PTSD-MDNN : Fusion tardive de réseaux de neurones profonds multimodaux pour la détection du trouble de stress post-traumatique

Long NGUYEN-PHUOC[1,2] Renald GABORIAU[2] Dimitri DELACROIX[2] Laurent NAVARRO[1]

[1]Mines Saint-Étienne, University of Lyon, University Jean Monnet, Inserm, U 1059 Sainbiose, Centre CIS, 42023 Saint-Étienne, France

[2]MJ Lab, MJ INNOV, 42000 Saint-Etienne, France

Résumé – Afin de proposer un moyen plus objectif et plus rapide de diagnostiquer le trouble de stress post-traumatique (TSPT), nous présentons `PTSD-MDNN` qui fusionne deux réseaux de neurones convolutifs unimodaux et qui donne un faible taux d'erreurs de détection. En ne prenant que des vidéos et des audios comme entrées, le modèle pourrait être utilisé dans la configuration de séances de téléconsultation, dans l'optimisation des parcours patients ou encore pour l'interaction humain-robot.

Abstract – In order to provide a more objective and quicker way to diagnose post-traumatic stress disorder (PTSD), we present `PTSD-MDNN` which merges two unimodal convolutional neural networks and which gives low detection error rate. By taking only videos and audios as inputs, the model could be used in the configuration of teleconsultation sessions, in the optimization of patient journeys or for human-robot interaction.

1 Introduction

1.1 Contexte Général

Tout individu, au cours de sa vie, peut rencontrer des situations potentiellement traumatogènes. [1] définit comme traumatogène toute situation qui implique « une mort effective, une menace de mort, une blessure grave ou des violences sexuelles ». En France, le TSPT toucherait entre 1 et 2% de la population. Les symptômes du TSPT causent des problèmes importants dans les situations sociales ou professionnelles.

Traditionnellement, le TSPT a été diagnostiqué par des professionnels de la santé impliquant des questionnaires. La collecte d'informations par le biais d'un questionnaire auto-déclaratif a des limites car souvent biaisés [2] : (1) les distorsions de la mémoire et de la perception de soi des patients rendent également le diagnostic difficile, et (2) les patients sont souvent gênés d'être diagnostiqués et ne veulent pas visiter les cliniques pour le diagnostic. Finalement, peu de mesures objectives ou qualitatives sont disponibles pour aider les cliniciens à diagnostiquer ce trouble. Un exemple d'un tel entretien est le Clinician-Administered PTSD Scale (CAPS) ou encore le PCL-5 [1].

La récente pandémie de SRAS-CoV-2 peut être considérée comme un événement traumatique mondial qui a fait émerger : (1) un fort impact sur la santé mental, (2) la réalisation de nombreux soins médicaux transformée en distanciel. Par conséquent, considérant le biais des modes de collecte de données auto-déclaratifs et le changement radical induit par les consultations en distanciel, il est nécessaire de trouver un moyen plus objectif et rapide pour diagnostiquer le TSPT.

1.2 Contexte Théorique

Nous distinguons deux catégories de modèles d'intelligence artificielle : les modèles de pronostic et les modèles de diagnostic du TSPT. Notre papier se concentre sur le diagnostic, c'est-à-dire la détection de l'état actuel des patients. Une grande majorité des études sur le diagnostic utilisent des techniques d'apprentissage automatique supervisé sur des données structurées. Par exemple, [13, 4] ont tous appliqué des algorithmes de Machine à Vecteurs de Support (SVM) sur des questionnaires. Toujours sur ces données auto-déclaratives et tabulaires, [11] ont utilisé l'optimisation minimale séquentielle, le perceptron multicouches et la classification naïve bayésienne. Concernant les données biométriques, il semble que les modifications de la variabilité de la fréquence cardiaque sont significativement associées au TSPT [6]. Enfin, la conductance cutanée peut être utilisée comme outil de diagnostic [7] car elle semble en particulier corrélée à son intensité.

À l'image des données non structurées qui alimentent la prochaine génération des modèles d'IA, de nombreuses études de détection du TSPT ont profité de ces avancées pour puiser l'information dans différentes sources disponibles dans le cadre clinique ou quotidien des patients. Parmi les moyens d'acquisition et de restitution d'imagerie médicale, les études d'imagerie par résonance magnétique (IRM) structurelle [15] et fonctionnelle [13] ont permis des progrès considérables dans la compréhension des mécanismes neuronaux sous-jacents au TSPT. Ainsi couplées avec l'agorithme SVM, elles permettent de détecter le TSPT avec de bon résultats.

D'autres études ont cherché des alternatives aux données médicales traditionnelles. Le diagnostic des patients atteints de TSPT par l'analyse des signaux vocaux a été étudié depuis ces dernières années [18]. Des approches de text mining ou de traitement du langage naturel ont été également utilisées [2, 17]. Même si la télémédecine via vidéo conférence peut réduire le délai de la prise en charge et être aussi efficace que le traitement en personne [10], seul [16] se concentre sur ce type de données audiovisuelles en utilisant différentes architectures de réseaux de neurones par types de modalités de données. Actuellement, la question de la fusion de ces différentes modalités audio-vidéo pour la détection du TSPT reste entière.

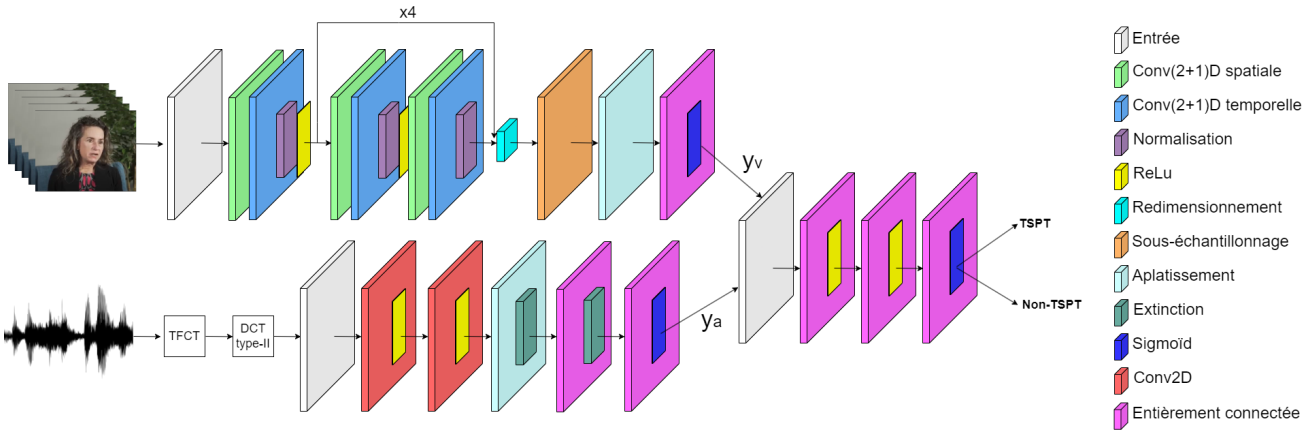

FIGURE 1 : L'architecture de PTSD-MDNN

1.3 Motivation

Malgré la variété des études citées ci-dessus, le diagnostic du TSPT nécessite des capteurs ou dispositifs médicaux à caractère invasif dont la disponibilité n'est pas garantie, notamment en temps de crise. Nous proposons, dans ce papier, un réseau de neurones profond multimodal pour la détection automatique de TSPT (`PTSD-MDNN: Post Traumatic Stress Disorder - Multimodal Deep Neural Network`) en utilisant comme entrées de simples vidéos et audios des patients en situation réelle. Nous montrons qu'en plus d'être adapté à ces données audiovisuelles facilement collectées, notre modèle obtient de meilleurs résultats en fusionnant ces deux modalités.

2 Méthode Proposée

Nous présentons un aperçu de `PTSD-MDNN` dans la Fig 1. Le modèle général se compose de deux sous-modèles qui prennent chacun en entrée une modalité différente. Le vecteur sortant de la dernière couche du classement vidéo y_v est concatené avec le vecteur sortant de la dernière couche du classement audio y_a pour former une matrice $M_{v,a}$. Après cette fusion tardive de modalités, la matrice $M_{v,a}$ est injectée dans un dernier réseau de neurones à deux couches afin de détecter le TSPT.

2.1 Classement Vidéo

Le sous-modèle de classification vidéo utilise un réseau de neurones convolutif (2+1)D [19] avec des connexions résiduelles à 18 couches de profondeur (ResNet18). La convolution (2+1)D permet la décomposition des dimensions spatiale et temporelle, créant ainsi deux étapes distinctes. Un avantage de cette approche est que la factorisation des convolutions en dimensions spatiales et temporelles permet de réduire le nombre de paramètres par rapport à la convolution 3D complète. La convolution spatiale prend les données sous la forme $(1, largeur, hauteur)$, tandis que la convolution temporelle prend les données sous la forme $(temps, 1, 1)$ comme illustré dans la Fig 2. Le redimensionnement de la vidéo est nécessaire pour : (1) effectuer un sous-échantillonnage des données, (2) examiner des parties spécifiques des images, (3) réduire la dimensionnalité pour un traitement plus rapide.

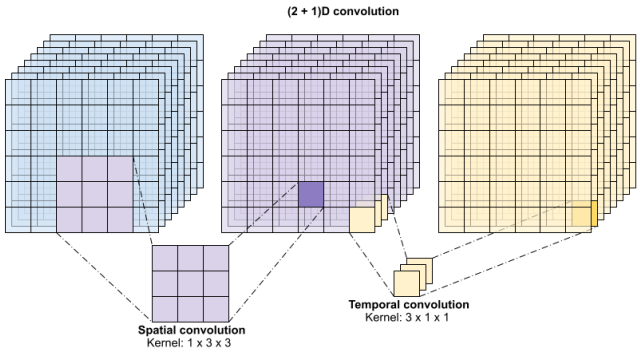

FIGURE 2 : Les convolutions spatiales et temporelles factorisées d'une convolution (2+1)D avec une taille de noyau $(3 \times 3 \times 3)$ nécessitent des matrices de poids de taille $(9 \times canaux^2) + (3 \times canaux^2)$. Ceci est moins de la moitié de celles nécessaires pour la convolution 3D complète $(27 \times canaux^2)$

2.2 Classement Audio

Nous avons adapté le modèle [9] basé sur une transformation de Fourier à court terme (TFCT) avec des fenêtres plus longues (64 ms) chevauchées à 75%, ce qui donne une meilleure résolution en fréquence pour la voix humaine. Nous avons ensuite converti les fréquences du spectrogramme en échelle logarithmique de Mel qui se rapproche de la perception humaine du son. Finalement, la transformée en cosinus discrète de type II (DCT) donne les coefficients Mel-Frequency Cepstral (MFCC) par 80 filtres triangulaires créés pour couvrir la plage de fréquences Mel. Nous sélectionnons uniquement 13 premiers qui sont utiles à la reconnaissance de la parole [3].

Le classement audio est entraîné à partir de ces MFCC en utilisant un réseau de neurones convolutif illustré dans la Fig 1. Le modèle commence par deux couches de convolution 2D avec 16 filtres chacune, une taille de noyau de (3,3) et une fonction d'activation ReLu. Ces couches convolutives 2D sont utilisées pour extraire des caractéristiques importantes des MFCC d'entrée, qui sont des matrices de taille (778, 13, 1). Ensuite, les couches d'Aplatissement (Flatten) et d'Extinction (Dropout) s'enchainent pour respectivement convertir la sortie de la dernière couche convolutive en un vecteur à une dimen-

sion et pour désactiver aléatoirement certains neurones afin de contourner le surapprentissage. Deux couches entièrement connectées sont ensuite combinées via une fonction Sigmoïd, ce qui permet d'obtenir les probabilités des classes à prédire.

2.3 Fusion Tardive Des Modalités

Nous avons choisi pour notre modèle une fusion tardive, dite « fusion orientée décisions ». En effet, [20] a montré que : (1) un meilleur réseau unimodal peut surpasser un réseau multimodal contre le problème de surapprentissage ; (2) différentes modalités se surajustent et se généralisent à des rythmes différents, donc les entraîner conjointement avec une seule stratégie d'optimisation n'est pas optimal. Nous avons imaginé un mécanisme de « correction d'erreur » qui fusionne des prédictions provenant de deux réseaux unimodaux qui sont entrainés séparément. La fonction de perte que nous avons utilisé pour ce réseau de fusion tardive est la même que pour les sous-modèles : Perte d'entropie croisée binaire définie comme la formule 1. Cette formule suppose que les p_i sont des probabilités et les y_i sont des labels (0 ;1).

$$L = -\frac{1}{N}\sum_{i=1}^{2} y_i log(p_i) = -\frac{1}{N}[y_1 log(p_1) + y_2 log(p_2)] \quad (1)$$

3 Évaluation

3.1 Base De Données

En général, il est difficile de collecter des données de haute qualité auprès de personnes qui présentent des symptômes de TSPT. Il peut y avoir aussi des considérations éthiques qui limitent la collecte et l'utilisation de données en milieu naturel. Cela peut être particulièrement difficile dans le contexte d'un trouble sensible comme le TSPT, où les participants peuvent être réticents à divulguer des informations personnelles.

Par conséquent, seules quatre bases de données non structurées pour la détection du TSPT existent : eDAIC-WOZ [5], FEMH [8], Aurora [14] et PTSD in-the-wild [16]. Parmi elles, seules les bases eDAIC-WOZ et PTSD in-the-wild disposent à la fois de modalités audio et vidéo. Nous avons choisi d'appliquer notre modèle `PTSD-MDNN` sur la base PTSD in-the-wild pour son caractère réel en milieu naturel. La base de données PTSD in-the-wild (EULA) contient 634 vidéos équilibrées : 317 vidéos de sujets avec TSPT et 317 vidéos de sujets témoins sains avec aucun symptôme de TSPT.

3.2 Résultats

Nous nous intéressons à l'évaluation d'une classification binaire avec deux classes : TSPT (positif) et Non-TSPT (négatif) avec différentes métriques de classification populaires : l'accuracy, la précision, le rappel. Nous avons suivi le processus train/validation/test proposé par [16] (80%/10%/10%) pour entraîner (`PTSD-MDNN`) sur une carte GPU NVIDIA A100 SXM 40Go avec taille de batch de 8, un taux d'apprentissage de 0.001, et un optimiseur Adam pour 50 époques. Ces paramètres sont optimaux pour éviter le surapprentissage lié à la taille des données. De plus, nous avons utilisé différentes méthodes de régularisation pour le classement audio (Table 1).

TABLE 1 : Résultats du classement sur les données test

Modalité	Régularisation	Accuracy	Précision	Rappel
Vidéo	NA	0,89	0,84	0,84
Audio	NA	0,72	0,68	0,81
Audio	L1	0,73	0,67	0,90
Audio	L2	0,75	0,72	0,81
Vidéo + Audio	L1	0,89	0,90	0,87
Vidéo + Audio	L2	0,92	0,88	0,97

3.3 Discussion

Le meilleur des modèles unimodaux est le classement vidéo avec une accuracy de 0,89. En revanche, le classement audio de base ne donne pas de très bons résultats même si les régularisations L1 et L2 l'améliorent respectivement à 0,73 et 0,75. Notre approche de fusion tardive des modalités apporte de réelles améliorations par rapport aux classements unimodaux car `PTSD-MDNN` donne la meilleure accuracy (0,92), avec une régularisation L2, et le meilleur rappel (0,97). Le principal avantage d'une fusion tardive est la prise en charge des différentes modalités non alignées ainsi nous ne dépendons pas de l'interopérabilité des capteurs. De plus, l'entrainement indépendant des deux sous-modèles permet de gagner du temps en effectuant des tâches en parallèle. Enfin, cette fusion permet d'apporter la flexibilité pour le choix des sous-modèles adaptés à chaque modalité. Il est à noter que la différence de taille des fichiers de la base PTSD in-the-wild (le plus court : 0 min 35 s et le plus long : 44 min 40 s) peut créer des difficultés pour l'extraction des variables en entrée du modèle de convolution.

4 Conclusion

Nous proposons `PTSD-MDNN`, un modèle qui fusionne tardivement des modalités audio et vidéo pour détecter le TSPT. Grâce à un mécanisme de correction d'erreur, notre modèle surpasse les modèles unimodaux. En plus d'être non invasif, `PTSD-MDNN` traite les informations sensibles sur les patients à très bas niveau (pixel, MFCC), ce qui permet de garder une certaine confidentialité pour les patients par rapport aux approches type NLP où les paroles sont transcrites.

Ce travail ouvre une multitude de travaux futurs. Premièrement, nous avons l'intention d'extraire des variables de haut niveau à partir d'aspects comportementaux subtils, tels que les mouvements du corps, les expressions faciales pour la vision ainsi que la prosodie et la parole pour l'audio. Deuxièment, d'autres directions concernent la fusion des modalités à travers le mécanisme de l'attention inter-modalité.

Références

[1] AMERICAN PSYCHIATRIC ASSOCIATION, Marc-Antoine CROCQ, Julien-Daniel GUELFI, Patrice BOYER, Marie-Claire PULL et Charles-Bernard PULL : *DSM-5 - Manuel diagnostique et statistique des troubles mentaux*. Elsevier Masson, juin 2015.

[2] Debrup BANERJEE, Kazi ISLAM, Keyi XUE, Gang MEI, Lemin XIAO, Guangfan ZHANG, Roger XU, Cai LEI,

Shuiwang JI et Jiang LI : A deep transfer learning approach for improved post-traumatic stress disorder diagnosis. *Knowl Inf Syst*, 60(3):1693–1724, septembre 2019.

[3] Jeroen BREEBAART et Martin F. MCKINNEY : Features for Audio Classification. *In* Wim F. J. VERHAEGH, Emile AARTS et Jan KORST, éditeurs : *Algorithms in Ambient Intelligence*, Philips Research, pages 113–129. Springer Netherlands, Dordrecht, 2004.

[4] Michael S. BREEN, Kevin G.F. THOMAS, David S. BALDWIN et Gosia LIPINSKA : Modelling PTSD diagnosis using sleep, memory, and adrenergic metabolites : An exploratory machine-learning study. *Human Psychopharmacology : Clinical and Experimental*, 34(2):e2691, 2019. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/hup.2691.

[5] Jonathan GRATCH, Ron ARTSTEIN, Gale LUCAS, Giota STRATOU, Stefan SCHERER, Angela NAZARIAN, Rachel WOOD, Jill BOBERG, David DEVAULT, Stacy MARSELLA, David TRAUM, Skip RIZZO et Louis-Philippe MORENCY : The Distress Analysis Interview Corpus of human and computer interviews. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, mai 2014. European Language Resources Association (ELRA).

[6] Marit HAUSCHILDT, Maarten J. V. PETERS, Steffen MORITZ et Lena JELINEK : Heart rate variability in response to affective scenes in posttraumatic stress disorder. *Biological Psychology*, 88(2):215–222, décembre 2011.

[7] Rebecca HINRICHS, Vasiliki MICHOPOULOS, Sterling WINTERS, Alex O. ROTHBAUM, Barbara O. ROTHBAUM, Kerry J. RESSLER et Tanja JOVANOVIC : Mobile assessment of heightened skin conductance in posttraumatic stress disorder. *Depression and Anxiety*, 34(6):502–507, 2017. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/da.22610.

[8] Kazi Aminul ISLAM, Daniel PEREZ et Jiang LI : A Transfer Learning Approach for the 2018 FEMH Voice Data Challenge. *In 2018 IEEE International Conference on Big Data (Big Data)*, pages 5252–5257, décembre 2018.

[9] Stefan KAHL, Amanda NAVINE, Tom DENTON, Holger KLINCK, Patrick HART, Hervé GLOTIN, Hervé GOËAU, Willem-Pier VELLINGA, Robert PLANQUÉ et Alexis JOLY : Overview of BirdCLEF 2022. *In Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, Bologna, Italy, septembre 2022.

[10] Jan A. LINDSAY, Michael R. KAUTH, Sonora HUDSON, Lindsey A. MARTIN, David J. RAMSEY, Lawrence DAILY et John RADER : Implementation of Video Telehealth to Improve Access to Evidence-Based Psychotherapy for Posttraumatic Stress Disorder. *Telemedicine and e-Health*, 21(6):467–472, juin 2015. Publisher : Mary Ann Liebert, Inc., publishers.

[11] Sevinç İlhan OMURCA et Ekin EKINCI : An alternative evaluation of post traumatic stress disorder with machine learning methods. *In 2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, pages 1–7, septembre 2015.

[12] Dhanesh RAMACHANDRAM et Graham W. TAYLOR : Deep Multimodal Learning : A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine*, 34(6):96–108, novembre 2017. Conference Name : IEEE Signal Processing Magazine.

[13] D. RANGAPRAKASH, Gopikrishna DESHPANDE, Thomas A. DANIEL, Adam M. GOODMAN, Jennifer L. ROBINSON, Nouha SALIBI, Jeffrey S. KATZ, Thomas S. DENNEY JR. et Michael N. DRETSCH : Compromised hippocampus-striatum pathway as a potential imaging biomarker of mild-traumatic brain injury and posttraumatic stress disorder. *Human Brain Mapping*, 38(6):2843–2864, 2017. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.23551.

[14] Niels RATHLEV : Correction : The AURORA Study : a longitudinal, multimodal library of brain biology and function after traumatic stress exposure. *All Scholarly Works*, septembre 2020.

[15] Lauren E. SALMINEN, Rajendra A. MOREY, Brandalyn C. RIEDEL, Neda JAHANSHAD, Emily L. DENNIS et Paul M. THOMPSON : Adaptive Identification of Cortical and Subcortical Imaging Markers of Early Life Stress and Posttraumatic Stress Disorder. *Journal of Neuroimaging*, 29(3):335–343, 2019. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/jon.12600.

[16] Moctar Abdoul Latif SAWADOGO, Furkan PALA, Gurkirat SINGH, Imen SELMI, Pauline PUTEAUX et Alice OTHMANI : PTSD in the Wild : A Video Database for Studying Post-Traumatic Stress Disorder Recognition in Unconstrained Environments, septembre 2022. arXiv :2209.14085 [cs].

[17] Jeff SAWALHA, Muhammad YOUSEFNEZHAD, Zehra SHAH, Matthew R. G. BROWN, Andrew J. GREENSHAW et Russell GREINER : Detecting Presence of PTSD Using Sentiment Analysis From Text Data. *Frontiers in Psychiatry*, 12, 2022.

[18] Chappidi SUNEETHA et Raju ANITHA : A Survey Of Machine Learning Techniques OnSpeech Based Emotion Recognition And Post Traumatic Stress DisorderDetection. *nq*, 20(14):1–11, décembre 2022.

[19] Du TRAN, Heng WANG, Lorenzo TORRESANI, Jamie RAY, Yann LECUN et Manohar PALURI : A Closer Look at Spatiotemporal Convolutions for Action Recognition, avril 2018. arXiv :1711.11248 [cs].

[20] Weiyao WANG, Du TRAN et Matt FEISZLI : What Makes Training Multi-Modal Classification Networks Hard ? *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020.