noise reduction techniques to prepare our data for deep learning algorithms.

2.1 Correlation

In the first phase of this study, we wanted to test whether there are correlations between EEGs segments using cross-correlation technique (explained below). Another technique that is widely used in the field of channel processing is Fourier transform (Hu and Zhang 2019). Roughly speaking, Fourier transform breaks a channel into an alternative representation that is characterized by sinus and cosines. However, using Fourier transform may result in losing an important portion of data.

Before explaining cross-correlation technique, we will explain very briefly EEG records of an ECT experiment. An EEG signal can be divided into pre-during-post shock phases.

Figure 1, shows the during shocking phase of an EEG channel after applying electroshock to an MDD patient's scalp. The during shocking phase starts with a hyperpolarization phase, followed by repolarization and, then depolarization phases. In Figure 1, the x-axis corresponds to time and the y-axis corresponds to the amplitude of the EEG channel. The depolarization phase finishes by oscillation around zero on the x-axis. It is worth mentioning that the trace in Figure 1 is the average output of about one hundred million neurons[1] activities after applying an ECT shock to the patient's brain.

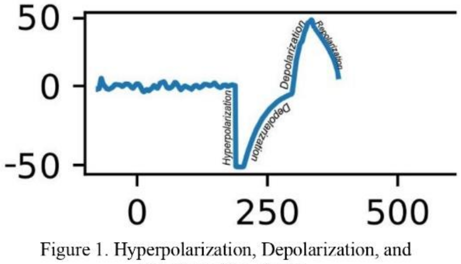


Figure 1. Hyperpolarization, Depolarization, and Repolarization phases.

Figure 2, demonstrates the last part of two complete EEG channels gathered from two patients' scalps during the electroshock procedure. The EEGs contain a patient's pre-during-post shock phases. To distinguish the good and bad ECTs, healthcare professionals use different criterion such as the quality of the pattern of the crisis, and/or the length of the crisis and/or the smoothness of the end of the EEG channel. If the end of the channels becomes smooth as demonstrated in Figure 2.A, the ECT session is considered successful.

However, if the end of the channel is not smooth (Figure 2.B), the electroshock procedure is considered unsuccessful. In this article, for convenience, the EEG associated with a good ECT test is called a good EEG and vice versa.

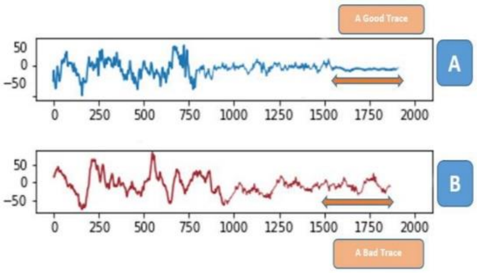


Figure2: The horizontal axis is time, and its length equals 2 seconds (2000ms). The vertical axis is amplitude. The last segment (which is underlined by an orange arrow) determines to what extent the trace is good or bad.

Postulated by one of our mental healthcare professional colleagues, the first hypothesis in our work was that there should be some type of correlation between EEGs' Hyperpolarization, Depolarization, Repolarization phases (Figure 1), and final phase of the EEG (Figure 2, the interval between 1500-2000). In our EEG database, each EEG file contains up to 4 channels. That is, to apply the ECT, our healthcare professional used two electrodes on the frontal lobe, and two on the temporal lobe, symmetrically.

[1] The human brain has about 100 billion neurons and an area of 1,200 square centimeters. Given that the area of an electrode is about one square centimeter. Therefore each electrode records the electrical activity of approximately 83 million neurons.

To verify our above hypothesis, we used the cross-correlation technique. The cross-correlation between two EEG channels measures the level of dependency between hyperpolarizations (HP), depolarizations (DP), and the repolarizations (RP) phases of the channels.

More specifically, If the cross-correlation of two EEG channels is 1 at a time t_0, the EEGs are either hyperpolarized or depolarized, or repolarized. That is, the amplitudes for the two channels are equal.

Similarly, if the cross-correlation between two EEG channels is 0.8 at time t_1, then the EEGs are hyperpolarized, depolarized, or repolarized at the t_1 and the amplitudes of their HP or RP or DP are equal to 80%.

Any negative correlation means the channels are correlated but their behavior is the opposite. That is if the HP value is increasing the RP or DP values are decreasing. Furthermore, the values around zero means EEG channels are acting independently (Dowdy, Wearden, & Chilko, 2011).

The EEGs we used in this study are gathered on a daily basis at the St-Marie hospital at Trois-Rivières (QC) by mental health professionals and then anonymized for research usage. So far, we have 290 EEGs traces. The gathering of EEGs by mental health professionals at the St-Marie hospital continues as the more EEG we have, the better results we obtain from our neural network (explained below). We applied cross-correlation technique to the above 290 EEG traces.

Every EEG has an average length of 75,000 milliseconds (75sec) in total. However, according to our hypothesis, we need to extract the HP , RP and DP parts of the EEGs. Once extracted, we apply cross-correlation technique to the aforementioned parts of the EEGs and the last part of the EEGs which contain the successful or failed ECT (Figure 2).

Because the total average of EEG files' length is 75 seconds, we need to divide them into smaller segments. In this article, we split the EEG files into two-seconds segments.

Therefore, we have (75÷2≈37)×(75÷2≈37)×202 cross-correlation plots. That is, after applying cross-correlation technique to the segmented EEGs, we will obtain 202×37×37 plots. As an example, Figure 3, shows the plot of two EEG segments with the length of two seconds (Figure 3. A and Figure 3. B).

Instead of 2 seconds, one can divide the EEGs into larger or smaller segments. Choosing smaller numbers than the number 2, makes cross-correlations comparison meaningless. That is, the segment has very little information that decreases the cross-correlation technique's performance. On the other hand, choosing larger numbers make comparisons difficult as every segment contains too much information. After dividing EEG channels into segments of two seconds, we: 1) compared all segments of the good and bad EEGs; 2) extracted the HPs, RPs, DPs, and the last segment of the good and bad EEGs.

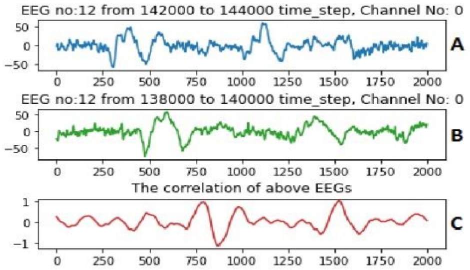


Figure 3: The red subplot (Figure 3.C) demonstrates the cross-correlation of two segments (Figure 3.A, Figure 3.B) from the twelfth EEG signal of our dataset. The horizontal axis is time (ms). In Figure 3.A and Figure 3.B, the vertical axis is EEG amplitude and in Figure 3.C is correlation value.

2.1.1 Comparing two seconds segments

In this subsection, using cross-correlation technique, we compared all the 2-seconds segments we extracted from the entire data set. One problem with the Cross-correlation technique is it compares every segment with themselves which causes redundancy in comparison.

So, in order to avoid calculating duplicate cross-correlations, we filtered the extracted segments so it only considered the unique combination of 2-second segments.

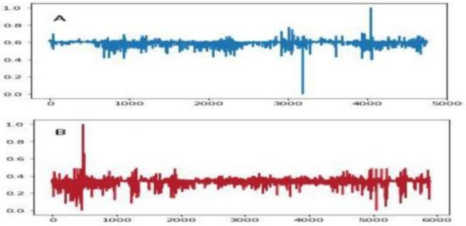


Figure 4: In Figure 4.A and B, the Y-axis shows computed averages of the cross-correlations of different segments for the first channel of the good EEGs (Figure 4.A) and bad EEGs (Figure 4.B). The X-axis shows the EEG segment pair numbers.

We then, averaged the cross-correlations of every 2-second segments for the entire data set. In our case every comparisons produced one single point which is demonstrated in Figure 4.A,B. Figure 4.A, B shows the result of the filtering process and averaging of the whole data set which resulted in more than 5000 points (each considered a point in the Figure 4).

It must be noted that according to the differences between channel size (Figure 4.A,B good EEGs with 5000 and bad EEG with 6000 length), we obtain different data points on the x-axis.

In Figure 4, the average correlation for the good EEGs is equal to 0.6 and 0.25 for the bad EEGs (maximum should be equal to 1).

In Figure 5, we see the average cross-correlation for the second channel of the good EEGs and bad EEGs which are equal to 0.6 and 0.4 respectively. We obtained similar results for the third and fourth channels. Our results show that there are some correlations in general between EEGs segments. However, we would like to remind you that so far, we have not extracted HPs, RPs and DPs from two seconds segments (see below).

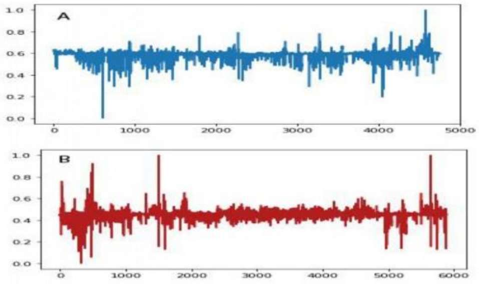


Figure 5: In Figure 5.A and B, The Y-axis shows computed averages of the cross-correlations of different segments for the second channel of the good EEGs (Figure 5.A) and bad EEGs (Figure 5.B). The X-axis shows the EEG segment pair numbers.

It is also worth mentioning that in Figures 4 and 5, at some points the spikes' values are almost one which are considered as outliers in our case. This is because the number of spikes is less than 10, while the total number of calculated cross- correlations is more than 5,000.

In the next subsection, we will analyze the possible correlations of HPs, RPs, DPs with ECT outcome.

2.1.2 HPs, RPs, DPs cross-correlation with ECT results

In this subsection, we will test our mental healthcare professional colleagues' hypothesis which postulates that there should be a logical connection between the HPs, DPs, RPs, and the ECT outcomes.

To do so, we extracted, concatenated, and averaged the HPs, RPs, and DPs segments from the entire EEGs data set.

Since, we obtained very similar results for the cross-correlation of HPs, DPs, RPs, here we will briefly explain RPs (Figure 6).

It must be noted that every EEG contains many HPs, DPs, RPs phases. We extracted and concatenated all RPs from the entire data set and obtained more than 100000 points (Figure 6). We then, calculated the cross-correlation of all RPs and the end of the good and bad EEGs.

Most of the average cross-correlation values in Figure 6 varies between -0.25 and 0.25, which is very low. However, there are some specific points that demonstrate good correlations. The number of these specific points are small comparing to the whole data set. Consequently, they cannot be considered as a solid indicator for the predictions of successful versus unsuccessful ECTs.

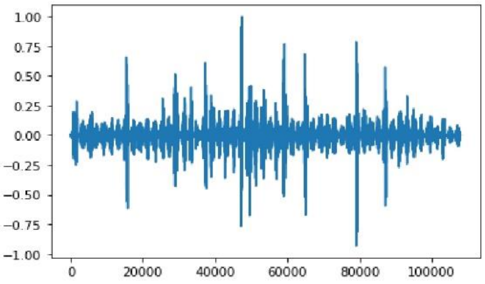


Figure 6: The Y-axis shows computed averages of the cross-correlations of RPs. The X-axis shows the EEG segment pair numbers.

We obtained similar results for HPs and DPs. That is, HPs, RPs, DPs cannot be considered as the good predictors of the successful and unsuccessful of ECTs (see the link to the code).

Our next hypothesis was the use of Deep Learning algorithms (DLs) that uses patients' EEGs pre-shock and during the shock phase in order to predict the successful and unsuccessful ECTs.

However, EEGs contain noises which degrade substantially DLs performance. Thus, we must ideally delete or reduce the noises. For instance, patients' fast or low winking results in different EEGs (Hu and Zhang 2019). The noises directly affect DLs performance. In a preliminary experiment, we directly applied Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) to ***anonymized*** EEGs, without obtaining good results.

Consequently, one crucial problem when processing the EEGs gathered during ECT sessions is how to separate the useful data and noises. Therefore, before we use the DLs to predict the ECTs results, we must reduce/suppress the noises. In the next section, we explain the noise reduction method we used in this study.

2.2 Noise Reduction (NR)

In order to do Noise Reduction (NR) in our data, we used Moving Averages (MAs) or Moving Mean (MM) technique (Booth, Mount and Viers 2006). MA takes as input a dataset and creates many subsets of it. It then, returns the average of the subsets by smoothing subsets' variations. This technique can be seen as noise reduction. Researchers use different versions of MA for noise reduction (Booth et al. 2006). We used a simple version of it (Figure 7):

$$MA_n = \frac{X_{(k-1)*n} + X_{((k-1)*n+1)} + X_{((k-1)*n+2)} + \cdots + X_{k*n}}{k}$$

Formula 1: The Simple Moving average formula

Where:

X_k= The average of input signal amplitudes in ith-period.
n= The nth point of output signal.
k= Length of the periods.

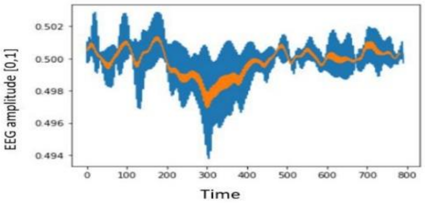


Figure 7: The blue signal is the EEG traces from our dataset and the orange signal is moving average of the blue signal. The x axis is time, and the y axis is EEG amplitude which is between (0, 1)

After applying the noise reduction (SMA), the data is ready to be fed into the Deep Learning algorithms (DLs). In the next section, we will create our DL.

3 Deep Learning algorithms

3.1 Predictor Neural Network Architecture

In this section, we will examine another hypothesis. That is, there is a strong logical relationship between the pre-shock segment of the EEGs and the success or failure of the ECT results. To do so, we need to design and implement a hybrid architecture that is capable of detecting the temporal features of the EEGs and the relationships between different phases of the EEGs.

Our hybride DL architecture (Figure 8) uses Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997), Convolutiona Neural Networks (CNN) (Sainath et al. 2020), and Multi-layer Perceptron (Gardner and Dorling 1998).

In this step, we used the same 2-second segments we used in subsection 2.1.1. Out of 290 EEG traces divided into 2-second segments, we considered 70% for training and 30% for test.

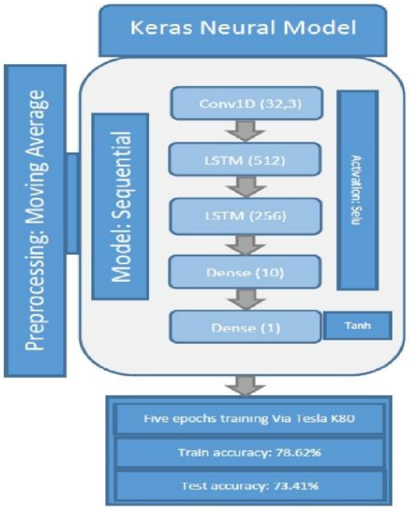
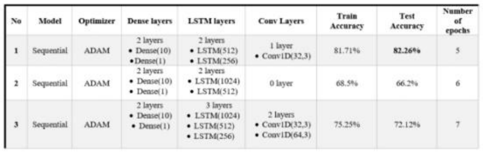


Figure 8: The Predictor CNN- Neural Architecture

In Figure 8, the denoised and averaged data (see previous section) is first fed into a one dimensional CNN (32 neurons and a kernel with size 3). It then fed into two LSTMs-*first having 512 neurons and the second 256 neurons.*

Table 1 shows different configurations for our hybrid DL architecture. We obtained the best performance (82%) using the configuration in the first line in Table 1.



No	Model	Optimizer	Dense layers	LSTM layers	Conv Layers	Train Accuracy	Test Accuracy	Number of epochs
1	Sequential	ADAM	2 layers • Dense(10) •Dense(1)	2 layers • LSTM(512) • LSTM(256)	1 layer • Conv1D(32,3)	81.71%	**82.26%**	5
2	Sequential	ADAM	2 layers • Dense(10) • Dense(1)	2 layers • LSTM(1024) • LSTM(512)	0 layer	68.5%	66.2%	6
3	Sequential	ADAM	2 layers • Dense(10) • Dense(1)	3 layers • LSTM(1024) • LSTM(512) • LSTM(256)	2 layers • Conv1D(32,3) • Conv1D(64,3)	75.25%	72.12%	7

Table1: Different configurations of our DL architecture

4 Conclusion

Currently, mental healthcare professionals (MHP) use trial and error method for the Electroconvulsive Therapy (ECTs) sessions. Consequently, a patient may experience many ECTs before noticing some results. This is a waste of time and resources. In this study, using only EEG traces gathered from MDD patients' scalps, we used hybrid Deep learning algorithms to predict successful and unsuccessful ECTs. This is the first attempt toward creating methodological technique for individualized ECTs.

It must be noted that we did not used expensive techniques such as MRI/ fMRI/rs-fMRI that are used in previous studies. Deep learning algorithms are capable of learning unseen patterns. Although, our data set was small, we obtained 82% precision for detecting good and bad ECTs. Thus, we believe DLs can help us to create a methodological approach to individualize ECT techniques for patients suffering from MDDs.

At this point, we demonstrated our results to other mental healthcare professionals who offered their support to this project by giving us more EEG traces gathered from MDD patients' scalps. Having more EEG traces will improve our DLs precision and prediction capability.

One problem with our current EEG files is that they do not have the degree to which ECTs are applied to the MDD patients' scalps. Our future work will be to change our current DL's architecture so by merely having the MDD patients pre-shock data it can assist MHPs to what degree they need to apply ECTs in order to get successful results.

5 References

Baziotis, C., N. Pelekis & C. Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 747-754.

Booth, E., J. Mount & J. H. Viers (2006) Hydrologic variability of the Cosumnes River floodplain. *San Francisco Estuary and Watershed Science,* 4.

Chen, H., J. D. Janizek, S. Lundberg & S.-I. Lee (2020) True to the Model or True to the Data? *arXiv preprint arXiv:2006.16234.*

Cruz-Roa, A. A., J. E. A. Ovalle, A. Madabhushi & F. A. G. Osorio. 2013. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 403-410. Springer.

Faghihi, U., S. Robert, P. Poirier & Y. Barkaoui (2020) From Association to Reasoning, an Alternative to Pearl's Causal Reasoning. *In Proceedings of AAAI-FLAIRS 2020. North-Miami-Beach (Florida).*

Gardner, M. W. & S. Dorling (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment,* 32, 2627-2636.

Gärtner, M., E. Ghisu, A. L. Herrera-Melendez, M. Koslowski, S. Aust, P. Asbach, C. Otte, F. Regen, I. Heuser & K. Borgwardt (2020) Using routine MRI data of depressed patients to predict individual responses to electroconvulsive therapy. *Experimental Neurology*, 113505.

Gong, J., L.-B. Cui, Y.-B. Xi, Y.-S. Zhao, X.-J. Yang, Z.-l. Xu, J.-B. Sun, P. Liu, J. Jia & P. Li (2020) Predicting response to electroconvulsive therapy combined with antipsychotics in schizophrenia using multi-parametric magnetic resonance imaging. *Schizophrenia research,* 216, 262-271.

Hochreiter, S. & J. Schmidhuber (1997) Long short-term memory. *Neural computation,* 9, 1735-1780.

Hu, L. & Z. Zhang. 2019. *EEG Signal Processing and Feature Extraction*. Springer.

Min, B., M. Kim, J. Lee, J.-I. Byun, K. Chu, K.-Y. Jung, S. K. Lee & J. S. Kwon (2020) Prediction of individual responses to electroconvulsive therapy in patients with schizophrenia: Machine learning analysis of resting-state electroencephalography. *Schizophrenia research,* 216, 147-153.

Robert, S., U. Faghihi, Y. Barkaoui & N. Ghazzali (2020) Causality in Probabilistic Fuzzy Logic and Alternative Causes as Fuzzy Duals. *ICCCI 2020,* Ngoc-Thanh Nguyen, et al.

Sainath, T. N., A. W. Senior, O. Vinyals & H. Sak. 2020. Convolutional, long short-term memory, fully connected deep neural networks. Google Patents.

Sun, H., R. Jiang, S. Qi, K. L. Narr, B. S. Wade, J. Upston, R. Espinoza, T. Jones, V. D. Calhoun & C. C. Abbott (2020) Preliminary prediction of individual response to electroconvulsive therapy using whole-brain functional magnetic resonance imaging data. *NeuroImage: Clinical,* 26, 102080.

Tsuchiyama, K., H. Nagayama, K. Yamada, K. Isogawa, S. Katsuragi & A. Kiyota (2005) Predicting efficacy of electroconvulsive therapy in major depressive disorder. *Psychiatry and clinical neurosciences,* 59, 546-550.

Van Waarde, J., H. Scholte, L. Van Oudheusden, B. Verwey, D. Denys & G. Van Wingen (2015) A functional MRI marker may predict the outcome of electroconvulsive therapy in severe and treatment-resistant depression. *Molecular psychiatry,* 20, 609-614.

Yu, F., G. S. Croso, T. S. Kim, Z. Song, F. Parker, G. D. Hager, A. Reiter, S. S. Vedula, H. Ali & S. Sikder (2019) Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques. *JAMA network open,* 2, e191860-e191860.

Causal Probabilistic Based Variational Autoencoders Capable of Handling Noisy Inputs Using Fuzzy Logic Rules

Usef Faghihi[1](✉), Cyrus Kalantarpour[1], and Amir Saki[2]

[1] University of Quebec, Trois-Rivières, QC, Canada
{usef.faghihi,cyrus.kalantarpour}@uqtr.ca
[2] Institute for Research in Fundamental Sciences, Tehran, Iran
amirsaki1369@ipm.ir

Abstract. Researchers and engineers may use inferential logic and/or fuzzy logic to solve real-world causal problems. Inferential logic uses probability theories, while fuzzy logic uses its membership functions and set theories to process uncertainty and fuzziness of the events. To benefit from both logics, some researchers in the past tried to create probabilistic fuzzy logic (PFL). Deep Learning algorithms (DLs) with their incredible achievements such as very high precision results in some specific tasks are at the center of the weak AI. However, DLs fail when it comes to causal reasoning. In order to equip Deep Learning algorithms (DLs) with reasoning capabilities, one solution would be to integrate non-classical logics such as PFL with DLs. In this paper, we will demonstrate the first step toward creating a deep causal probabilistic fuzzy logic architecture capable of reasoning with missing or noisy datasets. To do so, the architecture uses fuzzy theories, probabilistic theories, and deep learning algorithms such as causal effect variational autoencoders.

Keywords: Deep learning · Probabilistic fuzzy logic · Causal reasoning · Autoencoders

1 Introduction

As human beings, we are always in the search for the causes of events around us. For instance, was it the spicy food I had for my lunch that caused my abdominal discomfort? In causal reasoning, one uses previous information about an event or situation to predict its future state. However, discovering the real causes of events is usually difficult.

In order to solve the problem of causality, some researchers use inferential logic, which uses probability theory, while others use fuzzy logic, which outperforms inferential logic [3]. Probability theories deal with the uncertainty of human knowledge about an event. However, there is no gradient possible with probability theories [3]. Fuzzy logic makes it possible to take into account the vagueness of events. The ideal would be to use both inferential and fuzzy logic theories together and create probabilistic fuzzy logic [3]. In [3], Faghihi et al. used causal fuzzy rules belonging to fuzzy rule sets to

K. Arai (Ed.): SAI 2022, LNNS 507, pp. 190–202, 2022.
https://doi.org/10.1007/978-3-031-10464-0_12

find the influences of confounders on other variables. A confounder variable influences both dependent and independent variables causing fake correlations between variables. However, Faghihi et al.'s model does not have learning capabilities [3]. To equip PFL with learning capabilities, one must integrate them into the DLs [4]. One powerful generative deep learning algorithm that is widely used to deal with different real-world problems is Variational Autoencoders architectures families [2, 5]. In the following, we briefly explain Autoencoders, Variational autoencoders, and Causal Effect Variational Autoencoders [2].

Autoencoder (AE) is a specific type of generative artificial neural network that learns representation for a set of data in an unsupervised manner. An AE [2] has 1) Encoder module (inference network in causality context) which encodes or compresses the input data into a latent space representation (a reduced version of the original data); 2) Decoder module which tries to reconstruct the original input data from the latent encoded space.

Variational Autoencoders (VAEs) are similar to the AEs, except they consider a family of Gaussian distributions while sampling from the input data. VAEs work with both continuous and discrete data. Recently, researchers created Causal Variational Autoencoders (CEVAE) [2], which estimate the individual and average treatment effects (ITE, and ATE respectively) for unobserved confounders using proxy variables which are replacements for confounders [3].

In this paper, we first discuss the related works on causality using different DLs such as variational autoencoder families. In order to extract causal relationships from observational data, we then discuss two architectures that use PFL and causal effect variational autoencoders (CEVAEs) architecture [2] which we call FCEVAE-V1 and V2. The first architecture is called FCEVAE -V1: in this architecture PFL and CEVAE each separately applied to the dataset. That is, we cluster and fuzzify the data set, and then, it is feed-forwarded to the CEVAE architecture (Fig. 1). The second architecture is called FCEVAE -V2: we integrated association and causal rules from [3] to the CEVAE loss function (Fig. 2). That is, in the second architecture, we equipped CEVAE with a modified loss function that implements causal fuzzy logic rules from [3]. It must be noted that the initial CEVAE architecture developed by Amsterdam lab[1] uses TensorFlow and Edward (deprecated). In this study, we used a CEVAE version equipped with Pyro library[2]. Pyro is faster than Eward.

Finally, we compare the performance of our architecture with similar architectures and discuss the results and limitations of our work.

2 Related Works

Recently, learning causal relationships from observational data received lots of attention in the field of Artificial Intelligence [6, 7]. Moreover, some researchers try to address the identifiability[3] issue using neural networks [8]. However, the observational data may contain hidden confounder variables that may not or very difficult to be measured [2].

[1] https://github.com/AMLab-Amsterdam/CEVAE.

[2] https://github.com/pyro-ppl/pyro.

[3] If the true parameters of a statistical model can be learned after observing sufficient number of observations, the model is said to be identifiable. Wikipedia.

Take a study in which we are interested in individualized medicine and where we have to figure out the best medication for a patient from observed data [2]. In this example, the socio-economic status of the patient can influence the type of medication the patient has access to and her general health [2]. That is, the socio-economic status is a confounder, and we cannot compute its value [2]. It is worth mentioning that once we can estimate or calculate the confounder's value, another hurdle to overcome is to find which element(s) it influences the most.

Let's suppose we cannot measure the confounder which is the socioeconomic status of the patient. Roughly speaking, there are two main approaches to calculating confounders. The first one is a tree-based approach [9], wherein the authors use Bayesian Additive Regression Trees (BART) [10] to estimate average causal effects for solving causal inference problems such as individual treatment effects (ITE). The second approach uses Directed Acyclic Graphs (DAGs) as a causal structure and a Bayesian approach for reasoning.

TARnet [11] is one of the first architectures that is used for causal inference. It is based on weighting optimizations and using the feed-forward neural networks. However, TARnet is not robust enough to deal with noisy datasets [2, 12]. In 2017, Louizos et al. [2] created Causal Effect Variational Autoencoders (CEVAE) which estimates the individual and average treatment effects (ITE and ATE) for unobserved confounders using proxy variables. A confounder variable that can be hidden and/or have missing data, influences both dependent and independent variables, causing fake correlations between variables. The model suggested by the authors in [2] outperformed Tree-based approaches such as BART [2]. However, the model in [2] has problems with processing missing data.

To improve CEVAE, the authors in [12] created Identifiable VAE (iVAE) architecture. This architecture postulates that different model parameters must lead to the different marginal densities for the data. In 2021, the authors suggested Intact VAE [13], an improved version of iVAE. Intact VAE estimates ATE by using a modified version of propensity score (the probability of a subject receiving treatment) and B-score (The conditional distribution for the covariates receiving or not receiving treatment is the same). However, this study ignores computing confounders. As opposed to current DLs which can only process either noisy or missing data, a robust DL needs to be both tolerant to both noisy and missing data with hidden confounders. We will achieve this by integrating Non-Classical Logics such as probabilistic fuzzy logic rules with DLs.

Faghihi et al. [1] argued that in most real-life problems, the communication between nodes is two-way, something DAG does not support. In other words, the mere Bayesian approach to causation cannot answer the following problem: what is the probability that socio-economic status influences the type of medication the patient has access to and her general health, and to what degree?

Probabilistic Fuzzy logic (PFL), on the other hand, excels at reasoning with degrees of certainty and in real-life problems [14]. Importantly, this allows for degrees of dependency and membership. In PFL, Zadeh [14] proposes that a given set of elements always has a degree of membership and fits into an interval between [0,1]. PFL processes three types of uncertainty: randomness, probabilistic uncertainty, and fuzziness.

PFL can both manage the uncertainty of our knowledge (by the use of probabilities) and the vagueness inherent to the world's complexity (by data fuzzification) [14]. PFL has been used to solve many engineering problems, such as security intrusion detection [15, 16] and finding the causes of the events. However, PFL cannot learn by itself and needs experts to define intervals before applying fuzzification [3]. In [3], the authors used more than ten PFL rules to discover the causal relationship between variables from observational data. However, logic cannot learn a representation of the data [3]. One solution would be to integrate PFL with Deep Learning algorithms or use them in parallel. In the next section we explain how we used CEVAE architecture [2] with PFLs.

3 Fuzzy Cevae

We designed and implemented two versions of the CEVAE architecture [2] which we call FCEVAE: 1) FCEVAE-V1: in this architecture PFL and CEVAE separately applied to the dataset. That is, we clustered and fuzzified the dataset using PFL and then feed forwarded it to the CEVAE architecture (Fig. 1); and 2) FCEVAE-V2: in this architecture, we integrated clustering and causal rules with the CEVAE architecture (Fig. 2). That is, in the second architecture, we equipped CEVAE with a modified loss function that implements causal fuzzy logic rules from [3].

3.1 Fuzzy Causal Effect Variational Autoencoder (FCEVAE-V1) First Architecture

To cluster the dataset into “Low”, “Average”, and “High” clusters, we used the fuzzy c-mean algorithm [1] (Fig. 1A). It is worth mentioning that depending on the problem, one can use more than three clusters if needed.

However, the C-mean clustering only gives us the membership belongingness of the dataset elements to every clusters. Thus, C-mean’s output does not include any information about the nature of the dataset. Consequently, we multiplied the clustered

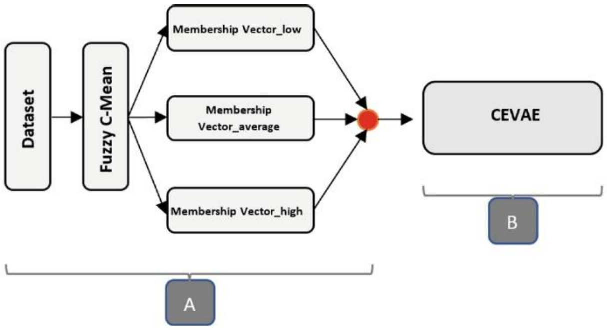


Fig. 1. Part A is a membership vector extractor according to the fuzzy C-mean algorithm [1]. It applies fuzzy c-mean on the dataset and computes memberships vectors for the elements of the dataset. Then, the red circle, which in our context is a ‘switch neuron’, selectively multiplies the memberships vectors calculated in the previous step to the original dataset. Part B is the original CEVAE proposed in [2].

data with the original dataset. This gave us a weighted fuzzy representation of the dataset elements describing how well each element belongs to the fuzzy clusters.

To justify our above multiplication, we briefly explain our Simple Probabilistic Fuzzy Logic Theory (SPFL) theory [17], a classical probability theory mainly useful for the problems with fuzzy concepts in their nature. For instance, the following problem could be solved by using the SPFL theory.

Question. Suppose the fuzzy attribute *Large* for the set $X = (1,...,20)$. In an experiment, what is the probability of randomly selecting 17 from X as Large?".

To answer the question, one can define random variable $\xi_{X,Large}$ so that $P(\xi_{X,Large} = 17)$. Hence, here the distribution of $\xi_{X,Large}$ matters. Randomly selecting the elements of X comes from the nature of the distribution on X, while selecting as Large comes from a two steps procedure consisting of fuzzifying data by some fuzzy attributes including Large, and then the distribution determining the chance of being selected as *Large*.

Another example follows:

Question. Suppose the fuzzy attribute *Large* for the set $X = (1,...,20)$. In an experiment, we are given $X = 17$. What is the probability of selecting 17 as Large?

The answer to this question is $\mathbb{P}(\mathrm{xisLarge})$, and it comes from a distribution. Now, a binary random variable is considered as follows:

$$\xi_{x,Large} = \begin{cases} x, & \mathbb{P}(\mathrm{xisLarge}) \\ 0, & 1 - \mathbb{P}(\mathrm{xisLarge}) \end{cases}$$

That is, $\mathbb{E}(\xi_{\mathrm{x,Large}}) = \mathrm{x}\mathbb{P}(\mathrm{x\ is\ Large})$, and it is interpreted as the quantity of x as being *Large.* Note that, in this paper we use a model with $\mathbb{P}(\mathrm{x\ is\ Large}) = \mu_{\mathrm{Large}}(\mathrm{x})$.

To calculate **Fuzzy Average Treatment Effect (FATE)** which will be used in our below FCEVAE (second architecture), we perform as follows: Suppose X, T and Y are the covariate, the treatment and the outcome of an experiment, respectively. Let A be a fuzzy attribute of X. We define the *fuzzy individual treatment effect* of any $X = x$ with respect to A as: $\mathrm{FITE}_A(x) = \mathrm{ITE}(\mathbb{E}(\xi_{\mathrm{x,A}}))$.

It follows that:

$$\mathrm{FITE}_A(x) = \mathbb{E}(\mathrm{Y|X = x}\mu_{\mathrm{A}}(\mathrm{x}),\mathrm{do(T=1)}) - \mathbb{E}(\mathrm{Y| = x}\mu_{\mathrm{A}}(\mathrm{x}),\mathrm{do(T=0)}).$$

Now, we define FATE of X with respect to A as $\mathrm{AFTE}_A(X) = \mathbb{E}_{\mathrm{X}}(\mathrm{FITE}_{\mathrm{A}}(\mathrm{X}))$.

Going back to the FCEVAE-V1 architecture (Fig. 1), by multiplying the clustered data with the original dataset, we obtain a weighted fuzzy representation of the dataset elements describing how well each element belongs to fuzzy clusters. As a result, FCEVAE-V1 produces three different average treatment effects values [18] each describing the fuzzy average treatment effect corresponding to the clusters such as "low", "average", and "high". Table 1 shows that FCEVAE-V1 outperforms Microsoft's DoWhy[4] project that implements Pearl's causal architecture [19] on Infant Health and Development Program (IHDP) [9] dataset. IHDP dataset contains information about the effect of specialists' home visits on premature infants' cognitive test scores [6]. In addition, the average number of ATEs obtained by FCEVAE-V1 is 4.006. This value is closer to the real IHDP's ATE of 4.021 [19]. It is worth noting that in the DoWhy project, the

[4] https://microsoft.github.io/dowhy/dowhy_ihdp_data_example.html.

ATE value for the IHDP dataset was calculated by subtracting the mean of the treated and controlled groups.

Table 1. Comparison of FCEVAE-v1 and DoWhy.

Cluster	Low	Average	High
FCEVAE-v1	3.812	4.015	4.192
Microsoft DoWhy	3.928		

However, our first architecture has two flaws: 1) similar to the original CEVAE architecture [2], to select the treatment and outcome columns, the architecture needs a human expert. However, in a real-world problem, humans may have no idea about the Treatment and Outcome columns; 2) because we fuzzify the dataset before feeding it to the CEVAE architecture, it cannot tolerate noisy data. We fixed the first architecture's flaws in our second architecture.

Unlike the first architecture that uses fuzzy weighted versions of datasets to create fuzzy-probabilistic-based CEVAE architecture (without using fuzzy rules), the second architecture incorporates fuzzy causal rules from [3, 20] in its loss function. This helps the CEVAE architecture discover the causal relationships between the dataset's elements.

3.2 Fuzzy Causal Effect Variational Autoencoder (FCEVAE-V2) Second Architecture

In order to create an architecture capable of dealing with noisy and missing data, we created FCEVAE-V2 architecture by integrating our fuzzy rules from [3] into the CEVAE's loss function. Our architecture is divided into two main components:

Figure 2, Part A: a conditional autoencoder that randomly generates equally unbiased samples from a dataset.

Figure 2, Part B: it takes the input from the previous step and uses fuzzy causal rules integrated into CEVAE's loss function to extract causal relationships.

We will briefly explain our architecture steps in the following:

Figure 2, Part A: Before explaining **Part A** in detail, we briefly explain the difference between Variational Autoencoders (VAE) and the conditional variational autoencoder (CVAE) we used in Fig. 2. Part A. Whereas the VAE architecture does not apply any condition during sampling from datasets, CVAE uses the conditioning method for the sampling process [21, 22].

The main goal behind the step A is to generate unbiased equal samples without missing data. To do so, Conditional VAE (Fig. 2A) takes a dataset with missing data and generates equal amount of sampling from conditional distribution of the dataset's columns. That is, we create a condition matrix (for which its columns are the output of the Conditional VAE that generates un-biased samples) so that it removes the missing data's bias ratio. For example, assume that for a given dataset $D = (X_0,\ldots,X_n)$, where X_i s are the columns with length $l(X_i)$. We have $M = (m_0,\ldots,m_n)$ where m_i s are the corresponding missing data ratio. We generate a condition matrix $\mathbb{C} = (\mathrm{C}_0,\ldots,\mathrm{C}_\mathrm{n})$, such that C_i s are binary vectors with length $l(X_i)$. If the corresponding dataset's element is missing, each element of C_i, such as c_{ij}, equals 0. Otherwise the value is equal to 1.

$$\mathbf{E}[\mathbf{log}(\mathbf{X}|\mathbf{z},\mathbb{C})]-\mathbf{D_{KL}}[\mathbf{Q}(\mathbf{z}|\mathbf{X},\mathbb{C})\,\|\,\mathbf{P}(\mathbf{z}|\mathbb{C})] \tag{1}$$

Equation (1) is the CVAE's objective function. Q and P are the conditional distribution of CVAE's encoder and decoder respectively. KL is the KL divergence. The model learns P and Q given the condition matrix $\mathbb{C}$ [21, 22].

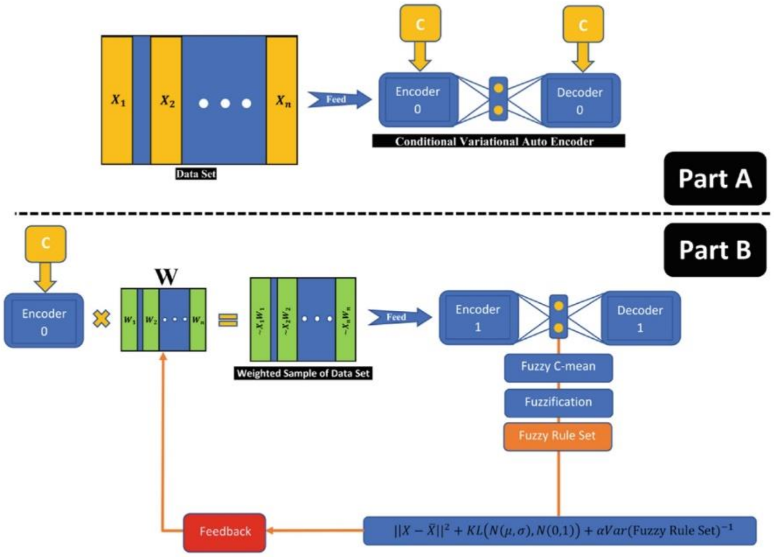


Fig. 2. FCEVAE-V2 where probabilistic fuzzy logic rules are integrated with the CEVAE loss function.

Figure 2, Part B: Part A's output is an unbiased sample S with no missing values. In Part B, we create a matrix W such that its columns will show possible causal relationships in the dataset D's columns (see Table 4). That is, once we calculate W, a higher value in a column (i.e., gestat10 in Table 4) shows a higher influence of the column on the outcome (see Table 4). We must emphasize that contrary to the previous works that used gestat10 as the cause, we used *all* columns to calculate possible causes.

To do so, we first initialized the randomly generated matrix W with size $(n \times n)$ where n is the number of the D columns. An important note is that since the matrix W's values are randomly generated, for different executions we get slightly different values for ITE, ATE and the values in Table 4. We then multiply W (see our above SPFL theory [17]) by the output from Part A. The result of the previous step feedforward into the CEVAE (Fig. 2, Part B).

After encoding the data in Fig. 2, Part B's encoder section, the resulting data is partitioned using the Fuzzy-C mean [1] algorithm. This partitioning is done to automatically find fuzzy membership intervals without the need for an expert to define them. We then use fuzzy rules from [3] to fuzzify the result. The next step is to add fuzzy rules to the CEVAE loss function (Eq. (2)):

$$||\mathbf{X} - \overline{\mathbf{X}}||^2 + \mathbf{KL}(\mathbf{N}(\mu,\sigma),\mathbf{N}(0,1)) + \alpha\mathbf{Var}(\mathbf{Fuzzy\ Ruleset})^{-1} \qquad (2)$$

In the above equation, the first term is the reconstruction error. The second term is KL divergence. The third term calculates the variance of the fuzzy rule set according to the association and causal fuzzy rules from [1]. α ϵ [0, 1] is the training hyper-parameter. It helps the model include the influences of the fuzzy ruleset from [3] to the loss function. For an α = 0, we have the original CEVAE architecture.

The above loss function's output is passed to the back-propagation algorithm to update the W(Fig. 2 Part B, red rectangle).

Table 2. FCEVAE-V2 performance on the noisy IHDP dataset.

Noise	$N{\sim}N(0.10{,}0.5)$	$N{\sim}N(0.15{,}0.5)$	$N{\sim}N(0.20{,}0.5)$
$ATE_{FCEVAE-v2}$	3.27542	2.65458	1.76581
ATE_{CEVAE}	3.35354	2.61252	1.91257
ATE_{Dowhy}	1.99664	1.37661	1.28898

The updated *Ws* are multiplied by the output from Part A. Again, the result are passed to the FCEVAE-V2 where the model applies C-mean and fuzzification and calculates fuzzy loss function before using back-propagation algorithm. FCEVAE-V2 continues the above steps until the result converges to a minimum value for the loss function.

3.3 Second Architecture's Experiments

Similar to the CEVAE project [2], and the DoWhy project [19], we tried FCEVAE-V2 with the IHDP [9] and TWINS [2] datasets. With the TWINS dataset, the goal is to find the possible causal relationships between the weight of twins and their death rate. The main difference between FCEVAE-V2, CEVAE, and DoWhy architectures is that while other architectures add noise to one specific column (gestat10 column), we added noises to the whole dataset. We did this to show that DLs equipped with non-classical logic rules are tolerant to multiple noise source.

After applying FCEVAE-V2 to the IHDP dataset, we obtained similar ATE and ITE values to the CEVAE and Dowhy's project's outputs (see GitHub). To try FCEVAE-V2 with noisy data, similar to [2], we applied the Gaussian noise $N\ N(\mu,0.5)$ where $\mu \in (0.10,0.15,0.20)$ on the IHDP dataset and passed it to FCEVAE-V2 in order to measure the network's noise tolerance level. Table 2 shows that compared to other architectures, our architecture is more tolerant to noises (a lower ATE is better).

We also applied the noise to the TWINS dataset and passed the noisy data to FCEVAE-V2. Table 3 shows that comparing to CEVAE and DoWhy, our model gives lower ATE values.

Table 3. FCEVAE_V2 Performance on Noisy TWINS Dataset

Noise	$N \sim N(0.10,\ 0.5)$	$N \sim N(0.15,\ 0.5)$	$N \sim N(0.20,\ 0.5)$
$ATE_{FCEVAE-v2}$	-0.02616	-0.02711	-0.05121
ATE_{Dowhy}	-0.06901	-0.11760	-0.17351
ATE_{CEVAE}	-0.02720	-0.02931	-0.06245

Similarly to the CEVAE [2] and DoWhy projects, we used the TWINS dataset with FCEVAE-V2. It must be noted that in the previous works the authors only used gestat10 column to calculate the possible cause of the twins' death rate. In this study we used all columns with FCEVAE-V2. Table 4 shows the most important relationships between columns (to see the full result, the reader is referred to[5]). We would like to remind the reader that although we used a heatmap to show the values in Table 4, these values are not correlations and/or covariance matrices. These values are the final values of the matrix W (see above), and they were obtained after using CEVAE's probability approach and many iterations of the c-mean clustering algorithm, fuzzification, and fuzzy rule integration to the CEVAE cost function.

In Table 4, all values belong to the [0, 1] interval. The higher value shows the stronger possible cause between columns. For instance, similar to [2], our model revealed a strong relationship (0.52%) between GESTAT10 and outcome which is one of the highest values in the outcome row. That is, the GESTAT10 column influences many other columns such as adequacy (adequacy of care) and incervix (risk factor, Incompetent cervix).

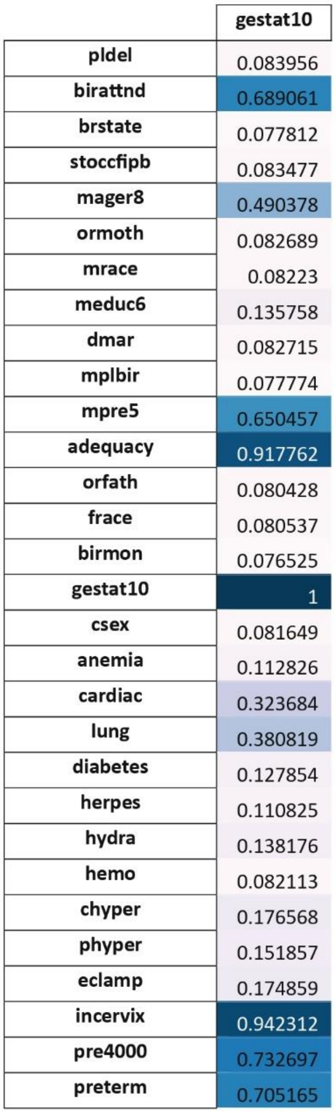
Limitations: Similar to previous work, FCEVAE-V2 is capable of finding the causal relationships between the TWINS dataset columns (TWINS' description). Since our model uses all columns in TWINS, it also found other possible causal relationships between columns that were not mentioned in previous works (Table 5).

We have found some health-related papers that could potentially suggest a scientific foundation for the results generated by our model. For instance:

However, this is only the very surface of what needs to be done next. Given that FCEVAE-V2 uses both probability and fuzzy approaches to calculate the casual relationship between columns in the dataset, at this point, we cannot provide an explanation for how these values are calculated precisely. We aim to do so in our future work. We also encourage the readers to contact us, should they find any explanation for our result (the code is on GitHub (see footnote 5)).

[5] https://github.com/josеffaghihi/Causal-fuzzy-CEVAE/blob/main/2021-12-14/Arch2/ARC2_Final_2021_12_14.ipynb.

Table 4. Partial FCEVAE-V2 output for TWINS Dataset. Each Element $\epsilon[0,1]$ Interval is the causality level of the associated columns and rows from matrix W. The dark blue color shows possible causal relationships. to see the full result, the reader is referred to[6].

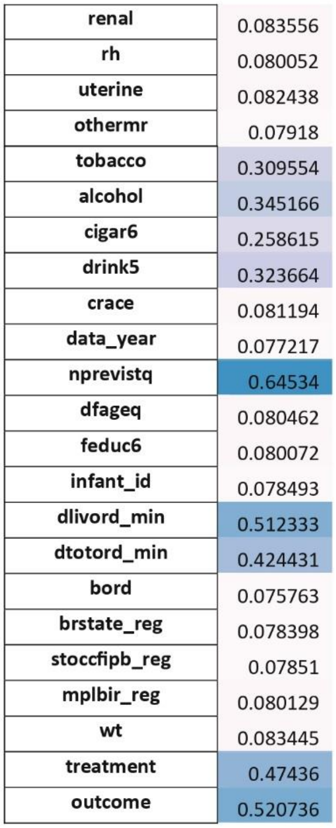


	gestat10
pldel	0.083956
birattnd	0.689061
brstate	0.077812
stoccfipb	0.083477
mager8	0.490378
ormoth	0.082689
mrace	0.08223
meduc6	0.135758
dmar	0.082715
mplbir	0.077774
mpre5	0.650457
adequacy	0.917762
orfath	0.080428
frace	0.080537
birmon	0.076525
gestat10	1
csex	0.081649
anemia	0.112826
cardiac	0.323684
lung	0.380819
diabetes	0.127854
herpes	0.110825
hydra	0.138176
hemo	0.082113
chyper	0.176568
phyper	0.151857
eclamp	0.174859
incervix	0.942312
pre4000	0.732697
preterm	0.705165

(*continued*)

[6] https://github.com/joseffaghihi/Causal-fuzzy-CEVAE/blob/main/2021-12-14/Arch2/ARC2_Final_2021_12_14.ipynb.

Table 4. (*continued*)



renal	0.083556
rh	0.080052
uterine	0.082438
othermr	0.07918
tobacco	0.309554
alcohol	0.345166
cigar6	0.258615
drink5	0.323664
crace	0.081194
data_year	0.077217
nprevistq	0.64534
dfageq	0.080462
feduc6	0.080072
infant_id	0.078493
dlivord_min	0.512333
dtotord_min	0.424431
bord	0.075763
brstate_reg	0.078398
stoccfipb_reg	0.07851
mplbir_reg	0.080129
wt	0.083445
treatment	0.47436
outcome	0.520736

Table 5. TWINS data set columns and their description according to TWINS' Description

TWINS dataset Column name and description		
mpre5 (trimester prenatal care begun)	adequacy (adequacy of care)	[23]
mpre5 (trimester prenatal care begun)	Eclamp (risk factor, Eclampsia)	[24]
mpre5 (trimester prenatal care begun)	Incervix (risk factor, Incompetent cervix)	[25]

4 Conclusion

In this paper, we have shown that Deep Learning algorithms (DLs) equipped with non-classical logics such as PFLs are capable of reasoning with multiple sources of missing data or noise. This had not been done in previous works; only one source of noise was previously used.

To do so, we created two architectures: 1) First, after applying probabilistic fuzzy logic association and causal rules (PFLs) to the dataset, the architecture feedforwarded the output to the Causal Variational Autoencoders (CEVAE) architecture [2]; 2) Second, we integrated PFLs into the CEVAE's loss function. Compared to the Microsoft DoWhy, and the original CEVAE architecture, our FCEVAE-V2 is more tolerant to datasets with missing data and multiple sources of noise.

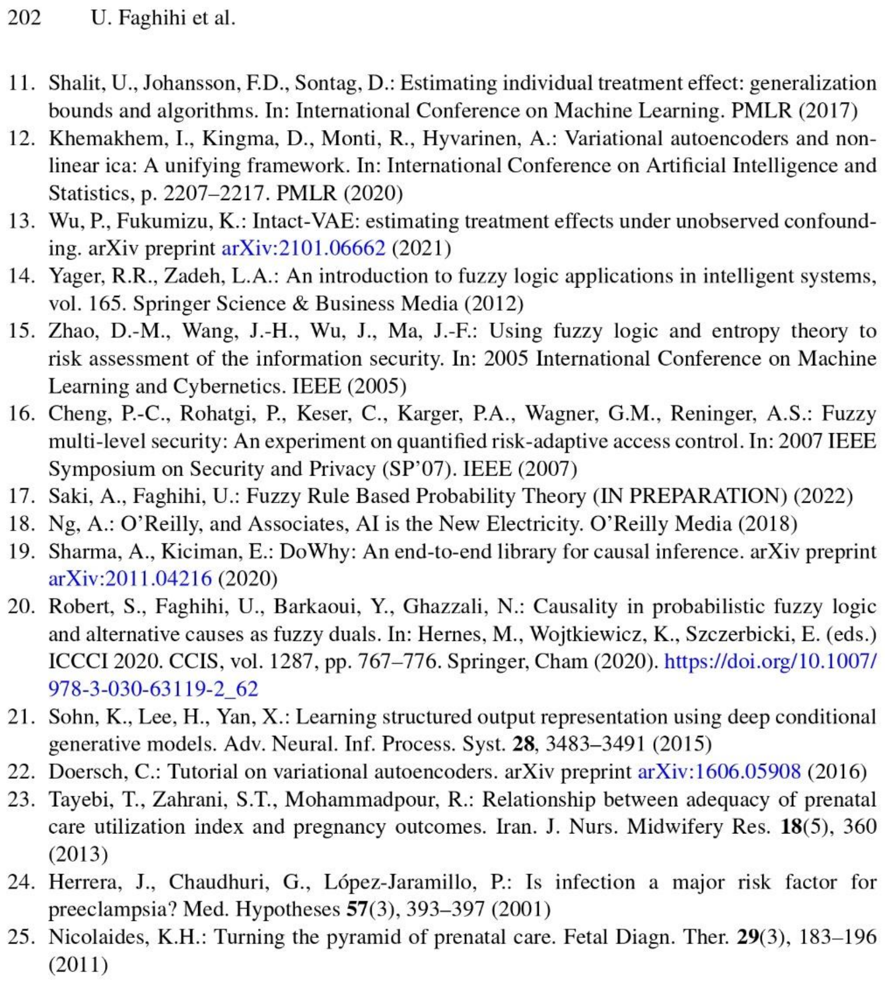
In contrast to the original CEVAE architecture which relies heavily on the treatment column to be determined by human experts, our FCEVAE-V2 does it automatically. That is, in order to reveal possible causal relationships between columns, our model applies causal rules to all columns. To prevent combinatorial problems when selecting treatment FCEVAE-V2 uses the CEVAE compression technique.

Much work remains to be done. An important limitation of our work is explaining the calculations of the causal relationships between columns, and their interpretation in real-life scientific contexts.

Acknowledgments. We thank Sioui Maldonado Bouchard for kindly proofreading this paper.

References

1. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: The fuzzy c-means clustering algorithm. Comput. Geosci. **10**(2–3), 191–203 (1984)
2. Louizos, C., Shalit, U., Mooij, J.M., Sontag, D., Zemel, R., Welling, M.: Causal effect inference with deep latent-variable models. In: Advances in Neural Information Processing Systems, p. 6446–6456 (2017)
3. Faghihi, U., Robert, S., Poirier, P., Barkaoui, Y.: From association to reasoning, an alternative to pearl's causal reasoning. In: Proceedings of AAAI-FLAIRS 2020. North-Miami-Beach (Florida) (2020)
4. Faghihi, U., Maldonado-Bouchard, S., Biskri, I.: Science of data: from correlation to causation. Springer (2021)
5. An, J., Cho, S.: Variational autoencoder based anomaly detection using reconstruction probability. Spec. Lect. IE **2**(1), 1–18 (2015)
6. Yao, L., Li, S., Li, Y., Huai, M., Gao, J., Zhang, A.: Representation learning for treatment effect estimation from observational data. Adv. Neural Inform. Process. Syst. 31 (2018)
7. Guo, R., Cheng, L., Li, J., Hahn, P.R., Liu, H.: A survey of learning causality with data: problems and methods. ACM Comput. Surv. (CSUR) **53**(4), 1–37 (2020)
8. Roeder, G., Metz, L., Kingma, D.: On linear identifiability of learned representations. In: International Conference on Machine Learning. PMLR (2021)
9. Hill, J.L.: Bayesian nonparametric modeling for causal inference. J. Comput. Graph. Stat. **20**(1), 217–240 (2011)
10. Chipman, H.A., George, E.I., McCulloch, R.E.: BART: Bayesian additive regression trees. Ann. Appl. Stat. **4**(1), 266–298 (2010)



11. Shalit, U., Johansson, F.D., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms. In: International Conference on Machine Learning. PMLR (2017)
12. Khemakhem, I., Kingma, D., Monti, R., Hyvarinen, A.: Variational autoencoders and nonlinear ica: A unifying framework. In: International Conference on Artificial Intelligence and Statistics, p. 2207–2217. PMLR (2020)
13. Wu, P., Fukumizu, K.: Intact-VAE: estimating treatment effects under unobserved confounding. arXiv preprint arXiv:2101.06662 (2021)
14. Yager, R.R., Zadeh, L.A.: An introduction to fuzzy logic applications in intelligent systems, vol. 165. Springer Science & Business Media (2012)
15. Zhao, D.-M., Wang, J.-H., Wu, J., Ma, J.-F.: Using fuzzy logic and entropy theory to risk assessment of the information security. In: 2005 International Conference on Machine Learning and Cybernetics. IEEE (2005)
16. Cheng, P.-C., Rohatgi, P., Keser, C., Karger, P.A., Wagner, G.M., Reninger, A.S.: Fuzzy multi-level security: An experiment on quantified risk-adaptive access control. In: 2007 IEEE Symposium on Security and Privacy (SP'07). IEEE (2007)
17. Saki, A., Faghihi, U.: Fuzzy Rule Based Probability Theory (IN PREPARATION) (2022)
18. Ng, A.: O'Reilly, and Associates, AI is the New Electricity. O'Reilly Media (2018)
19. Sharma, A., Kiciman, E.: DoWhy: An end-to-end library for causal inference. arXiv preprint arXiv:2011.04216 (2020)
20. Robert, S., Faghihi, U., Barkaoui, Y., Ghazzali, N.: Causality in probabilistic fuzzy logic and alternative causes as fuzzy duals. In: Hernes, M., Wojtkiewicz, K., Szczerbicki, E. (eds.) ICCCI 2020. CCIS, vol. 1287, pp. 767–776. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63119-2_62
21. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Adv. Neural. Inf. Process. Syst. **28**, 3483–3491 (2015)
22. Doersch, C.: Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908 (2016)
23. Tayebi, T., Zahrani, S.T., Mohammadpour, R.: Relationship between adequacy of prenatal care utilization index and pregnancy outcomes. Iran. J. Nurs. Midwifery Res. **18**(5), 360 (2013)
24. Herrera, J., Chaudhuri, G., López-Jaramillo, P.: Is infection a major risk factor for preeclampsia? Med. Hypotheses **57**(3), 393–397 (2001)
25. Nicolaides, K.H.: Turning the pyramid of prenatal care. Fetal Diagn. Ther. **29**(3), 183–196 (2011)

Références

1. Pearl, J., *Causality*. 2009: Cambridge university press.
2. Imbens, G.W. and D.B. Rubin, *Rubin causal model.* Microeconometrics, 2010: p. 229-241.
3. Faghihi, U., et al. *From Association to Reasoning, an Alternative to Pearls' Causal Reasoning*. in *The Thirty-Third International Flairs Conference*. 2020.
4. Pearl, J. and D. Mackenzie, *The book of why: the new science of cause and effect*. 2018: Basic Books.
5. Kingma, D.P. and M. Welling, *An introduction to variational autoencoders.* Foundations and Trends® in Machine Learning, 2019. **12**(4): p. 307-392.
6. Louizos, C., et al., *Causal effect inference with deep latent-variable models.* arXiv preprint arXiv:1705.08821, 2017.
7. Miller, D.B. and J.P. O'Callaghan, *Personalized medicine in major depressive disorder—opportunities and pitfalls.* Metabolism, 2013. **62**: p. S34-S39.
8. Fink, M., *Electroconvulsive therapy resurrected: its successes and promises after 75 years*. 2011, SAGE Publications Sage CA: Los Angeles, CA.
9. Koutsouleris, N., et al., *Predicting response to repetitive transcranial magnetic stimulation in patients with schizophrenia using structural magnetic resonance imaging: a multisite machine learning analysis.* Schizophrenia Bulletin, 2018. **44**(5): p. 1021-1034.
10. Min, B., et al., *Prediction of individual responses to electroconvulsive therapy in patients with schizophrenia: Machine learning analysis of resting-state electroencephalography.* Schizophrenia research, 2020. **216**: p. 147-153.
11. Oh, H.S., et al., *Machine Learning Algorithm-Based Prediction Model for the Augmented Use of Clozapine with Electroconvulsive Therapy in Patients with Schizophrenia.* Journal of Personalized Medicine, 2022. **12**(6): p. 969.
12. Xi, Y.-B., et al., *Neuroanatomical features that predict response to electroconvulsive therapy combined with antipsychotics in schizophrenia: a magnetic resonance imaging study using radiomics strategy.* Frontiers in psychiatry, 2020. **11**: p. 456.
13. Ying, Y.-b., et al., *Electroconvulsive therapy is associated with lower readmission rates in patients with schizophrenia.* Brain Stimulation, 2021. **14**(4): p. 913-921.
14. Zhao, Q. and S. Yuan, *Magnetic Resonance Imaging (MRI) Based on Machine Learning Algorithms for the Diagnosis in Efficacy of Dexmedetomidine along with Modified Electroconvulsive Therapy Nursing on First Episode Schizophrenia.* Scientific Programming, 2021. **2021**.
15. Faghihi, U., C. Kalantarpour, and A. Saki. *Causal Probabilistic Based Variational Autoencoders Capable of Handling Noisy Inputs Using Fuzzy Logic Rules*. in *Science and Information Conference*. 2022. Springer.
16. Sainath, T.N., et al. *Convolutional, long short-term memory, fully connected deep neural networks*. in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2015. Ieee.

17. Sharma, A. and E. Kiciman, *DoWhy: An end-to-end library for causal inference.* arXiv preprint arXiv:2011.04216, 2020.