Transformer multimodal pour la détection du stress

Kevin Feghoul [1,2], *Deise Santana Maia* [2], *Mohamed Daoudi* [2,3], *Ali Amad* [1]
[1]Univ. Lille, Inserm, CHU Lille, UMR-S1172 - LilNCog, F-59000 Lille, France
[2]Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[3]IMT Nord Europe, Institut Mines-Télécom, Centre for Digital Systems, F-59000 Lille, France
{kevin.feghoul, deise.santanamaia, mohamed.daoudi, ali.amad}@univ-lille.fr

Résumé

Le stress peut entraîner des conséquences nocives sur la santé mentale et physique des individus, ainsi que sur leur qualité de vie en général. Par conséquent, il est bénéfique de développer des outils automatisés pour aider à y faire face. Dans cette perspective, nous avons proposé différentes architectures Transformer multimodales sur l'ensemble de données WESAD afin de détecter le stress de manière automatique. Les résultats de notre étude démontrent l'adaptabilité des modèles proposés à cette tâche. En utilisant la méthode de fusion intermédiaire, nous avons dépassé l'état de l'art, avec une précision de $98,69\%$ *et un score F1 de* $98,73\%$*. Les résultats obtenus mettent en évidence l'efficacité de notre méthode et ouvrent des perspectives intéressantes pour le développement de techniques de reconnaissance des émotions basées sur des architectures Transformer multimodales.*

Mots clefs

Multimodal, Transformer, Stress, Données physiologiques

1 Introduction

Selon la *National Institutes of Health*, le stress se définit comme une réponse du corps à une pression physique, mentale ou émotionnelle. Les facteurs de stress peuvent inclure des pressions liées au travail, des difficultés financières, des problèmes relationnels, et bien plus encore [1]. Le stress peut être classé en deux catégories en fonction de sa durée : (1) le stress aigu, qui est de courte durée et qui est provoqué par un événement inhabituel ou une menace immédiate ; et (2) le stress chronique, qui est une réponse prolongée de l'organisme face à des facteurs de stress maintenus sur une longue période. Le stress chronique peut augmenter la susceptibilité à certains types de cancer [2], ralentir la guérison des plaies [3], et accroître la vulnérabilité aux infections [4].

Étant donné l'impact généralisé du stress sur les individus et la société dans son ensemble, il est de plus en plus important de développer des outils de détection automatique du stress. Ces outils permettront aux professionnels de santé de prendre des mesures préventives et d'offrir des traitements adaptés aux patients présentant des signes de stress. Pour développer de tels outils, il est important de comprendre que le stress est une réaction physiologique déclenchée par le système nerveux sympathique (SNS) en réponse à un stimulus, qui déclenche une réaction hormonale en cascade. Cette réaction hormonale implique la libération d'hormones telles que l'ACTH, le cortisol et l'adrénaline. À la suite de cette libération d'hormones, une accélération de la fréquence cardiaque et respiratoire, ainsi qu'une tension musculaire peuvent être observées.

Grâce aux récentes avancées dans les techniques d'apprentissage automatique, l'apprentissage profond est désormais largement utilisé pour le traitement des séries temporelles [5–7]. Les méthodes d'apprentissage profond offrent plusieurs avantages, notamment : (1) la capacité à capturer des caractéristiques complexes dans les séries temporelles, ce qui peut être difficile à détecter avec les méthodes traditionnelles ; (2) la possibilité de modéliser une large gamme de séries temporelles, telles que des données continues et discrètes, des séries multivariées et des séries à fréquence variable ; et (3) la capacité de faire des prévisions à long terme.

Comme le stress peut être détecté à l'aide de plusieurs types de capteurs, chacun possédant des propriétés différentes en termes de fréquence et de types de signaux enregistrés, nous avons traité la détection du stress comme un problème multimodal en combinant les signaux provenant de différents capteurs. Dans cette étude, nous nous sommes intéressés au modèle d'apprentissage profond Transformer [8] pour la tâche de détection automatique du stress, en adoptant une approche multimodale. Nous avons effectué nos expérimentations sur le jeu de données WESAD [9].

Les contributions de ce travail sont doubles et peuvent être résumées comme suit : (1) évaluation comparative des différentes méthodes de fusion de signaux physiologiques multimodaux en utilisant le modèle Transformer ; (2) amélioration significative des résultats par rapport à l'état de l'art.

2 Etat de l'art

Au cours des dernières années, de nombreuses études se sont intéressées à l'utilisation des données physiologiques pour la détection du stress. Le jeu de données WESAD [9] est couramment utilisé dans les travaux de recherche relatifs à cette problématique. Les auteurs de [9] ont proposé

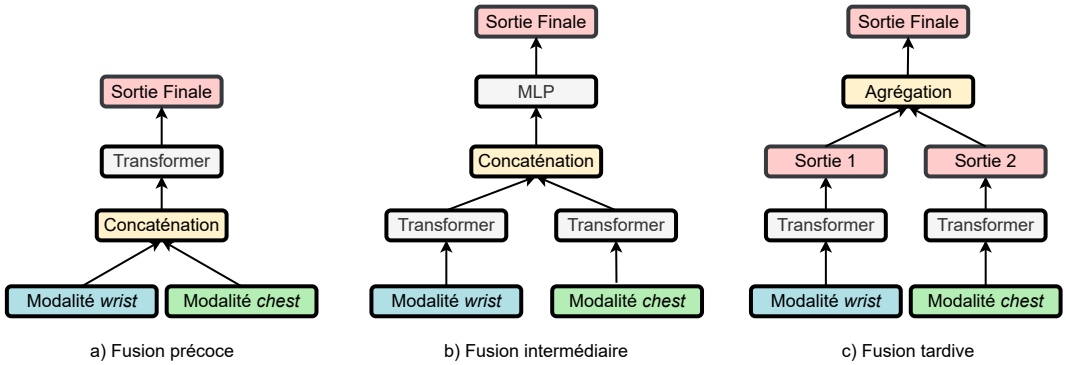

FIGURE 1 – *Illustration des trois principales stratégies de fusion multimodale utilisant le Transformer, à savoir la fusion précoce (a), la fusion intermédiaire (b) et la fusion tardive (c).*

un benchmark complet utilisant des statistiques provenant des domaines temporel et fréquentiel pour entraîner différents modèles d'apprentissage automatique traditionnels.

Plusieurs autres approches basées sur l'apprentissage profond ont été proposées en utilisant le jeu de données WESAD. Samyoun et al. [10] ont proposé d'utiliser des réseaux de neurones de type GAN, RNN et MLP pour traduire les signaux physiologiques du capteur du poignet en signaux provenant des capteurs placés au niveau de la poitrine. Les données traduites ont ensuite été utilisées pour détecter le stress à l'aide de méthodes d'apprentissage automatique. Gil-Martin et al. [11] ont proposé l'utilisation d'un réseau de neurones de type CNN pour la détection du stress, ainsi que l'analyse de plusieurs techniques de traitement du signal pour générer les entrées du modèle. Dans leur étude, Huynh et al. [12] ont proposé l'utilisation d'un schéma d'entraînement de réseau de neurones profonds optimisé basé sur des CNN, en utilisant la méthode de recherche d'architecture neuronale. L'étude menée par Lai et al. [13] a utilisé un réseau de neurones à convolution temporelle résiduelle pour traiter les différents signaux, et a proposé plusieurs stratégies de fusion. Wu et al. [14] ont étudié l'utilisation de matrices SPD pour fusionner efficacement des signaux physiologiques et comportementaux, permettant ainsi de capturer simultanément les informations de corrélation au sein et entre les différentes modalités. Les résultats de leur étude ont démontré l'impact positif de l'utilisation de plusieurs modalités sur les performances par rapport à l'utilisation d'une seule modalité.

Pour notre part, nous proposons l'utilisation d'un Transformer multimodal pour la tâche de détection du stress, où nous avons traité chaque ensemble de signaux provenant d'un capteur comme une modalité distincte.

3 Approche proposée

La présente section expose les différentes techniques de fusion multimodale que nous avons employées pour la détection du stress, en utilisant le Transformer. Ces techniques incluent la fusion précoce, la fusion intermédiaire et la fusion tardive, lesquelles sont illustrées dans la Figure 1.

3.1 Transformer

Le modèle Transformer est une architecture de réseau de neurones introduite par [8] qui est maintenant considérée comme la référence en matière de traitement de tâches liées au langage naturel. De plus, le Transformer a été étendu avec succès à d'autres domaines tels que le traitement des séries temporelles [15] et l'apprentissage multimodal [16]. Le Transformer repose sur le mécanisme d'auto-attention, qui permet au modèle de se concentrer sur différentes parties de la séquence d'entrée pour effectuer des prédictions. Ce mécanisme calcule une somme pondérée de la séquence d'entrée, où les poids sont appris pendant l'entraînement. Grâce à cela, le modèle est capable de capturer des dépendances à long terme et de faire des prédictions en se basant sur l'ensemble de la séquence d'entrée plutôt que sur des représentations passées limitées, contrairement aux réseaux de neurones récurrents.

En plus du mécanisme d'auto-attention, le Transformer utilise également des techniques telles que l'attention multi-tête et l'encodage de position pour améliorer sa performance en terme de prédiction. L'attention multi-tête permet au modèle de calculer plusieurs poids d'attention en parallèle pour différents sous-espaces de caractéristiques de l'entrée, ce qui permet au modèle de mieux capturer les relations entre les différentes parties de la séquence. L'encodage de position est une technique qui permet au modèle de prendre en compte l'ordre des éléments de la séquence, en ajoutant des informations sur leur position relative. Cela permet au modèle de mieux comprendre les relations entre les différentes parties de la séquence et de capturer les informations séquentielles.

3.2 Transformer multimodal

Nous avons adopté une approche multimodale basée sur le modèle Transformer pour détecter le stress. Cette méthode nous a permis d'intégrer de manière efficace différents signaux physiologiques provenant de divers capteurs.

L'un des principaux défis de l'apprentissage automatique multimodal est la fusion efficace de données provenant de

TABLEAU 1 – *Détection du stress : comparaison avec des méthodes de pointe.*

Methodes	Wrist		Chest	
	Acc	F1 score	Acc	F1 score
Schmidt et al. [9]	87.12	84.11	92.83	91.07
Samyoun et al. [10]	89.90	87.60	91.10	90.20
Gil-Martin et al. [11]	92.70	92.55	93.10	93.01
Huynh et al. [12]	93.14	-	-	-
Wu et al. [14]	94.65	93.99	95.54	94.76
Lai et al. [13]	94.16	93.62	**96.69**	**96.61**
Transformer	**95.74 ± 7.22**	**96.4 ± 5.34**	96.07± 4.81	95.93 ± 4.96

TABLEAU 2 – *Détection du stress (approches multimodales) : comparaison avec des méthodes de pointe.*

Methodes	Wrist + Chest	
	Acc	F1 score
Schmidt et al. [9]	92.28	90.74
Samyoun et al. [10]	94.70	93.40
Gil-Martin et al. [11]	96.62	96.63
Wu et al. [14]	96.88	96.44
Lai et al. [13]	97.75	97.74
MMT-early (ours)	98.13 ± 3.00	98.07 ± 3.09
MMT-inter (ours)	**98.69 ± 2.85**	**98.73 ± 2.62**
MMT-late (ours)	98.34 ± 3.31	98.33 ± 3.28

sources différentes. Les stratégies de fusion multimodale sont généralement classées en trois catégories : la fusion précoce, la fusion intermédiaire et la fusion tardive, chacune ayant ses avantages et ses inconvénients en fonction de la tâche à accomplir et des caractéristiques des données. En ce qui concerne la fusion précoce, les deux modalités d'entrée sont d'abord concaténées avant d'être traitées par un Transformer. Pour la fusion intermédiaire, en revanche, les deux modalités sont traitées de manière indépendante par un Transformer, permettant ainsi de découvrir les corrélations intra-modales avant de les fusionner pour découvrir les corrélations inter-modales. Les caractéristiques extraites sont ensuite concaténées et traitées par un réseau de neurones de type MLP pour la classification finale. Enfin, pour la fusion tardive, les deux modalités sont traitées par un Transformer jusqu'à la prédiction, suivie d'une fonction d'agrégation pour la prédiction finale. Cette approche s'avère utile lorsque les modalités d'entrée sont très différentes et ne peuvent pas être facilement combinées en une représentation conjointe. Nous désignerons respectivement ces méthodes sous les noms de MMT-early, MMT-inter et MMT-late (MMT pour *Multimodal Transformer*).

4 Résultats expérimentaux

4.1 Jeu de données

WESAD est un ensemble de données multimodal bien connu pour la détection du stress et de l'affect. Il contient des données physiologiques et de mouvement provenant de 15 sujets, qui ont été capturées à l'aide d'un bracelet Empatica E4 porté au poignet (*wrist*) et d'un dispositif RespiBAN placé au niveau de la poitrine (*chest*). Le bracelet Empatica enregistre l'activité électrodermale (EDA), le volume sanguin pulsé (BVP), la température corporelle (TEMP) et l'accélération sur trois axes (ACC) à des fréquences respectives de 4, 64, 4 et 32 Hz. En complément, le RespiBAN mesure l'électrocardiogramme (ECG), l'électromyographie (EMG), la respiration (RESP), la température de la peau (TEMP), l'EDA et l'ACC, échantillonnés à une fréquence de 700 Hz.
Le protocole d'étude a été conçu pour induire trois états émotionnels chez les sujets : neutre, stressé et amusé. En nous appuyant sur des travaux antérieurs [9–14], nous avons formulé un problème de détection de stress binaire (stress vs non-stress) en utilisant les séquences de stimuli catégorisé comme neutres et amusants pour constituer la classe "non-stress", conformément à la littérature.

4.2 Prétraitement

Nous avons tout d'abord appliqué un filtre passe-bas aux différents signaux physiologiques, afin de réduire le bruit et de conserver les fréquences d'intérêt. Ensuite, nous les avons sous-échantillonnés à une fréquence de 4 Hz. Les signaux ont ensuite été segmentés en fenêtres glissantes de 60 secondes, sans chevauchement entre les fenêtres successives. Chaque échantillon correspond à l'ensemble des signaux provenant des deux capteurs pendant une période de 60 secondes, ce qui donne un total de 240 points par échantillon.

4.3 Résultats

Conformément aux travaux antérieurs sur WESAD [9–14], nous avons utilisé la stratégie d'évaluation *Leave-One-Subject-Out Cross Validation* (LOSO-CV) pour valider nos différents modèles.
Le modèle Transformer a surpassé toutes les autres méthodes en termes de précision et de score F1 lorsqu'il a été entraîné avec les données provenant du bracelet, dépassant de près de $1,09\%$ et $2,41\%$ respectivement le modèle affichant la meilleure performance [14].
En ce qui concerne le capteur attaché au niveau de la poitrine, notre modèle se classe deuxième en termes de performance, avec une précision et un score F1 inférieurs de seulement $0,62\%$ et $0,68\%$ respectivement par rapport à [13].
Nous avons observé que les méthodes basées sur l'apprentissage profond [11–14] ont obtenu de meilleures performances que les méthodes basées sur l'apprentissage automatique [9, 10] pour les deux capteurs.
De plus, nous avons réalisé des expériences en utilisant à la fois les données provenant du bracelet et celles du capteur attaché à la poitrine. Les résultats de ces expériences sont présentés dans le tableau 2. Nous pouvons constater que les trois modèles que nous proposons, à savoir MMT-early, MMT-inter et MMT-late, affichent des performances

supérieures à toutes les autres méthodes. En particulier, le modèle MMT-inter a obtenu les meilleurs résultats pour les deux métriques d'évaluation, surpassant le modèle de [13] avec une amélioration de $0,94\%$ et $0,99\%$ en termes de précision et de score F1, respectivement. Ces résultats confirment notre hypothèse selon laquelle l'utilisation de modèles multimodaux est appropriée pour le traitement de groupes de signaux provenant de différents capteurs.

5 Conclusion

Dans cette étude, nous avons traité de la détection du stress en utilisant des architectures Transformer multimodales. À la suite de multiples expérimentations sur l'ensemble de données WESAD, nous avons établi un nouvel état de l'art, en atteignant des taux de précision et de score F1 respectifs de $98,69\%$ et $98,73\%$. Dans un futur travail, nous envisageons d'étendre nos recherches à l'utilisation de différents types de données tels que des données vidéo, audio et textuelles pour des tâches d'informatique affective.

Références

[1] Longfei Yang, Yinghao Zhao, Yicun Wang, Lei Liu, Xingyi Zhang, Bingjin Li, et Ranji Cui. The effects of psychological stress on depression. *Current neuropharmacology*, 13(4) :494–504, 2015.

[2] Alison N Saul, Tatiana M Oberyszyn, Christine Daugherty, Donna Kusewitt, Susie Jones, Scott Jewell, William B Malarkey, Amy Lehman, Stanley Lemeshow, et Firdaus S Dhabhar. Chronic stress and susceptibility to skin cancer. *Journal of the National Cancer Institute*, 97(23) :1760–1767, 2005.

[3] Ronald Glaser et Janice K Kiecolt-Glaser. Stress-induced immune dysfunction : implications for health. *Nature Reviews Immunology*, 5(3) :243–251, 2005.

[4] Sheldon Cohen, David AJ Tyrrell, et Andrew P Smith. Psychological stress and susceptibility to the common cold. *New England journal of medicine*, 325(9) :606–612, 1991.

[5] John Cristian Borges Gamboa. Deep learning for time-series analysis. *arXiv preprint arXiv :1701.01887*, 2017.

[6] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, et Pierre-Alain Muller. Deep learning for time series classification : a review. *Data mining and knowledge discovery*, 33(4) :917–963, 2019.

[7] Guansong Pang, Chunhua Shen, Longbing Cao, et Anton Van Den Hengel. Deep learning for anomaly detection : A review. *ACM computing surveys (CSUR)*, 54(2) :1–38, 2021.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, et Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[9] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, et Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. Dans *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[10] Sirat Samyoun, Abu Sayeed Mondol, et John A Stankovic. Stress detection via sensor translation. Dans *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 19–26. IEEE, 2020.

[11] Manuel Gil-Martin, Ruben San-Segundo, Ana Mateos, et Javier Ferreiros-Lopez. Human stress detection with wearable sensors using convolutional neural networks. *IEEE Aerospace and Electronic Systems Magazine*, 37(1) :60–70, 2022.

[12] Lam Huynh, Tri Nguyen, Thu Nguyen, Susanna Pirttikangas, et Pekka Siirtola. Stressnas : Affect state and stress detection using neural architecture search. Dans *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, pages 121–125, 2021.

[13] Kenneth Lai, Svetlana N Yanushkevich, et Vlad P Shmerko. Intelligent stress monitoring assistant for first responders. *IEEE Access*, 9 :25314–25329, 2021.

[14] Yujin WU, Mohamed Daoudi, Ali Amad, Laurent Sparrow, et Fabien D'Hondt. Fusion of physiological and behavioural signals on spd manifolds with application to stress and pain detection. *arXiv preprint arXiv :2207.08811*, 2022.

[15] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, et Wancai Zhang. Informer : Beyond efficient transformer for long sequence time-series forecasting. Dans *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021.

[16] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, et Boqing Gong. Vatt : Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34 :24206–24221, 2021.