20ème édition du colloque ORASIS, journées francophones des jeunes chercheurs en vision par ordinateur

Intelligence Artificielle et Emotionnelle pour la santé mentale

Dr. HDR. Alice OTHMANI

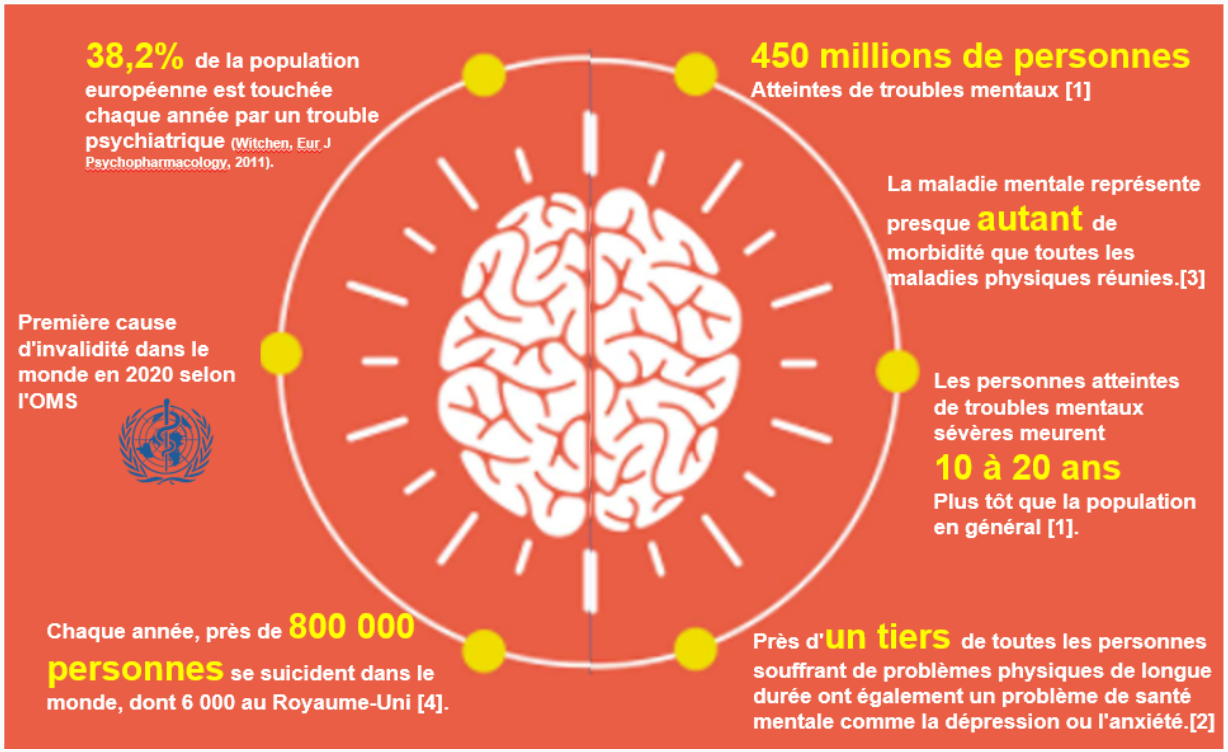
June 10, 2025

Plan de la présentation

Plan de la présentation

Contexte et Motivations

Pourquoi s'intéresser à la santé mentale et aux troubles mentaux ?



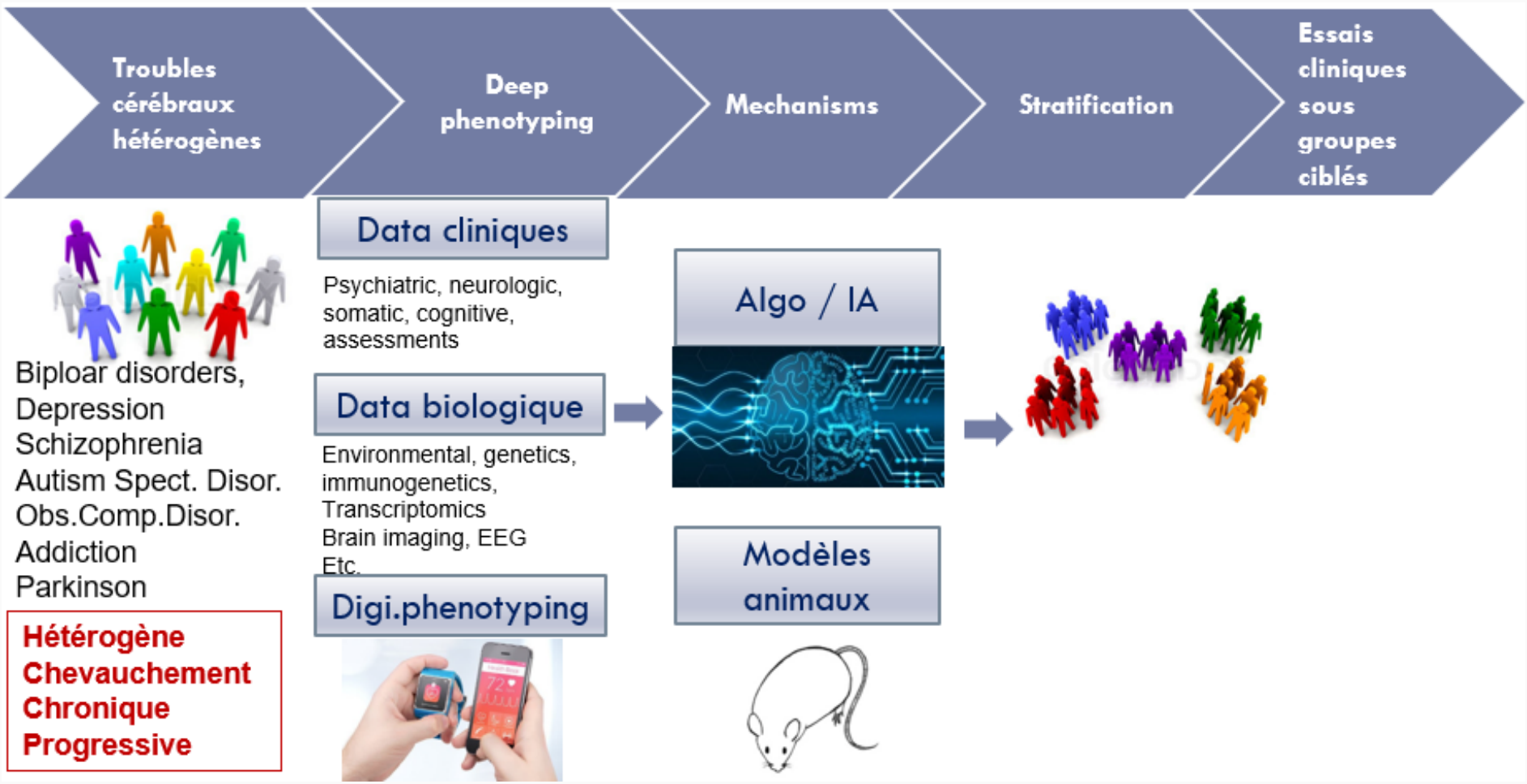
Fréquent, à début précoce, grave, **Un enjeu majeur de santé publique**

[1] Brown, et al., (2010) 'Twenty-five year mortality of a community cohort with schizophrenia', British Journal of Psychiatry 196: 116-121; Parks, et al., (2006) Morbidity and Mortality in People with Serious Mental Illness, 13th technical report, Alexandria, Virginia: National Association of state Mental Health Program Directors.

[2] Naylor, C., et al. (2012), 'Long-term conditions and mental health: The cost of co-morbidities', The King's Fund and Centre for Mental Health.

[3] Kessler et al., (2005), 'Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication', Archives of General Psychiatry 62: 593-602.

Vers une médecine de précision et personnalisée



La médecine de précision en neuro/psychiatrie transformera le diagnostic, le traitement et le pronostic des patients atteints de troubles mentaux graves.

Objectifs

- Développer des solutions basées sur l'intelligence artificielle (IA) pour le diagnostic, le pronostic, l'assistance et la découverte de médicaments (drug discovery) pour la santé mentale.
- Développer des systèmes intelligents basés sur des approches d'intelligence artificielle intégrées dans des capteurs et des dispositifs portables (Digital Phenotyping) pour étudier la maladie mentale et proposer des interventions potentielles.

Positionnement à l'échelle national et international

- L'imagerie médicale est le premier domaine d'applications médicales pour lequel l'utilisation de l'IA est des plus prometteuses.
- Peu d'équipes en vision par ordinateur s'intéressent aux applications IA pour la psychiatrie et la santé mentale.

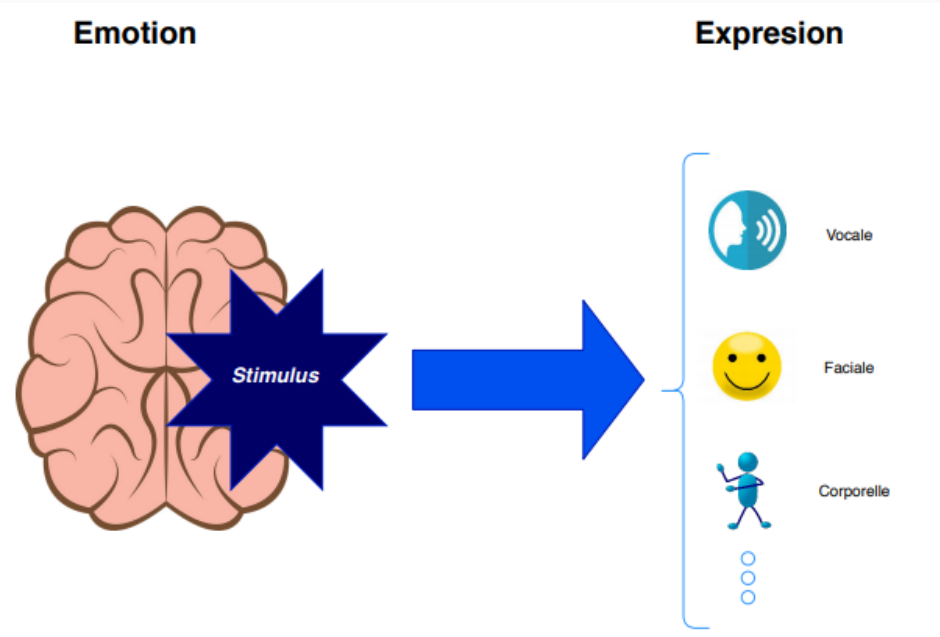
Spécificité et Originalité de mon projet

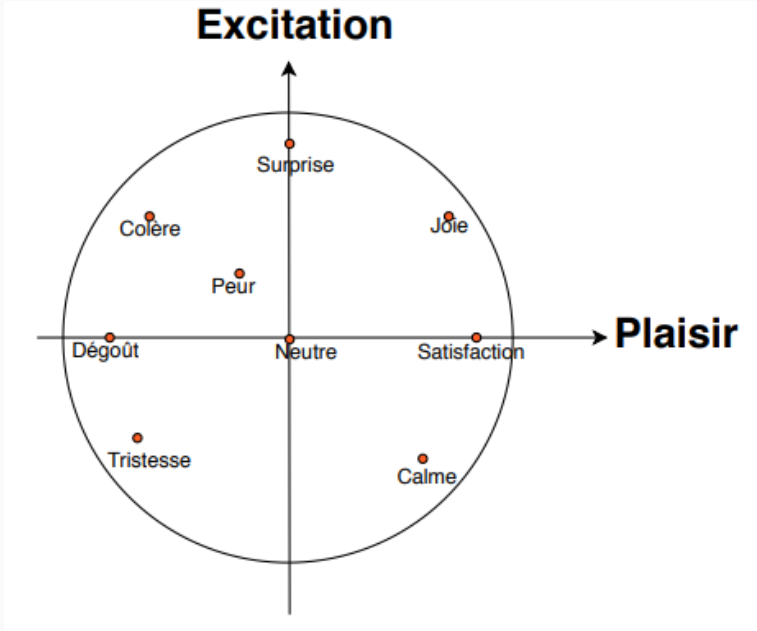
- Utilisation de l'IA pour couvrir toute la chaîne allant du diagnostic des troubles mentaux à la découverte de médicament par le criblage à haut débit (high-throughput screening, HTS).
- Le développement de biomarqueurs basés sur l'intelligence émotionnelle pour le diagnostique et le suivi des personnes souffrantes de troubles mentaux.

Plan de la présentation

L'intelligence Emotionnelle - Définition

L'intelligence émotionnelle artificielle (IE) ou **l'informatique affective**[1] est l'étude et le développement de systèmes intelligents pour la reconnaissance, l'interprétation, le traitement et/ou la simulation des affects ou émotions humaines.





Projection de classes d'émotion discrète dans l'espace excitation-plaisir.

Une intelligence artificielle pour les expressions faciales et les emotions

- Le visage humain transmet une quantité importante d'informations sur l'identité, le sexe, l'appartenance ethnique, l'âge et les émotions.
- L'expression de l'émotion sur nos visages est universelle chez les humains, *la théorie de Darwin*.
- IE a un large éventail d'applications en Interaction Homme-Machine telles que en robotique[2], en éducation[3] et en informatique médicale[4].

Une intelligence artificielle pour les expressions faciales - les limitations

- **La variabilité des manifestations d'expressions** spontanées chez les individus,
- **La variabilité des conditions d'acquisition non contrôlées** telles que l'éclairage, les occlusions et les postures de la tête.
- Le besoin de **grandes bases** d'images labelisées pour entraîner les réseaux de neurones profonds pour éviter le risque de sur-ajustement ou surapprentissage du modèle.
- La plupart des méthodes existantes manquent encore de **généralisabilité**.

Analyse des variations des régions faciales pour la reconnaissance de l'expression faciale

- Une approche basée sur un réseau de neurones convolutifs peu profonds pour REF,
- Une approche performante basée sur l'apprentissage profond, entraînée sur un petit ensemble de données,
- Utiliser des techniques d'augmentation de données pour améliorer les performances,
- Un réseau profond d'agrégation de patchs faciaux pour REF mais qui pourrait être utilisé pour la reconnaissance d'autres traits du visage comme l'âge, le sexe, l'origine ethnique ou l'identité,
- Une analyse des contributions des différentes parties du visage à l'affichage des émotions humaines.

Analyse des variations des régions faciales pour la reconnaissancede l'expression faciale

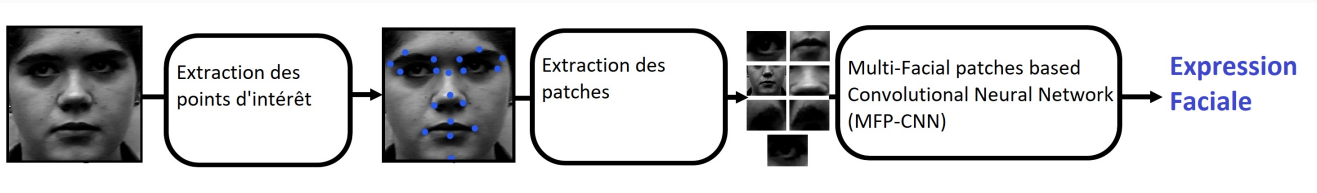
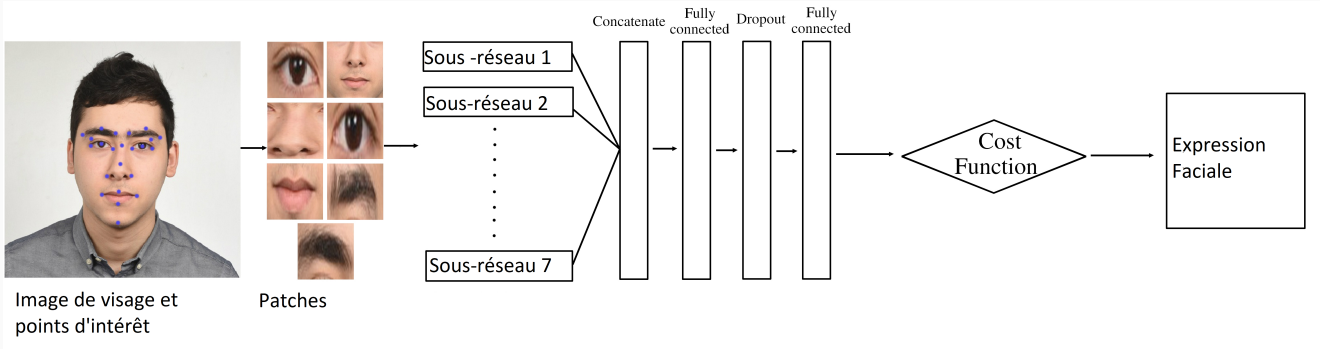
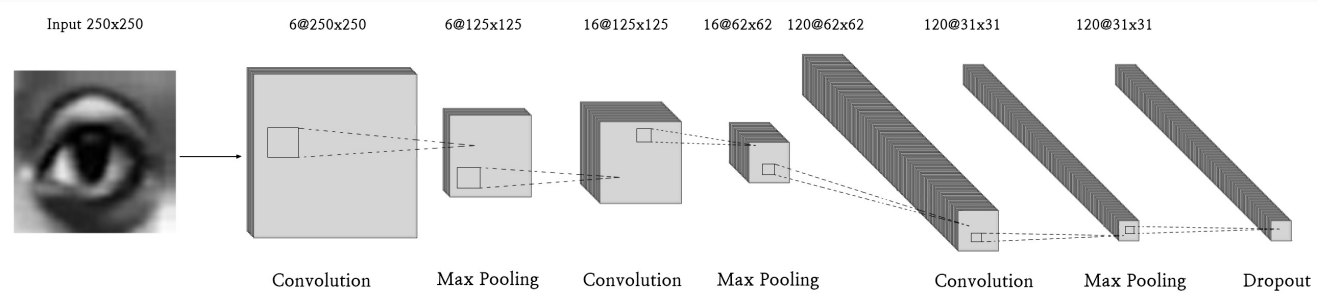


Illustration de l'approche proposée.

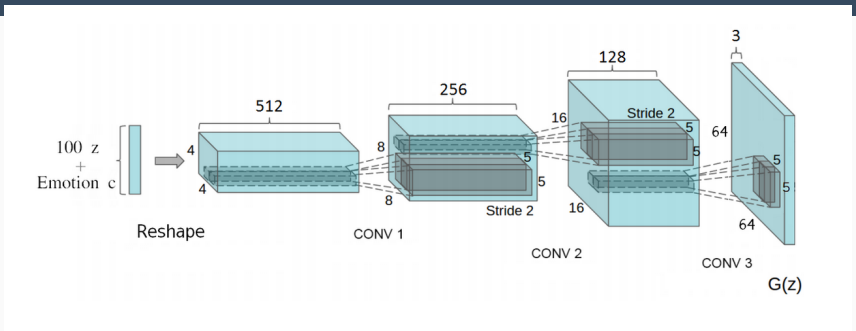


Réseau de neurones convolutifs d'agrégation de patchs faciaux (MFP-CNN).

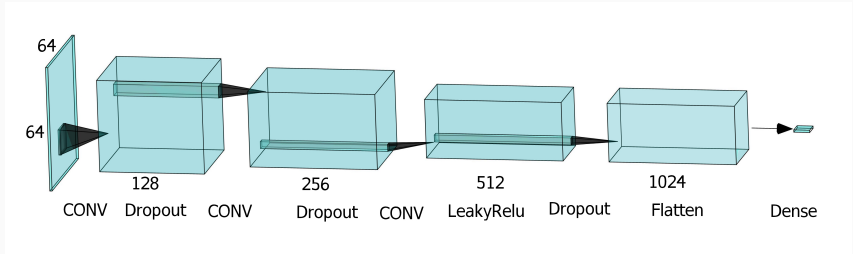


Structure de chaque sous-réseau pour chaque patch facial dans l'architecture MFP-CNN proposée.

Notre réseau antagoniste génératif conditionnel cGAN



Generator G.



Discriminator D.

A partir d'une image observée x et d'un vecteur de bruit aléatoire z, cGAN apprend une fonction de correspondance avec l'image de sortie y : $G(x, z) \longrightarrow y$. Cela peut être formulé comme un problème d'optimisation dont le but est de résoudre un problème min-max:

$$G^* = \arg\min_G \max_D \left(\mathcal{L}_{cGAN}(D) + \lambda \mathcal{L}_{cGAN}(G)\right) \tag{1}$$

où $\mathcal{L}_{cGAN}(D)$ est la fonction de perte du discriminateur D et $\mathcal{L}_{cGAN}(G)$ est la fonction de perte ou fonction objectif (loss function en anglais) du générateur G telle que définie ci-après:

$$\mathcal{L}_{cGAN}(G) = 1/N \sum_{i=1}^{N} \left(\mathcal{L}_{ad} + \alpha \mathcal{L}_{MSE} + \beta \mathcal{L}_{PEP}\right) \tag{2}$$

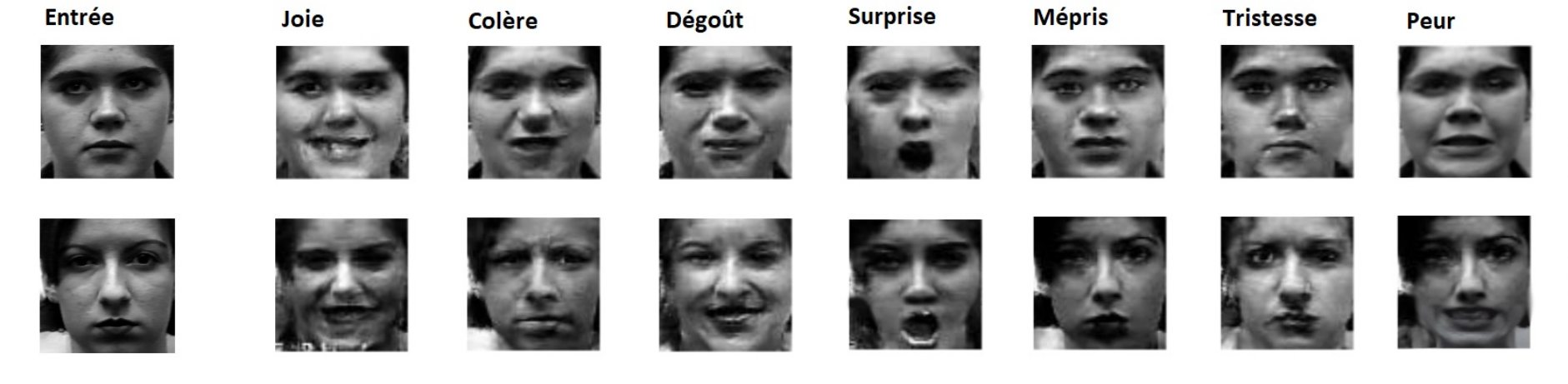
où N est le nombre total d'images d'entraînement, $\mathcal{L}_{ad}$ est la perte contradictoire, $\mathcal{L}_{MSE}$ et $\mathcal{L}_{PEP}$ sont respectivement la perte MSE pixel par pixel et la perte perceptive entre les échantillons régénérés et celles de l'apprentissage. La fonction de perte du discriminateur peut être exprimée comme suit:

$$\mathcal{L}_{cGAN}(D) = 1/N \sum_{i=1}^{N} \left(\log D(x, y) + \log(1 - D(x, G(x, z)))\right) \tag{3}$$

Évaluation des performances du réseau MFP-CNN

Expérience	Base de l'apprentissage	Base de test	Taux de bonne classification
1	CK+	CK+	89.77%
2	CK+ et images générées par cGAN	CK+	96.61%
3	CK+ et patches générés	CK+	97.96 %
4	CK+ et images et patches générés	CK+	**98.07**%

Résumé des performances du MFP-CNN dans différentes expériences et configurations.



Contributions des régions à l'expression des émotions

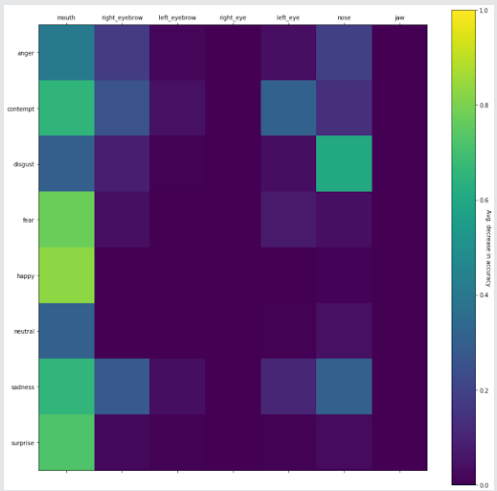
Contribution des différentes régions faciales à la reconnaissance des expressions

- Une étude d'ablation avec sept expériences sur la base de données CK+ pour étudier l'importance de chaque région faciale.
- L'importance de chaque zone faciale est définie comme **la différence de performance entre le MFP-CNN avec ses sept sous-réseaux et le MFP-CNN après suppression du sous-réseau correspondant à la zone faciale.**

Le sous-réseau supprimé	Taux de bonne classification of MFP-CNN
La mâchoire	96.6%
Le nez	87%
L'œil gauche	90%
L'œil droite	88%
Sourcil gauche	91%
Sourcil droite	86%
La bouche	**40.7%**

Participation des différentes régions du visage dans l'expression de chaque émotion

- la base de données est corrompue et sept expériences sont effectuées. Dans chaque expérience, le MFP-CNN est utilisé pour prédire l'ensemble de données après avoir remplacé à chaque fois l'une des régions du visage (par exemple la bouche) par des échantillons aléatoires de la même région du visage tirés de différents classes d'émotion.
- L'importance de chaque région dans chaque classe d'émotion est définie comme **la diminution moyenne du taux de bonne classification après corruption.**



Généralisation à d'autres bases de données

Deux bases de données sont considérées.

- JAFFE [9] : c'est une base de données d'expression faciale féminine japonaise, acquise dans des conditions contrôlées en laboratoire exactement comme le jeu de données CK+.
- SFEW [2]: acquise dans des conditions non contrôlées de laboratoire.

Expérience	Base d'entraînement	Base de Test	Performance
Originale	CK+	CK+	98.07%
EXP1: Généralisation	CK+	JAFFE	61, 97%
EXP2: Généralisation	CK+	SFEW	32, 42%

Lorsqu'un le transfer de connaissances est appliqué à EXP2, la précision du FER augmente considérablement et atteint 86, 36%.

Manque de Généralisation

Vers plus de généralisabilité des réseaux de neurones profonds

Motivations

- Manque de capacité de généralisation à travers des ensembles de données acquis dans différentes conditions.
- La majorité des approches existantes :
  - apprend ou affine un réseau de bout en bout qui ne peut être utilisé que pour un ensemble de données spécifique,
  - et/ou utilise des caractéristiques plus générales en entrée d'un modèle plus basique.

Objectifs

- Apprendre un extracteur de caractéristiques indépendant et général d'expressions faciales.
- Capacité de généralisation à d'autres ensembles de données non vus et non utilisés pendant l'entraînement, et sans aucun ajustement ou réglage (fine-tuning en anglais).

Vers plus de généralisabilité des réseaux de neurones profonds

- Nous avons proposé le Deep Facial Expression Vector ExtractoR (DeepFEVER).
- DeepFEVER est un réseau de neurones convolutifs (CNN) entraîné à l'aide de:
  - plusieurs bases de données REF étiquetées,
  - en employant la distillation des connaissances (en particulier, *auto-distillation*),
  - en considérant des données supplémentaires non étiquetées pour la reconnaissance des expressions faciales.

Deep Facial Expression Vector ExtractoR (DeepFEVER)

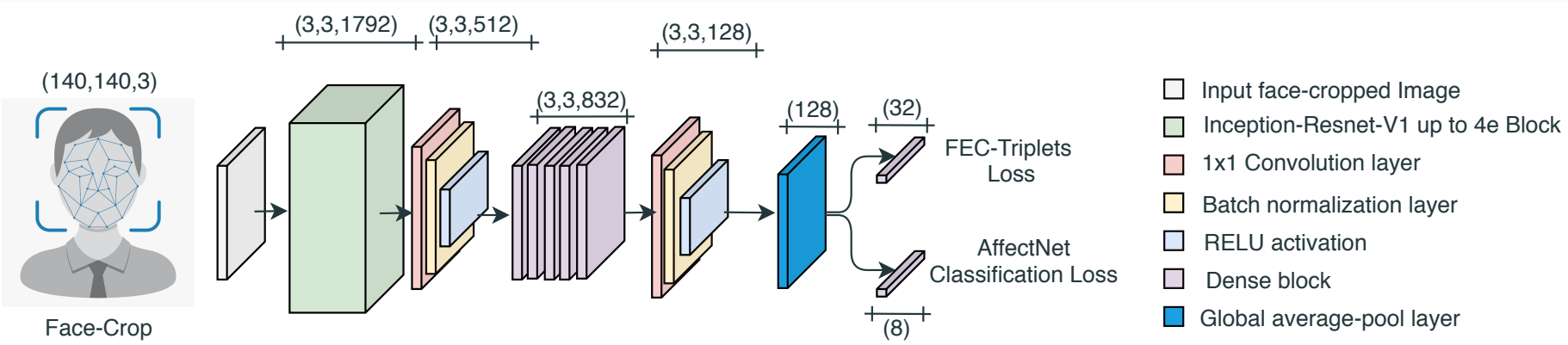
Algorithm 1: DeepFEVER : training the teacher network. Given feature extractor network f_Θ, Google FEC output head g_ϕ, AffectNet output head h_θ, number of training steps N, AffectNet loss weight α.

for *iteration in range(N)* **do**

- $(\mathbf{X}_{FEC}, \mathbf{y}_{FEC}) \leftarrow$ batch of Google FEC triplets and labels
- $(\mathbf{X}_{Aff}, \mathbf{y}_{Aff}) \leftarrow$ batch of AffectNet images and class labels
- $\mathbf{e}_{FEC} \leftarrow f_\Theta(\mathbf{X}_{FEC})$ Face embeddings for FEC images
- $\mathbf{e}_{Aff} \leftarrow f_\Theta(\mathbf{X}_{Aff})$ Face embeddings for AffectNet images
- $\mathbf{v}_{FEC} \leftarrow g_\phi(\mathbf{e}_{FEC})$ Predict vectors for triplet loss
- $\mathbf{p}_{Aff} \leftarrow h_\theta(\mathbf{e}_{Aff})$ Predict class probabilities for AffectNet
- $L_{FEC} = \texttt{triplet\_loss}(\mathbf{v}_{FEC}, \mathbf{y}_{FEC})$
- $L_{Aff} = \texttt{cross\_entropy\_loss}(\mathbf{p}_{Aff}, \mathbf{y}_{Aff})$
- $L = L_{FEC} + \alpha * L_{Aff}$ Total loss for training step
- Obtain all gradients $\Delta_{all} = (\frac{\partial L}{\partial \Theta}, \frac{\partial L}{\partial \phi}, \frac{\partial L}{\partial \theta})$
- $(\Theta, \phi, \theta) \leftarrow \texttt{SGD}(\Delta_{all})$ Update feature extractor and output heads' parameters simultaneously

end

Deep Facial Expression Vector ExtractoR (DeepFEVER)



L'architecture du réseau de neurones pour la reconnaissance des expressions faciales, avant la distillation (c'est-à-dire le *réseau maître*).

- Les visages sont détectés et rognés à l'aide de MTCNN[40].
- Les images RGB et de taille 140x140 résultantes sont ensuite transmises à un Inception Resnet V1 jusqu'au 4 ème bloc Inception. Ceci est suivi d'une couche de convolution 1x1 (1x1 Conv), d'une normalisation (BN) et d'une activation ReLU. Ensuite, cinq blocs DenseNet sont appliqués. Enfin, un autre ensemble de 1x1 Conv, BN et ReLU sont appliqués.
- La sortie est ensuite moyennée sur les dimensions spatiales, ce qui donne un vecteur de taille D_{face} (dans la figure, $D_{\text{face}} = 128$). Deux couches linéaires distinctes donnent ensuite les sorties finales du modèle: un vecteur pour la tâche des triplets Google FEC et des logits de classe pour AffectNet. Le modèle est formé pour minimiser simultanément les pertes AffectNet et Google FEC. Les nombres dans chaque bloc représentent la forme de sortie du tenseur après l'application de ce bloc.

Vers plus de généralisabilité des réseaux de neurones profonds

Méthodes	8-classes	7-classes
CNNs and BOVW + local SVM [3]	59.6%	
Pyramid with Super Resolution [30]	60.7%	63.8%
PAENet [8]		65.3%
DeepFEVER (Réseau maître)	61.3%	65.4%
DeepFEVER (Réseau élève, sans distillation)	58.8%	62.6%
DeepFEVER (Réseau élève avec distillation, sans PowderFaces)	61.1%	65.2%
DeepFEVER (Réseau élève avec distillation)	**61.6%**	**65.4%**

TAB1: Taux de bonne classification de DeepFEVER sur l'ensemble de validation AffectNet (7 ou 8 classes) par rapport aux méthodes de pointe existantes

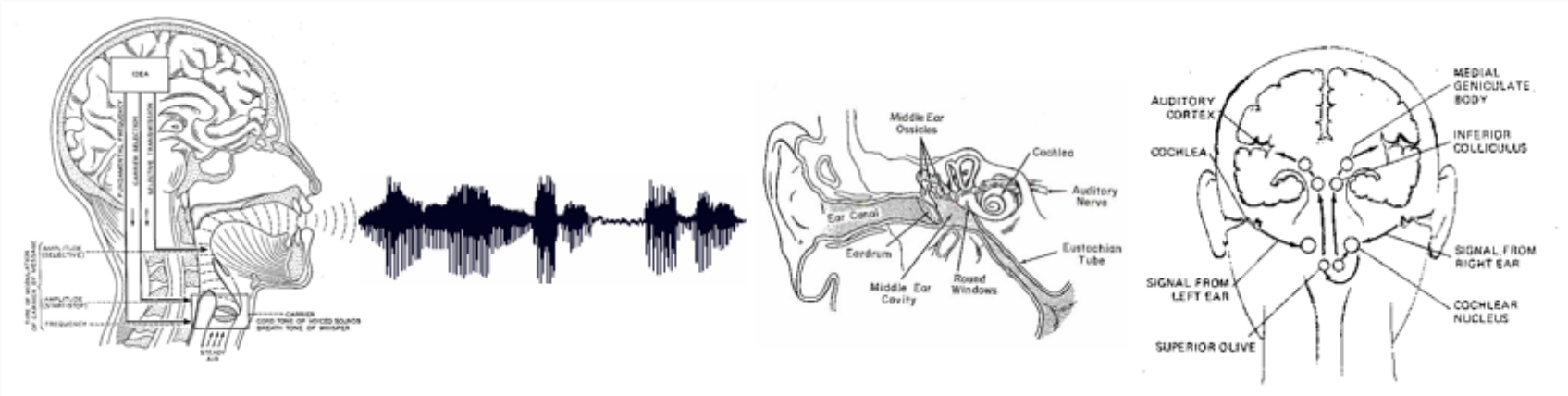
Méthodes	Taux de bonne classification
FaceNet avec fine-tuning [29]	81.8%
DeepFEVER (Le réseau maître)	84.5%
DeepFEVER (Le réseau élève sans distillation)	85.0%
DeepFEVER (Le réseau élève après distillation, sans PowderFaces)	86.4%
DeepFEVER (Le réseau élève après distillation)	**86.5%**

TAB2:Acc. classification des triplets de DeepFEVER sur l'ensemble de test GoogleFEC par rapport aux approches existantes

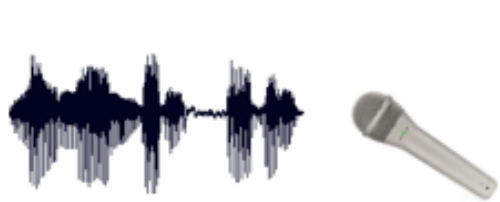
Méthodes	Taux de bonne classification
IPA2LT [38]	86.8%
RAN-ResNet18 [31]	86.9%
PSN [30]	**89.0%**
DeepFEVER (Le réseau élève après distillation, **pas de fine-tuning**)	87.4%

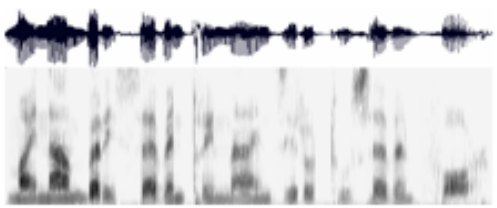
TAB3: Performances de DeepFEVER sur l'ensemble de test standard de la base RAF (7 classes) par rapport aux méthodes de pointe existantes

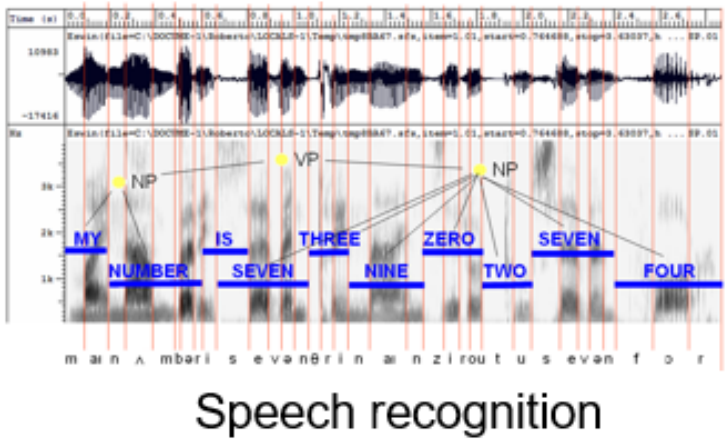
Une intelligence pour la voix et les emotions



L'articulation produit des ondes sonores que l'oreille transmet au cerveau pour traitement.







Après la numérisation, l'analyse acoustique du signal de parole est réalisé par machine.
Tout comme le visage, la voix véhicule une information importante sur l'identité, l'âge, le sexe et l'émotion.

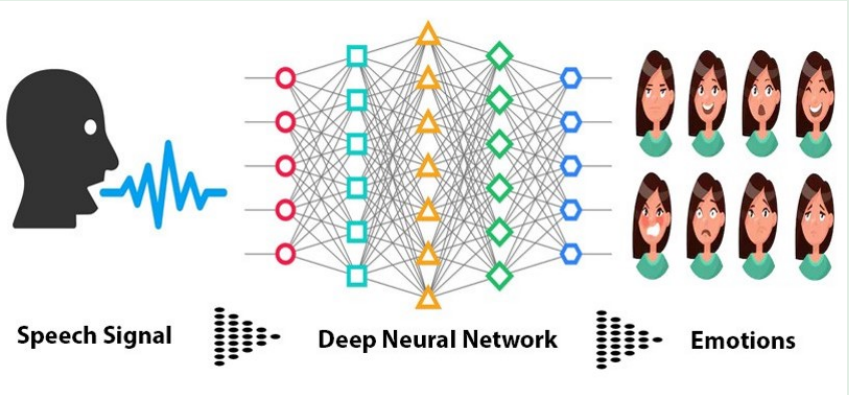
Première approche automatique pour la reconnaissance des emotions à partir des signaux audio

Motivations

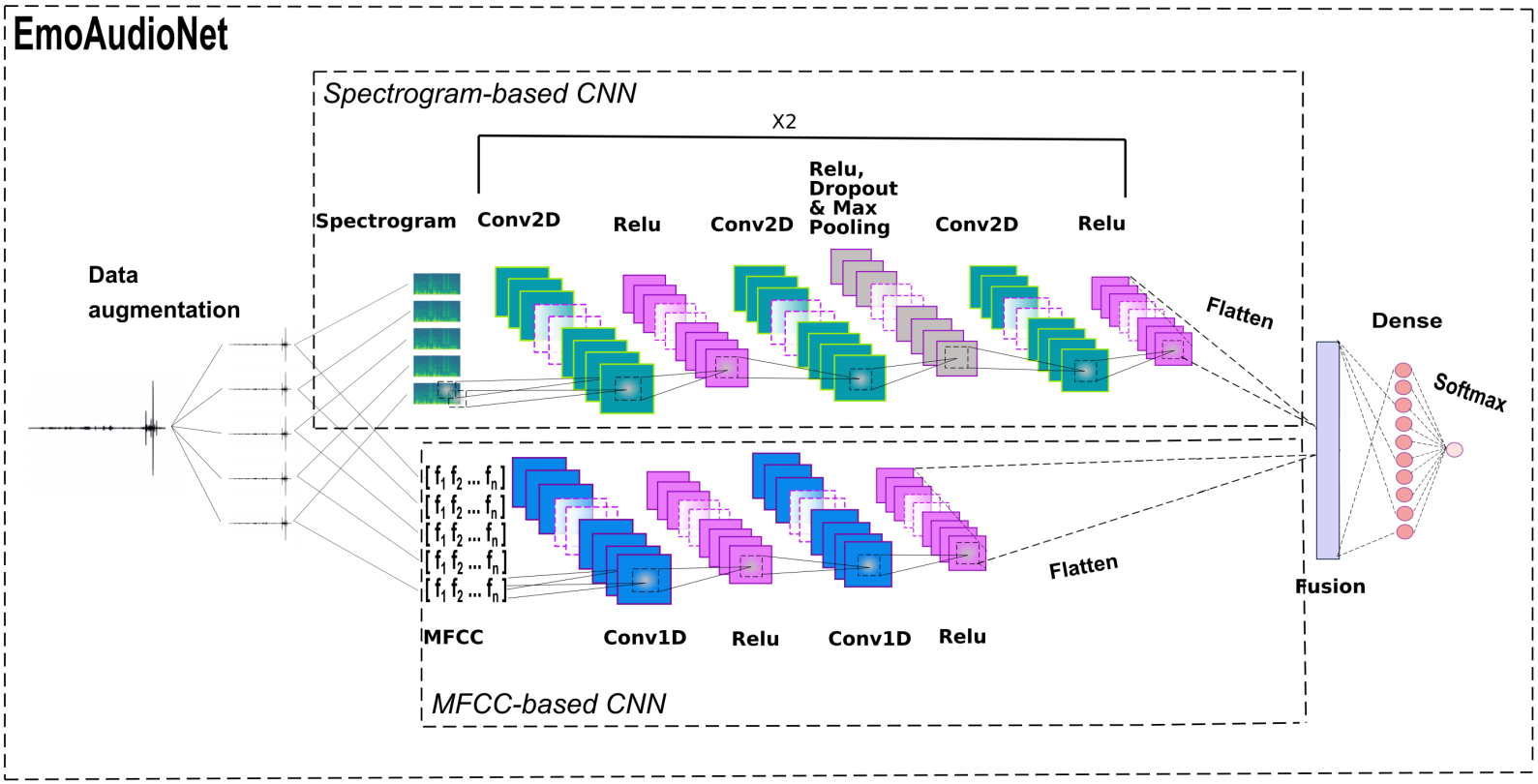
- Une bonne représentation paramétrique du signal audio améliora la reconnaissance des émotions.
- L'analyse spectrale à court terme est le moyen le plus courant de caractériser le signal vocal à l'aide des coefficients MFCC.
- Les signaux audio dans leurs représentations temps-fréquence, présentent souvent des motifs intéressants dans le domaine visuel [37].

Objectifs

- Développer un réseau neuronal profond qui combine des informations du temps, de fréquence et du domaine spacial (visuel) pour la reconnaissance des émotions.



Première approche automatique pour la reconnaissance des emotions à partir des signaux audio



Le schéma de l'architecture proposée des réseaux de neurones profonds appelée EmoAudioNet. La couche de sortie est une couche dense de neurones de taille n neurones avec une fonction d'activation Softmax. n est défini en fonction de la tâche, $n = 10$ pour la prédiction d'excitation ou de valence.

Résultats expérimentaux avec EmoAudioNet

	Développement			Test		
	Acc	CC	RMSE	Acc	CC	RMSE
CNN basé sur MFCC	83.37%	0.8289	0.1405	71.12%	0.6965	0.2082
CNN basé sur le spectrogramme	78.32%	0.7984	0.1446	73.81%	0.7598	0.2132
EmoAudioNet	95.42%	0.9568	0.0625	**91.44%**	**0.9221**	**0.1118**
EmoAudioNet (*)	87.55%	0.9028	0.1222	83.07%	0.8624	0.1508

Performances des réseaux de neurones profonds dans la prédiction de la valence. Les résultats obtenus pour les ensembles de données de développement et de test de la base de données RECOLA en termes de trois métriques : la précision, le coefficient de corrélation de Pearson (CC) et l'erreur quadratique moyenne (RMSE). (*) EmoAudioNet est pré-entraîné sur le jeu de données LibriSpeech et affiné sur le jeu de données RECOLA.

	Développement			Test		
	Acc	CC	RMSE	Acc	CC	RMSE
CNN basé sur les MFCC	81.93%	0.8130	0.1501	70.23%	0.6981	0.2065
CNN basé sur le spectrogramme	80.20%	0.8157	0.1314	75.65%	0.7673	0.2099
EmoAudioNet	94.49%	0.9521	0.0082	89.30%	0.9069	0.1229
EmoAudioNet (*)	95.16%	0.9555	0.07895	**90.37%**	**0.9156**	**0.1180**

Performances des réseaux de neurones profonds dans la prédiction d'excitation. Les résultats obtenus pour les ensembles de données de développement et de test de la base de données RECOLA en termes de trois métriques : la précision, le coefficient de corrélation de Pearson (CC) et l'erreur quadratique moyenne (RMSE). (*) EmoAudioNet est pré-entraîné sur le jeu de données LibriSpeech et affiné sur le jeu de données RECOLA.

Deuxième approche automatique pour la reconnaissance des émotions à partir des signaux audio

Algorithm 1: Notre algorithme d'ajustement du VGGish pour la prédiction du niveau d'excitation/arousal. Etant donné un extracteur de caractéristiques du réseau VGGish f_Θ, une tête de prédiction d'excitation f_ϕ, un nombre d'étapes d'entraînement N.

for *iteration in range(N)* **do**

- $(\mathbf{X}, \mathbf{y}) \leftarrow$ batch of RECOLA spectrograms and targets
- $\mathbf{e} \leftarrow f_\Theta(\mathbf{X})$ ▷ Calculate VGGish embeddings for batch
- $\mathbf{p} \leftarrow f_\theta(\mathbf{e})$ ▷ Predict arousal for all elements in batch
- $Loss = -\texttt{concordance\_correlation\_coeff}(\mathbf{p}, \mathbf{y})$
- Obtain all gradients $\Delta_{\text{all}} = (\frac{\partial Loss}{\partial \Theta}, \frac{\partial Loss}{\partial \theta})$
- $(\Theta, \theta) \leftarrow \texttt{Adam}(\Delta_{\text{all}})$ ▷ Update VGGish model, output head

end

L'approche est basée sur un ajustement/réglage (fine-tuning en anglais) du modèle VGGish [7] sur l'ensemble de données RECOLA [21].

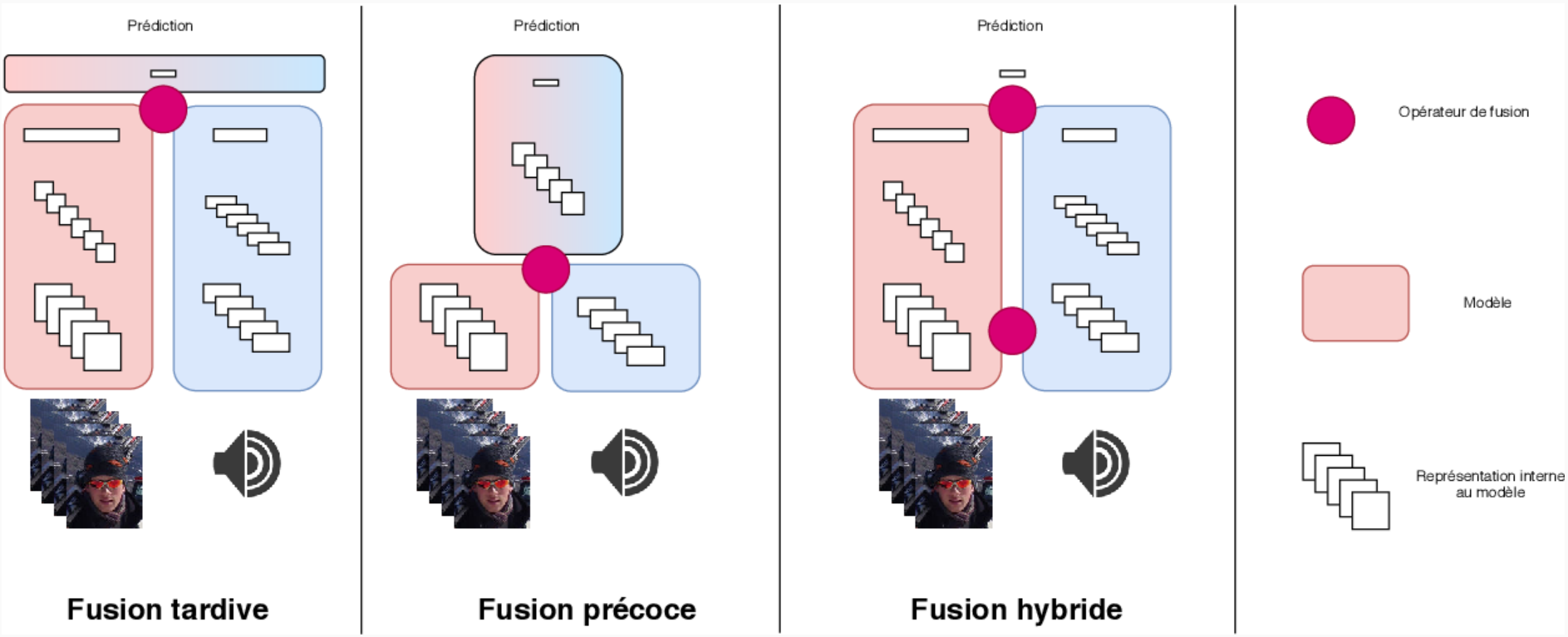
Résultats expérimentaux de la deuxième approche

Méthodes	Excitation	Valence
[26]	.70 (.75)	.31 (.41)
[5]	.67 (.76)	.36 (.48)
[6]	——(.80)	——(.40)
Notre réseau VGGish re-entraîné	**.70 (.80)**	**.52 (.46)**

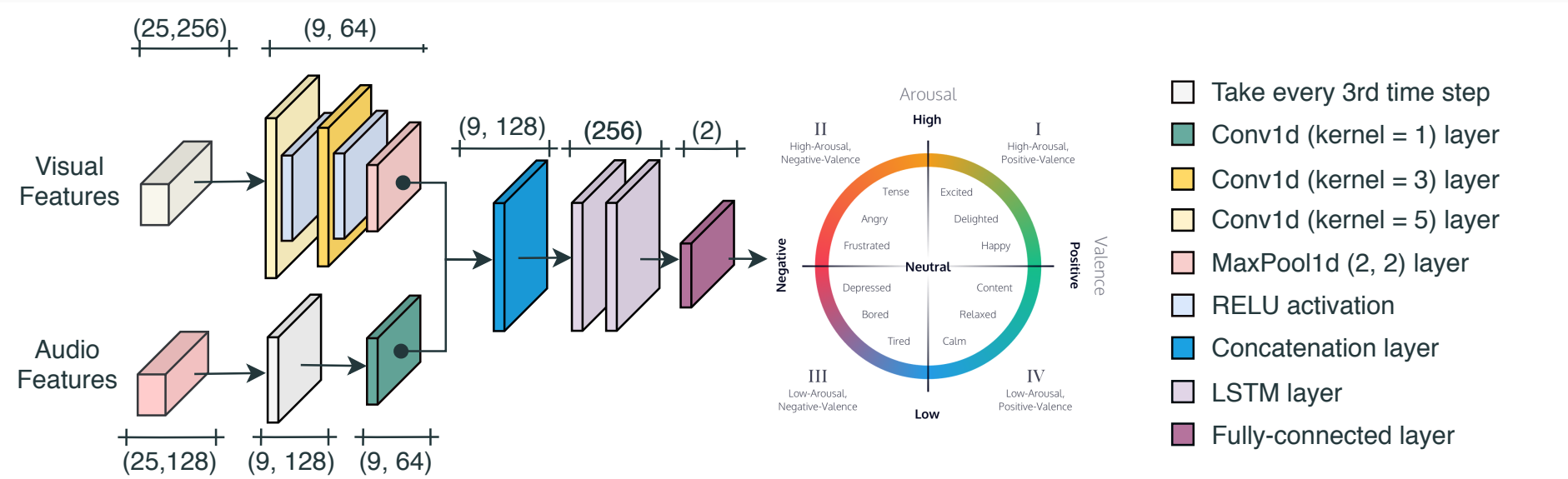
Performances du réseau VGGish ré-entrainé et testé sur l'ensemble de données RECOLA en termes de coefficient de corrélation CC par rapport aux méthodes de pointe existantes. Entre parenthèses sont les performances obtenues dans l'ensemble de développement. —— : aucun résultat rapporté dans les articles originaux.

$CC(X,Y) = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$ avec $COV(X,Y)$ est la covariance, σ_X est la déviation standard de X et σ_Y est la déviation standard de Y.

Fusion de données multimodales



Fusion des données audiovisuelles pour la reconnaissance des émotions



Notre approche de fusion multimodale basée sur un réseau de neurones profond.

Les représentations audio et visuelles unimodales sont concaténées et passées à un réseau profond récurrent type LSTM.

Fusion des données audiovisuelles pour la reconnaissance des émotions

	Valence			**Excitation**		
	Train	Dev	Test	Train	Dev	Test
Visuel	.6	.55	.66	.49	.57	.57
Audio	.55	.46	.52	.78	.80	.70
Audio-visuel	.69	.63	.74	.78	.81	.72

$$CCC(Tr, Pr) = \frac{2CC\sigma_{Tr}\sigma_{Pr}}{\sigma_{Tr}^2 + \sigma_{Pr}^2 + (\mu_{Tr} - \mu_{Pr})^2} \quad (4)$$

σ_{Tr} and σ_{Pr} représentent les deviations standard des variables Tr et Pr, et μ_{Tr} and μ_{Pr} représentent leurs moyennes respectives.

Plan de la présentation

Intelligence Emotionnelle et troubles mentaux

- Il existe toute une gamme de troubles mentaux, qui se manifestent sous des formes différentes.
- Ils se caractérisent généralement par **un ensemble anormal de pensées**, **de perceptions**, **d'émotions**, **de comportements et de relations avec autrui**.
- Parmi les troubles mentaux figurent:
  - la dépression,
  - les troubles affectifs bipolaires,
  - la schizophrénie et autres psychoses,
  - la démence,
  - la déficience intellectuelle et les troubles du développement, y compris l'autisme.
  - Le syndrome de stress post-traumatique (PTSD) ou trouble stress post-traumatique (TSPT)

Intelligence Emotionnelle et troubles mentaux

On considère que la pathologie survient lorsque l'émotion est dérégulée, que l'affect est réprimé, inaudible. L'intensité émotionnelle perçue ou non par le sujet est un point fondamental pour le clinicien.

—Aurélie Pasquier, Les troubles de l'émotion, Article publié dans la revue Santé mentale en avril 2013 (n°177, p. 32-36).

Le trouble Dépressif Majeur ou Dépression clinique

1
La dépression est le 1er facteur d'incapacité sur le plan mondial (OMS)

1 sur 5
Le nombre de personnes qui souffira d'une dépression au cours de sa vie

2
La dépression touche 2 fois plus les femmes que les hommes

3 millions
le nombre de personnes ayant vécu une dépression en France au cours des 12 derniers mois (INPES)

30
Le risque de tentative de suicide est multiplié par 30 en cas d'épisode dépressif (Inserm).

350 millions de personnes dans le monde sont diagnostiqués avec une dépression

50% Le nombre de personnes ayant une dépression qui ne sont pas traitées.

70% Les traitements sont efficaces dans près de 70% des cas

75% Le risque de rechute après une première dépression.

Le trouble Dépressif Majeur ou Dépression clinique

- Le diagnostic des troubles psychiatriques repose sur **l'inventaire de symptômes psychologiques en auto-évaluation**.
- Le DSM (Diagnostic and Statistical Manual of Mental Disorders) à sa cinquième version souffre de:
  - insuffisance et le flou des définitions des troubles.
  - incapacité à identifier systématiquement les faux positifs.
  - Définition basée sur les symptômes et qui ne prend pas en compte l'historique du patient,

  $\longrightarrow$ Implications thérapeutiques et pronostiques.
- Manque d'approches standardisées de diagnostic de dépression.
  - Hospital Anxiety and Depression Scale (HADS),
  - Quick Inventory of Depression Symptomatology (QIDS),
  - Beck's Depression Inventory (BDI)
  - Patient Heath Questionnaire (PHQ)
- **2/3** des personnes atteintes de dépression ne sont pas identifiées [12]

Le trouble Dépressif Majeur ou Dépression clinique

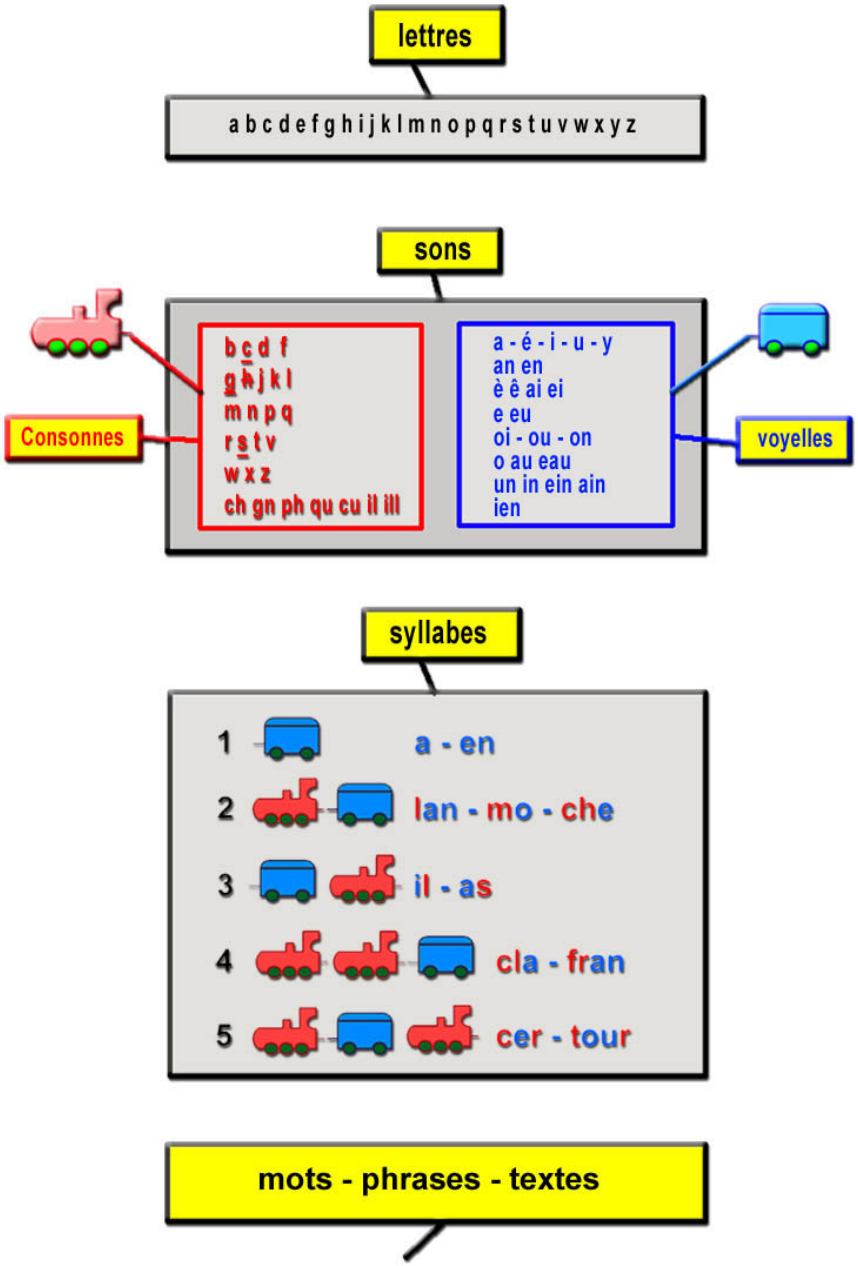
- La taxonomie psychiatrique (classification des troubles mentaux) classe la dépression parmi les humeurs basses [11]

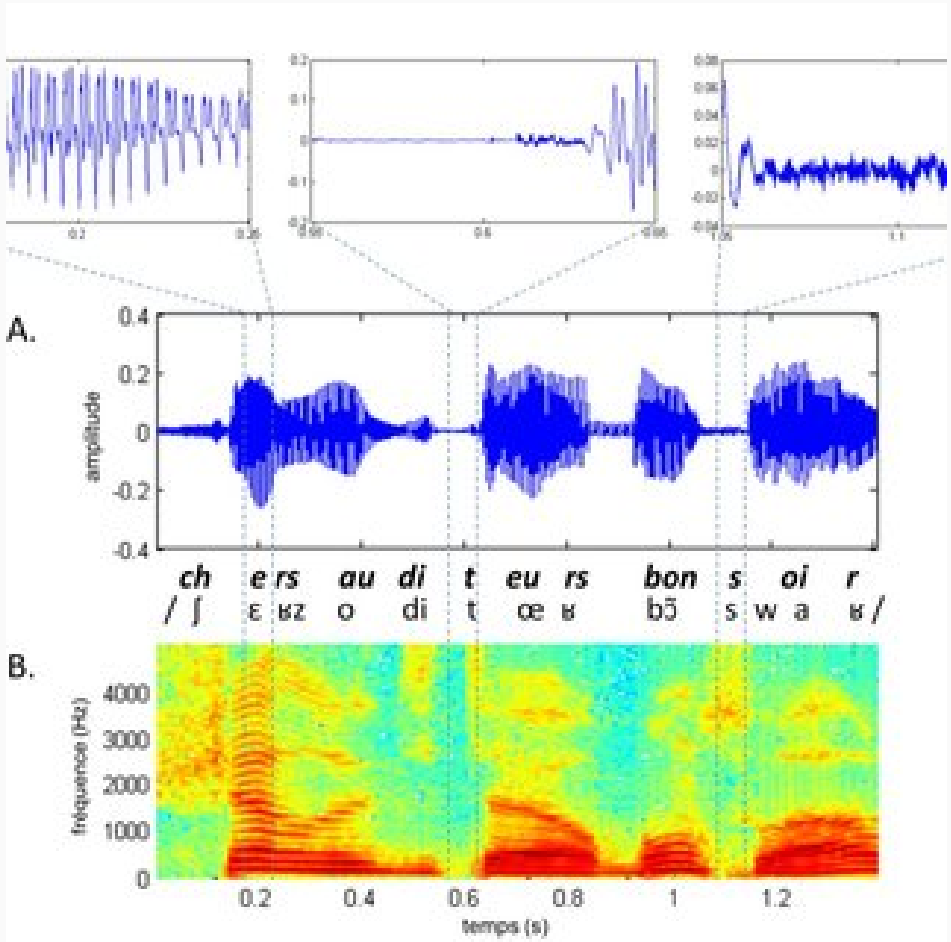
  $\longrightarrow$ une condition caractérisée par une fatigue et un **ralentissement global physique, intellectuel, social et émotionnel.**

La dépression décrit l'expression constante de certaines émotions: émotions négatives (**faible Valence**) et manque d'intérêt (**faible excitation ou faible arousal**)

Le discours des sujets dépressifs est ralenti, les pauses sont allongées et le ton de la voix (prosodie) est plus monotone.

Reconnaissance de la dépression à partir des signaux audio





Une architecture CNN profonde basée sur le niveau de phonème pour le diagnostic de la dépression

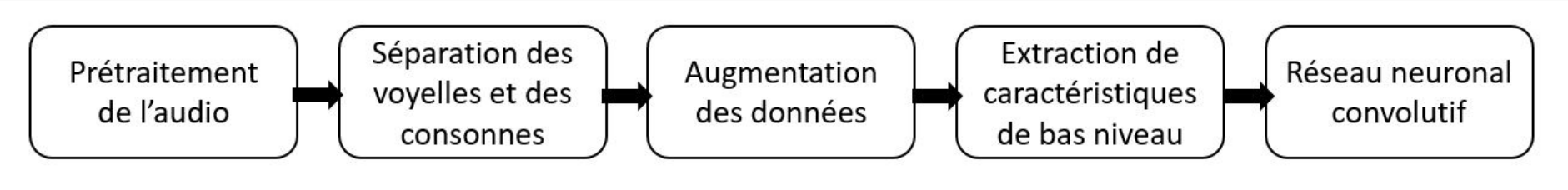


Schéma de l'approche proposée basée sur les voyelles et les consonnes extraites à partir d'un signal audio pour la reconnaissance de la dépression.

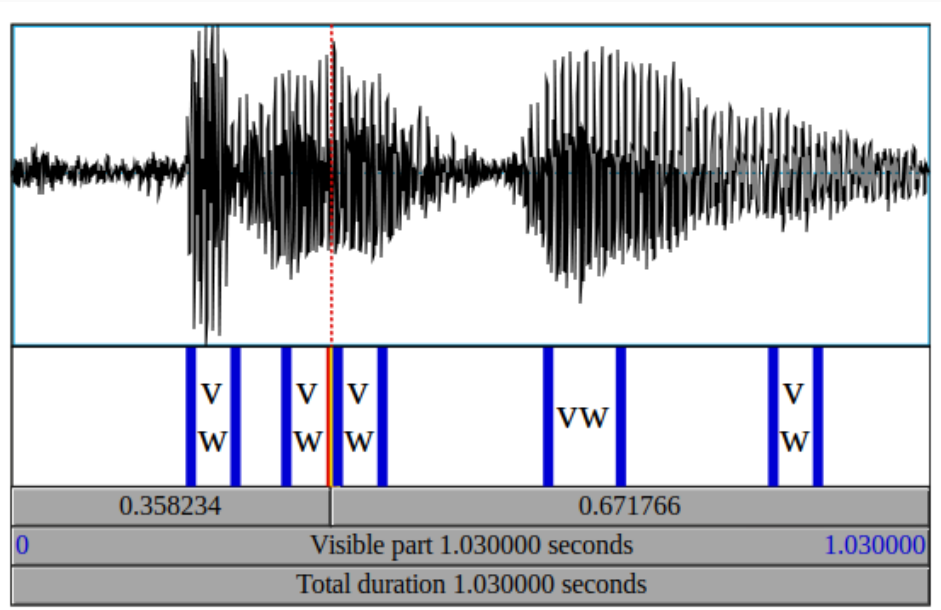
Extraction de la parole du patient

- Evaluation de la dépression à partir des réponses du patient aux questions cliniques.
- Un prétraitement des enregistrements audio est effectué pour extraire le discours du patient de celui de l'intervieweur.

Une architecture CNN profonde basée sur le niveau de phonème pour le diagnostic de la dépression

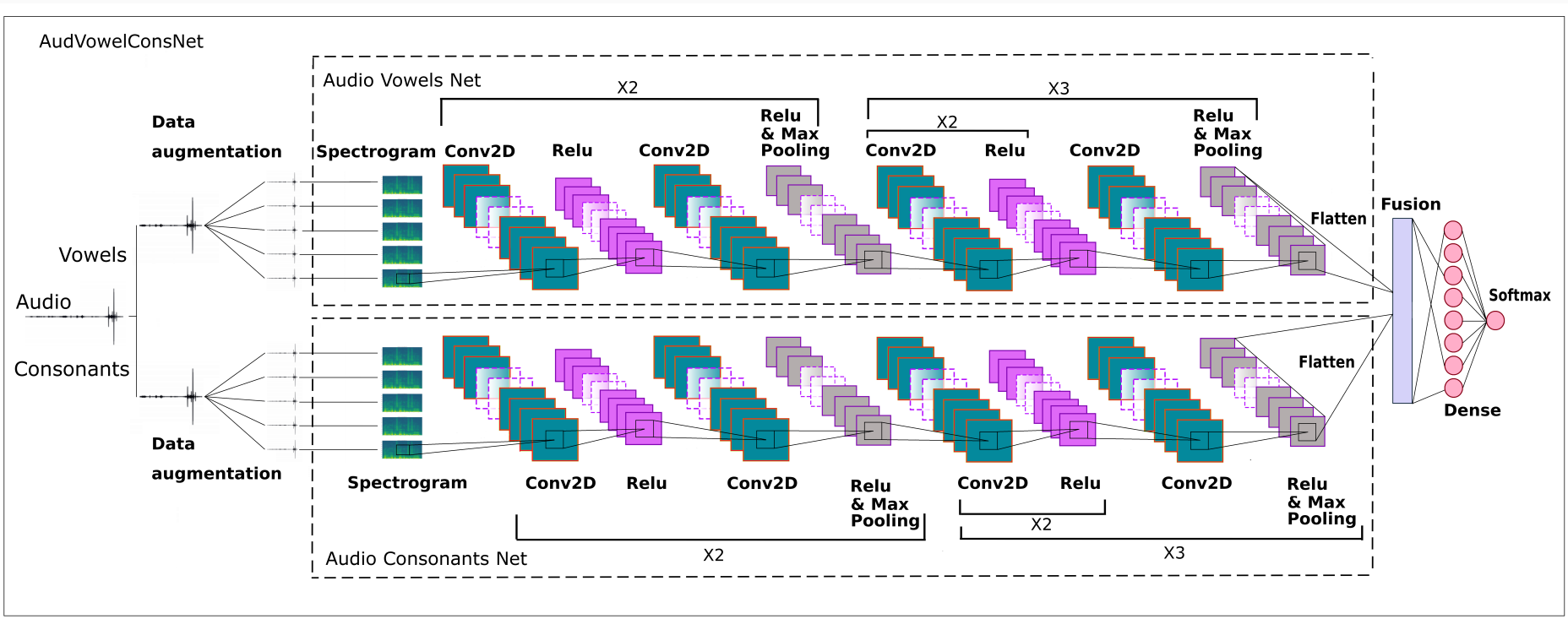
Segmentation de la parole

- Fenêtrage,
- Seuillage basé sur l'amplitude et la force du signal.



Séparation des voyelles et consonnes audio. Les lignes bleu foncé montrent les limites des voyelles.

Une architecture CNN profonde basée sur le niveau de phonème pour le diagnostic de la dépression



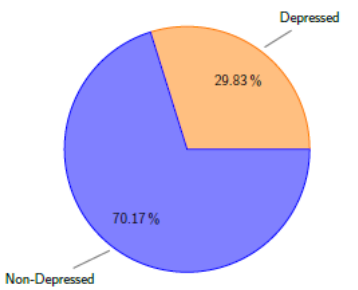
L'architecture AudVowelConsNet: un CNN basé sur le phonème qui fusionne les caractéristiques acoustiques des voyelles et des consonnes via l'apprentissage profond.

Une architecture CNN profonde basée sur le niveau de phonème pour le diagnostic de la dépression

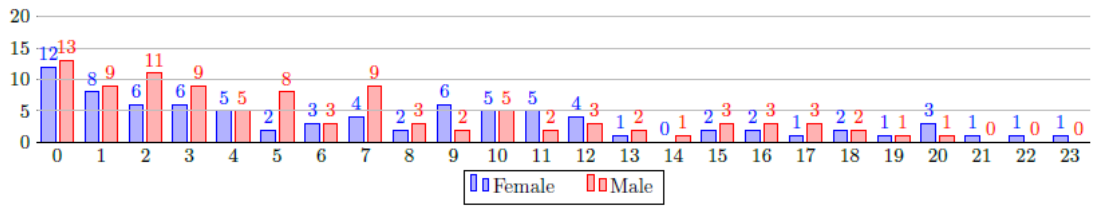
Dataset DAIC-WOZ

Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) [4]:

- Des entretiens cliniques enregistrés pour enquêter sur différentes conditions de détresse psychologique (dépression, l'anxiété et le trouble de stress post-traumatique).
- Enregistrements audio de 189 participants avec un intervieweur virtuel.
- Étiquetage par les scores PHQ-8 (niveau de gravité de la dépression, plage [0-23]) et le binaire PHQ-8 (1/0 dépression vs non dépression).
- La durée moyenne des enregistrements audio est de 15 minutes avec un taux d'échantillonnage de 16 kHz.



(a) Répartition des participants déprimés et non déprimés.



(b) Répartition des participants selon les niveaux de gravité du test PHQ.

Une architecture CNN profonde basée sur le niveau de phonème pour le diagnostic de la dépression

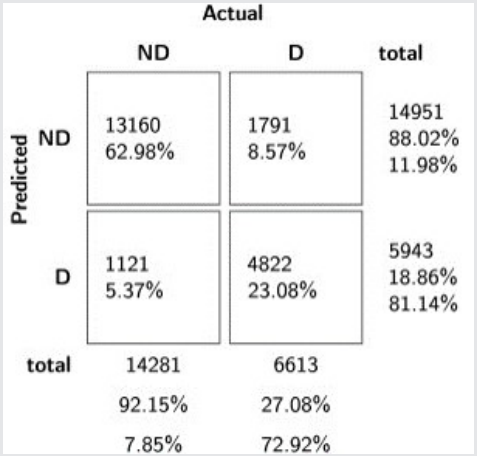
Réseau	PHQ-8 Binaire			PHQ-8 Score		
	Acc. (%)	CC	CCC	Accuracy (%)	CC	CCC
Audio Vowels Net	78.77	0.58	0.57	54.26	0.59	0.58
Audio Consonants Net	80.98	0.62	0.61	57.57	0.61	0.61
AudVowelConsNet	**86.06**	**0.72**	**0.72**	**70.86**	**0.73**	**0.73**

Performances des réseaux de neurones profonds proposés pour la détection de la dépression (PHQ-8 Binaire) et évaluation de sa gravité (PhQ-8 Score).

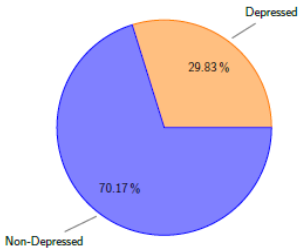
Réseau	Acc (%)	AUC	Precision (%)		Rappel (%)		F1-Score (%)		
			D	ND	D	ND	D	ND	Av.
Audio Vowels Net	78.77	0.75	67	84	66	85	66	85	78.99
Audio Consonants Net	80.98	0.76	73	84	64	89	68	86	80.30
AudVowelConsNet	**86.06**	**0.83**	**81**	**88**	**73**	**92**	**77**	**90**	**85.85**

Comparaison des architectures de réseaux de neurones profonds proposées pour la détection de la dépression pour les deux classes binaires de dépression (D) et de non-dépression (ND).

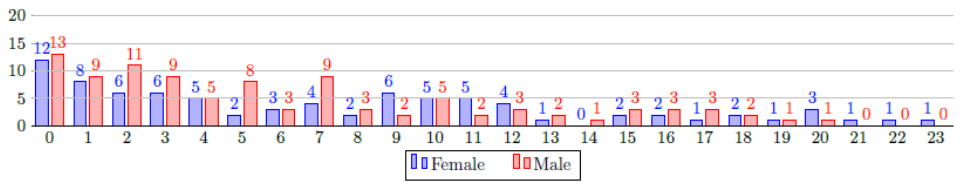
Une architecture CNN profonde basée sur le niveau de phonème pour le diagnostic de la dépression



Matrice de confusion d'AudVowelConsNet pour la classification binaire de la dépression.



(a) Répartition des participants déprimés et non déprimés.



(b) Répartition des participants selon les niveaux de gravité du test PHQ.

Répartitions des participants selon leurs niveaux de dépression dans la base DAIC-WOZ.

Une architecture CNN profonde basée sur le niveau de phonème pour le diagnostic de la dépression

Méthode	Précision (%)		Rappel (%)		F-Score (%)			Acc. (%)	RMSE	CC
	D	ND	D	ND	D	ND	Av.			
[10]	35	100	100	54	—	70	52	—	—	—
[23]	—	—	—	—	—	—	85.44	**96.7**	—	—
[16]	56.28	79.48	45.11	85.85	50	83	75.39	74.13	0.47	0.51
[20]	69	78	35	94	46	85	80.00	76.27	0.41	—
AudVowelConsNet	**81**	**88**	**73**	**92**	**77**	**90**	**85.85**	86.06	**0.37**	**0.72**

Comparaison des performances du réseau de neurones profonds proposé avec des méthodes de l'état de l'art pour la reconnaissance de la dépression en termes de précision, rappel et F-score.
(D) et (ND) sont respectivement pour les classes de dépression et de non-dépression.

Méthode	RMSE
[**valstar2016**]	7.78
[34]	5.59^{DM}
[**yang2017b**]	1.46^{DM}
[**EmnaRejaibi2019**]	0.18^{Norm}
[**rejaibi2019**]	0.17^{Norm}
AudVowelConsNet	0.14^{Norm} — **3.22**

Comparison des performances d'AudVowConsNet avec les méthodes existantes dans la prédiction du score du test PhQ. (^{DM}) : Homme déprimé. (^{Norm}): RMSE normalizé

Reconnaissance multimodale de la dépression

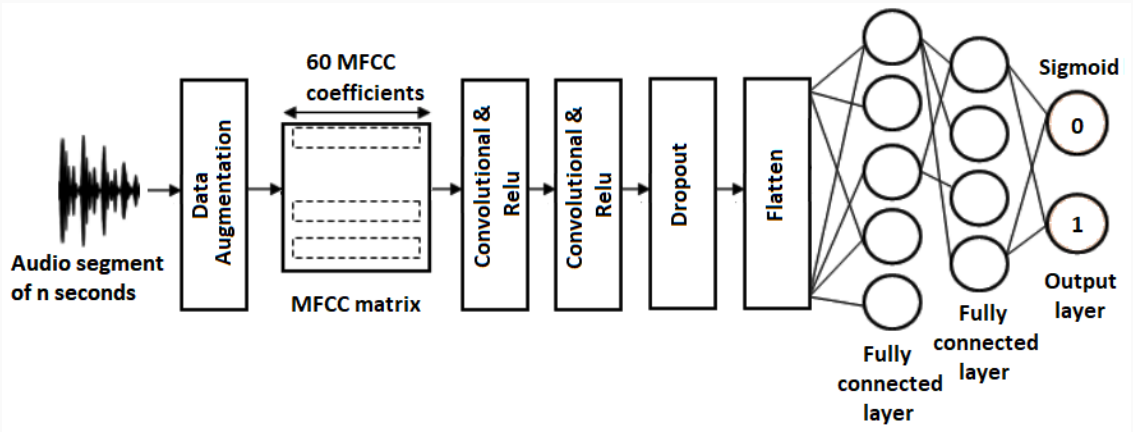
Motivations

- Il a été démontré que les approches basées sur la prédiction de la dépression à l'aide du signal audio de la parole dépassent les approches qui utilisent l'expression faciale ou le texte [32, 35].
- Les coefficients MFCC ont prouvé leur grande efficacité dans la détection de la dépression clinique par rapport à d'autres caractéristiques audio de bas niveau [20].

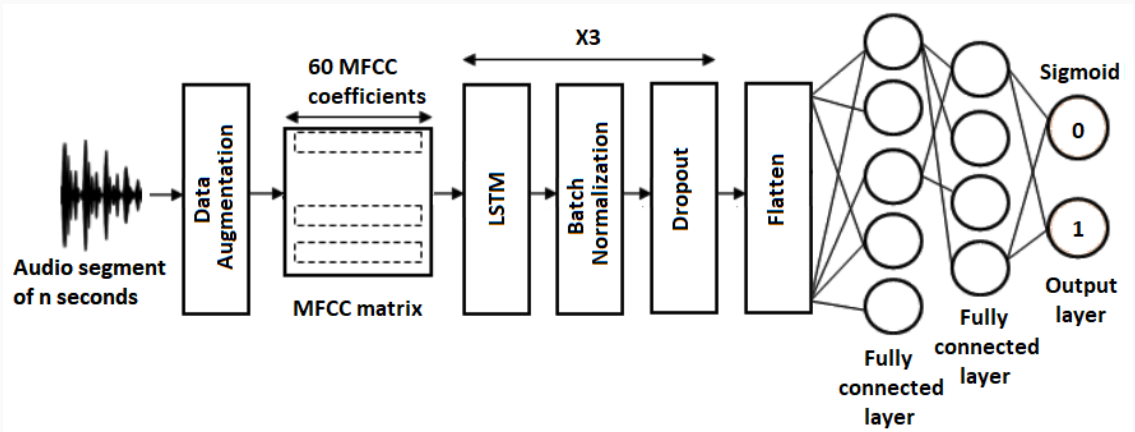
Objectifs

- Comparer plusieurs architectures de réseaux de neurones profonds.
- Comparer différentes stratégies d'apprentissage profond des caratéristiques et de fusion pour la reconnaissance multimodale de dépression.
- Apprendre des représentations de la dynamique temporelle à partir de signaux multidimensionnels et multimodaux.
- Développer une approche automatique multimodale et hautement performante, basée sur des très courtes séquences vidéo.

Représentations Unimodales profondes basées sur le MFCC pour la reconnaissance de la dépression



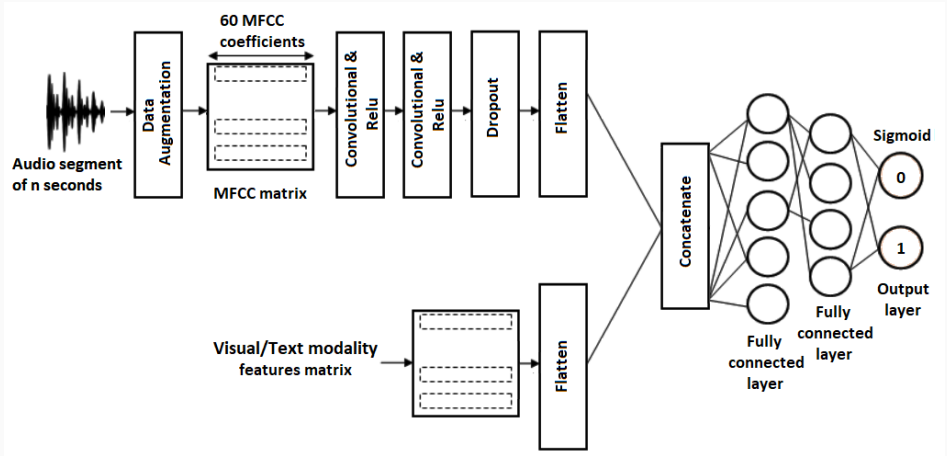
(a)



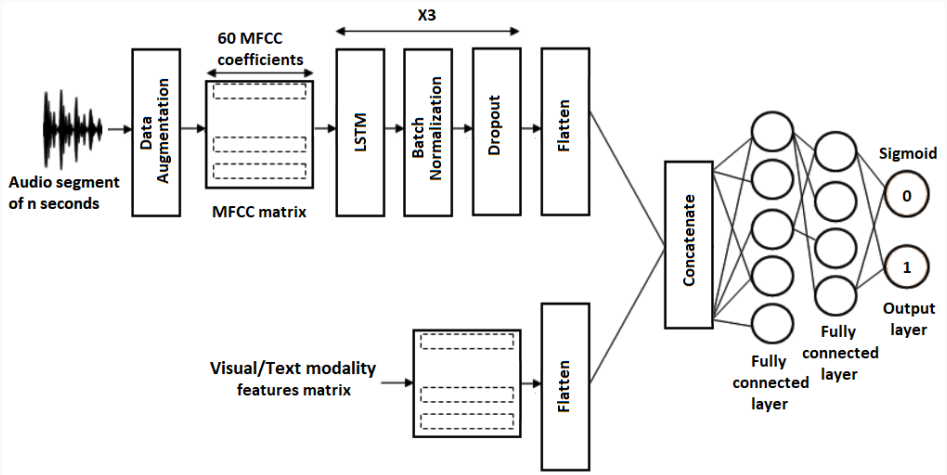
(b)

Représentations unimodales profondes basées sur la matrix (Image) des coefficients MFCC pour la reconnaissance de la dépression. 60 coefficients MFCC sont extraits du segment audio de taille 7, 6 secondes. Ensuite, ils sont transmis à un CNN (a) ou à un LSTM (b) suivi de deux couches Fully connected.

Fusion Multimodale pour la reconnaissance de la dépression



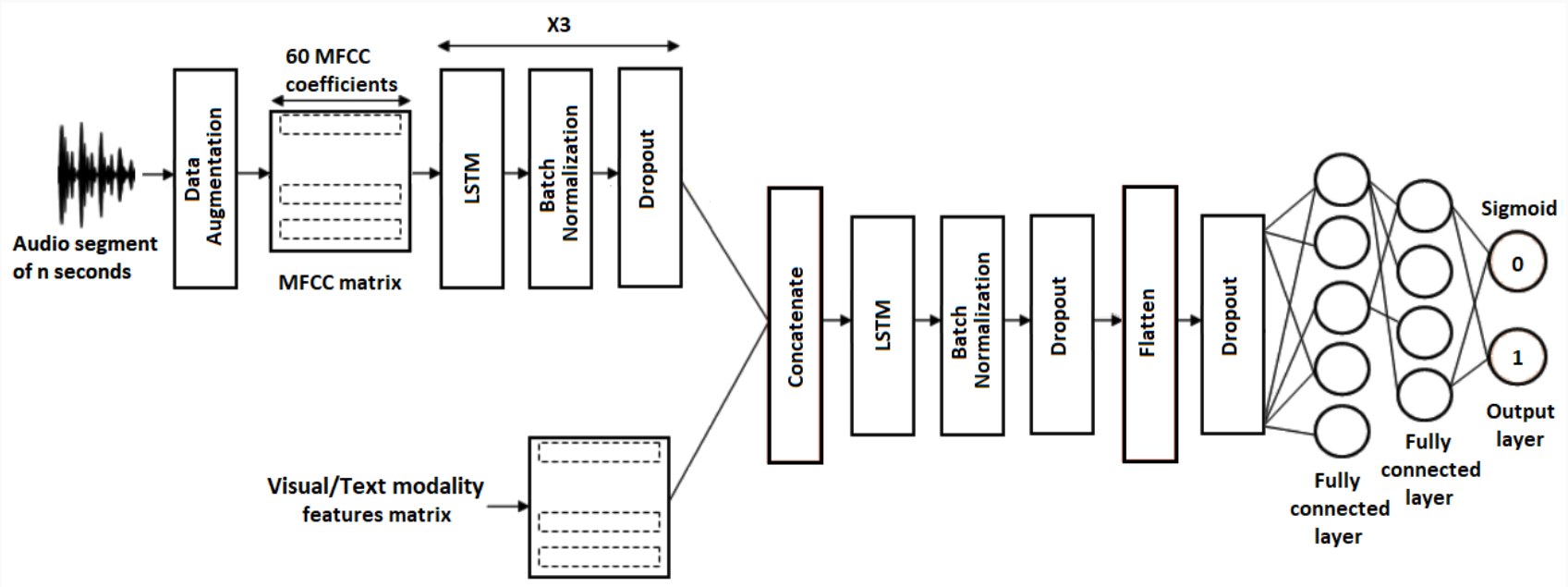
(a)



(b)

Approches multimodales basées sur la fusion précose pour la reconnaissance de la dépression. Après avoir extrait les représentations unimodales basées sur MFCC, les caractéristiques visuelles ou textuelles sont concaténées avec les caractéristiques audio profondes dans un seul vecteur de caractéristiques de grande dimension, puis transmises à deux couches FC.

Approche proposée de fusion basée sur modèle pour la reconnaissance de la dépression



L'approche de fusion proposée basée sur modèle pour la reconnaissance de la dépression. Les caractéristiques audio profondes basées sur MFCC sont concaténées avec les caractéristiques visuelles ou textuelles, puis transmises à un réseau de neurones profond basé sur LSTM pour apprendre une représentation multimodale conjointe pour la reconnaissance de la dépression.

Résultats Expérimentaux

Fusion	Caractéristiques	Réseau de neurones profond	AUC Score	Acc. (%)	RMSE	CC	CCC
	MFCC	CNN	0.4866	65.60	0.49	0.149	0.06
	MFCC	LSTM	0.4816	66.25	0.49	0.154	0.07
Précose	MFCC-Word2Vec	CNN	0.4740	64.29	0.49	0.12	0.06
	MFCC-Word2Vec	LSTM	0.4678	65.98	0.48	0.13	0.07
	MFCC-AU	CNN	0.5059	68.32	0.47	0.22	0.09
	MFCC-AU	LSTM	0.5391	71.33	0.46	0.35	0.17
Basée sur un modèle	MFCC-Word2Vec	LSTM	0.4690	68.79	0.46	0.20	0.12
	MFCC-AU	LSTM	**0.6575**	**77.16**	**0.42**	**0.54**	**0.34**

Performance des modèles d'apprentissage profond proposés pour la reconnaissance binaire de la dépression en termes de score AUC, de précision, de RMSE, de CC et de CCC sur l'ensemble de test. Acc. : taux de bonne classification.

Évaluation des niveaux de gravité de la dépression

Fusion	Caractéristiques	Réseau de neurones profond	RMSE
	MFCC	CNN	$0.2041^{N}/4.69$
	MFCC	LSTM	$0.2093^{N}/4.81$
Précose	MFCC-Word2Vec	CNN	$0.2175^{N}/5.00$
	MFCC-Word2Vec	LSTM	$0.2109^{N}/4.85$
	MFCC-AU	CNN	$0.1978^{N}/4.55$
	MFCC-AU	LSTM	$0.1862^{N}/4.28$
Basée sur un modèle	MFCC-Word2Vec	LSTM	$0.1945^{N}/4.47$
	MFCC-AU	LSTM	$\mathbf{0.1519}^{N}/\ \mathbf{3.49}$

Performances des modèles d'apprentissage profond proposés pour la prédiction des scores du test PhQ pour l'évaluation du niveau de gravité de la dépression en terme de RMSE sur l'ensemble de test. (^{N}): RMSE Normalisé.

L'expérience de validation croisée Leave-One-Subject-Out

Assessment task	Metric	Value
Binary	**AUC Score**	0.94
	Acc. (%)	95.38
	RMSE	0.22
	CC	0.94
	CCC	0.89
Severity Level	**RMSE**	**0.15**^{Norm} / 3.40

Expérience Leave-One-Subject-Out (LOSO) pour l'architecture la plus performante (MFCC-AU LSTM) pour les deux tâches d'évaluation de la dépression (classification binaire et niveau de gravité de la dépression) en termes de précision, RMSE, CC et CCC. (^{Norm}): RMSE Normalisé.

(a) MFCC AU

Predicted \ Actual	ND	D	total
ND	7190 (68.44)	357 (3.40)	95.27 4.73
D	128 (1.22)	2830 (26.94)	95.67 4.33
total	98.25 1.75	88.80 11.20	

La matrice de confusion de l'architecture la plus performante (MFCC-AU LSTM) en se basant sur une stratégie d'apprentissage Leave-One-Subject-Out. ND: Non-Depression, D: Depression.

Comparison avec les méthodes existantes

Method	Precision		Recall		F1-Score			Acc.	RMSE	CC
	D	ND	D	ND	D	ND	Av.			
DepAudioNet [10]	35	**100**	**100**	54	52	70	—	—	—	—
EmoAudioNet [15]	52	80	46	84	49	82	72.89	73.25	0.47	—-
MFCC-based RNN [20]	**69**	78	35	**94**	46	**85**	73.65	76.27	**0.41**	—
MFCC-AU LSTM	53	83	44	88	48	**85**	76.09	77.16	0.42	**0.54**
BLSTM-MIL* [23]	—	—	—	—	—	—	85.44	**96.7**	—	—
MFCC-AU LSTM*	96	95	89	98	92	97	**95.48**	95.38	0.22	0.94

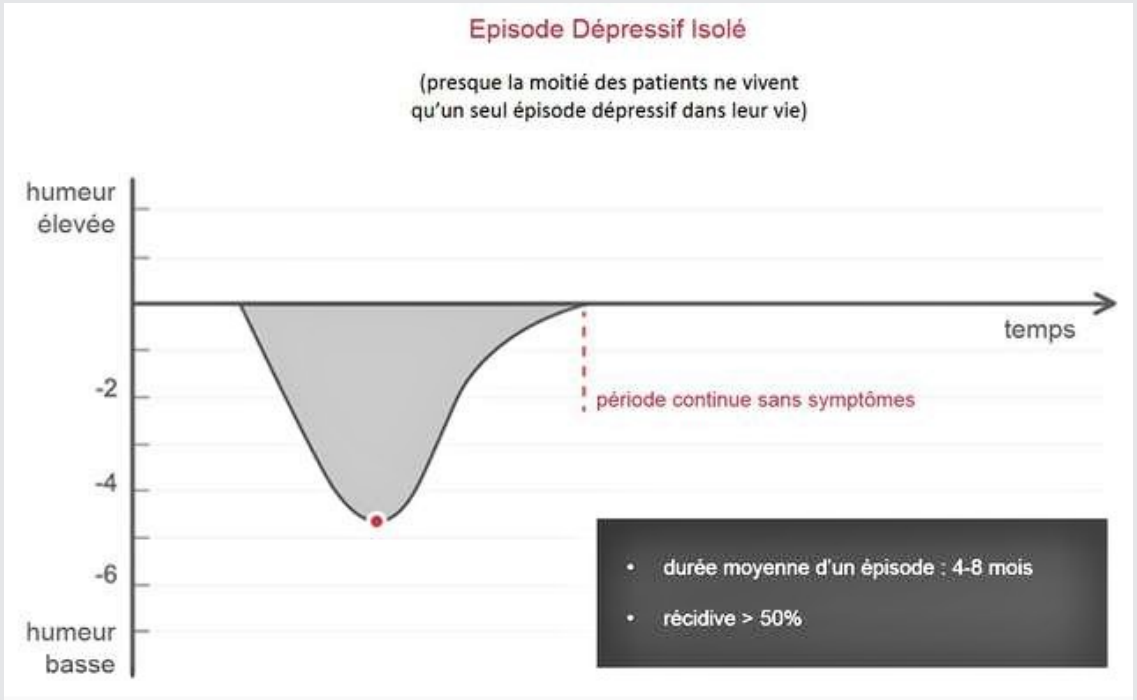
Comparaison du réseau proposé avec les méthodes de pointe pour la classification binaire du PHQ-8 en termes de précision, de rappel et de F1-score (%) pour les classes Dépression (D) et non dépression (ND). Le tableau résume également la moyenne du F-score, la précision, le RMSE et le CC. (*) Évaluation avec la stratégie Leave-One-Subject-Out.

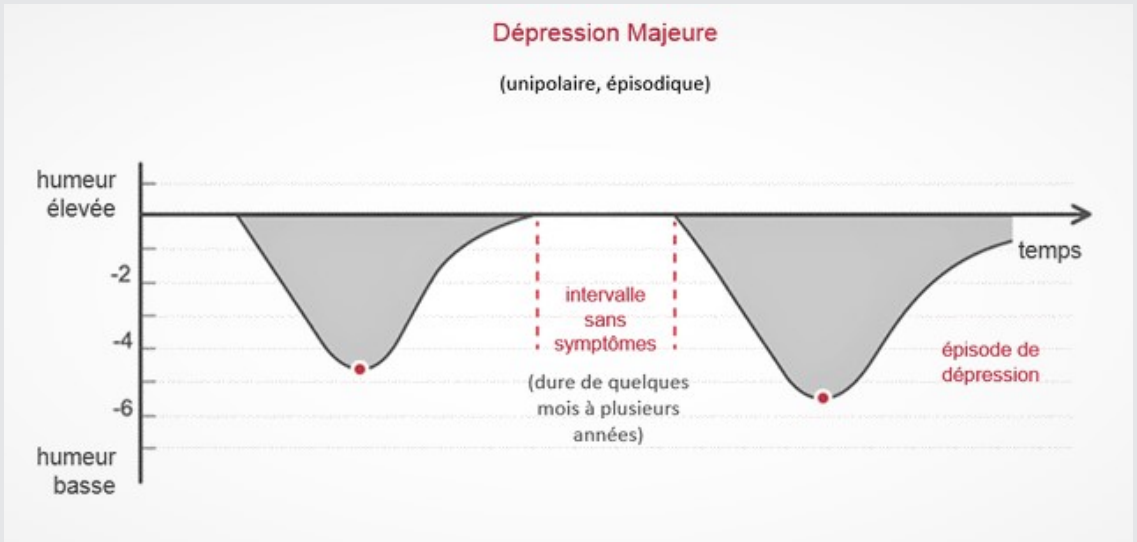
Comparison avec les méthodes existantes

Method	RMSE
Valstar et al. (2016) [27]	7.78
Yang et al. (2017)[36]	5.59^{DM}
Yang et al. (2017b)[34]	1.46^{DM}
Othmani et al. (2020)[15]	0.18^{*}
Rejaibi et al. (2019)[20]	0.17^{*}
MFCC-AU LSTM	0.15^{Norm} / 3.49

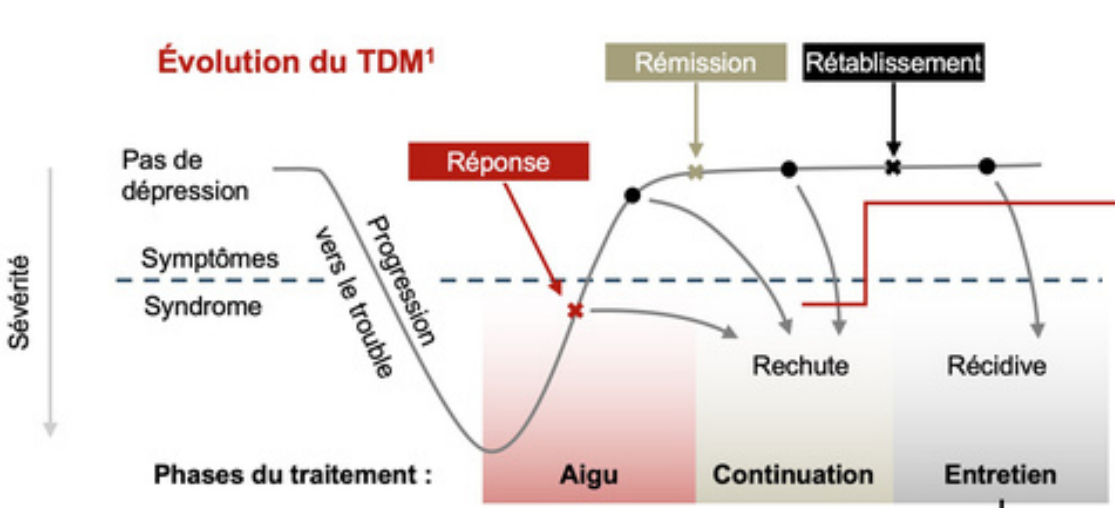
Comparison des performances de notre réseau le plus performant avec les approches existantes pour la prédiction des niveaux de gravité de la dépression sur l'ensemble de données DAIC-Woz. (^{DM}) : Homme déprimé, (^{Norm}): RMSE normalisé.

Episode Dépressif Vs. Dépression Majeure





Définitions de l'évolution clinique et des résultats du traitement



Rechute : épisode de trouble dépressif majeur qui survient dans les 6 mois suivant une réponse ou une rémission ;[2,3] théoriquement, un retour de l'épisode initial[4,5]

Récidive : épisode dépressif qui survient après 6 mois suite à une réponse ou une rémission ;[2,3] théoriquement, un nouvel épisode[4,5]

1. TDM = trouble dépressif majeur
2. Kupfer. J Clin Psychiatry 1991;52(Suppl 5):28–34;
3. Nierenberg et al. J Clin Psychiatry 2003; 64 (suppl 15): 13–17;
4. Riso et al. J Affect Disord 1997; 43 (2): 131–142;
5. Nierenberg & DeCecco. J Clin Psychiatry 2001; 62 (suppl 16): 5–9;
6. Frank et al. Arch Gen Psychiatry 1991; 48 (9): 851–855;

Lundbeck Institute Campus

Reconnaissance d'une rechute après une dépression à l'aide des données audiovisuelles

Motivations

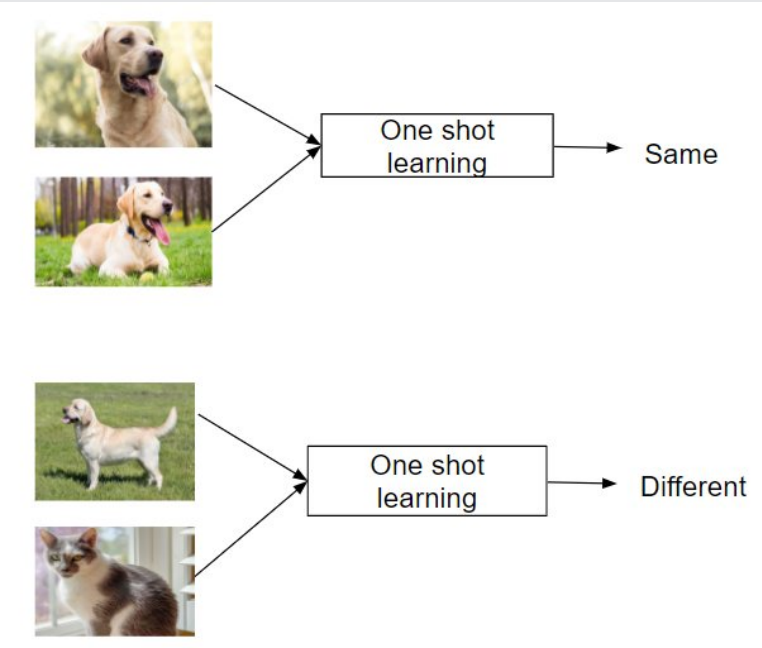
- Les épisodes dépressifs isolés sont assez rares : une première dépression annonce souvent une récidive. Cela se produit au cours des cinq prochaines années dans 50% à 80% des cas.

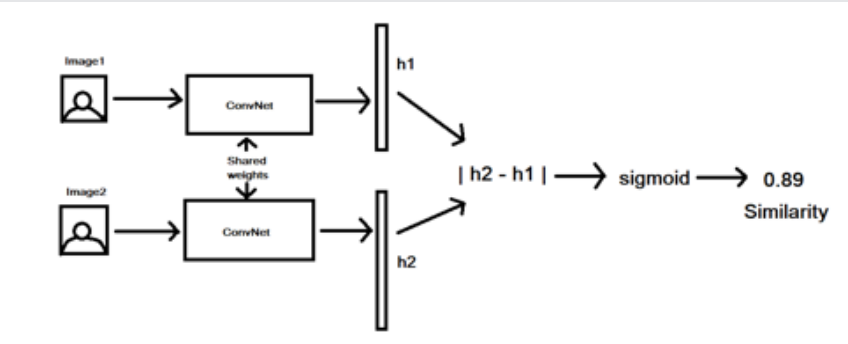
Objectifs

- Un système de surveillance prospective pour l'identification des signes de rechute après la dépression

Reconnaissance d'une rechute après une dépression à l'aide des données audiovisuelles

One-shot learning

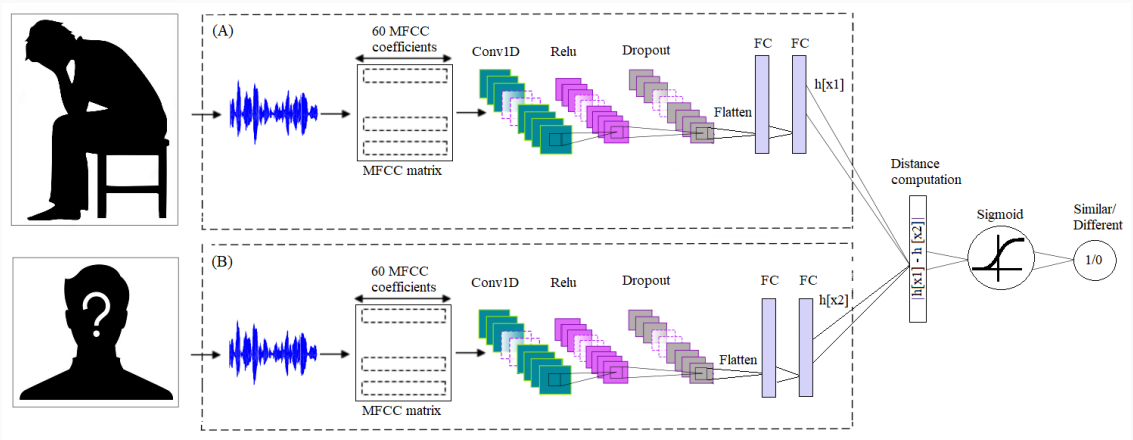




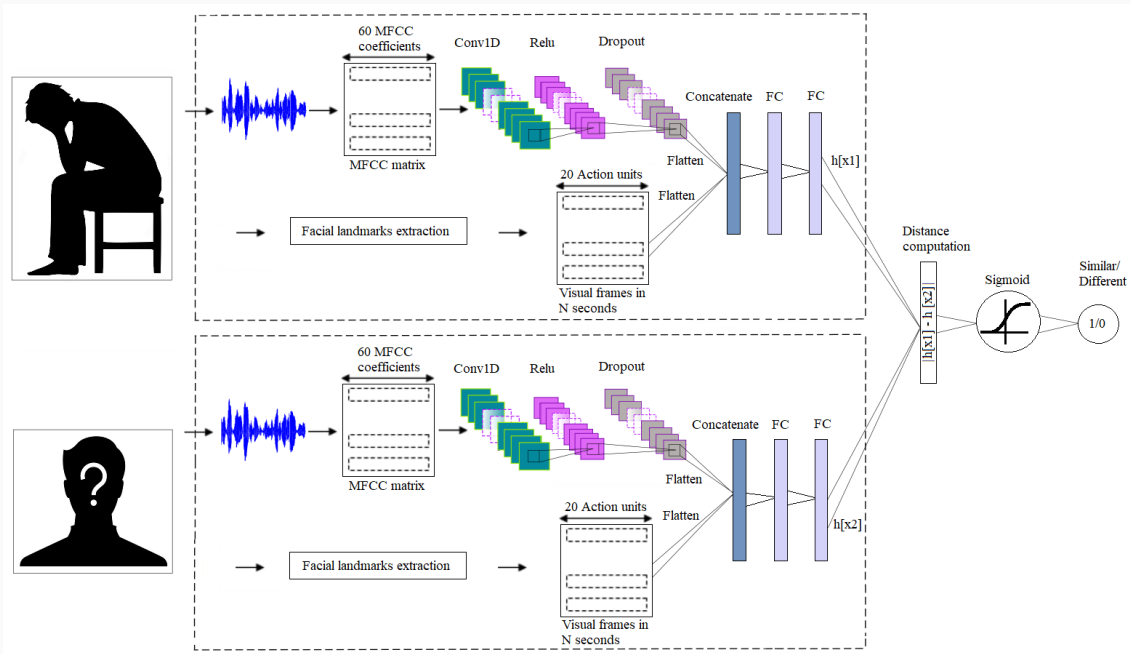
Comment définissons une rechute ?

- la similitude ou la proximité de l'encodage audio/visuel d'un sujet non diagnostiqué avec l'encodage audiovisuel d'un sujet en dépression
- la dissemblance ou l'éloignement de l'encodage audio/visuel du sujet non diagnostiqué et l'encodage audio/visuel d'un sujet sain.

Réseaux de type Siamois basés sur les données audio et audiovisuelles pour la détection des rechutes de dépression



(a)



(b)

Reconnaissance d'une rechute après une dépression à l'aide des données audiovisuelles

Réseau	Acc. (%)	RMSE	CC	CCC
MFCC Siamese	66.35	0.50	0.23	0.09
MFCC-AU Siamese	73.21	0.45	0.42	0.20

Performance des réseaux neuronaux siamois proposés pour la reconnaissance de similarité des états dépressifs en termes de taux de bonne classification (ou accuracy), d'erreur quadratique moyenne (RMSE), CC et CCC.

Réseau	Classification des échantillons			
	D-D	D-ND	ND-D	ND-ND
Actuel	10	252	674	128
Prédict	8	201	547	23
Percentage (%)	80	79.76	81.16	17.97

Performance du réseau neuronal siamois proposé avec les caractéristiques audiovisuelles pour la classification des rechutes de dépression (paires D-D).

Système de diagnostic assisté par ordinateur multimodal utilisant des vidéos biomédicales pour la prédiction des rechutes de dépression

- Système prospectif de monitorage, de suivi et/ou surveillance continue des patients atteints de trouble dépressif majeur.
- Développer un modèle de normalité de dépression pour la prédiction des rechutes de dépression à l'aide de données audiovisuelles.
- Un framework facilement extensible pour intégrer plus de modalités.
- Définir un nouveau biomarqueur pour la reconnaissance et l'évaluation de la dépression et la prédiction des rechutes de dépression sur la base des données audiovisuelles, qui soit facile à enrichir avec d'autres paramètres.

Système de diagnostic assisté par ordinateur multimodal utilisant des vidéos biomédicales pour la prédiction des rechutes de dépression

- Nous définissons la rechute de la dépression par la proximité des représentations ou des encodages audiovisuels d'un sujet en rémission avec **une réprésentation générique** des encodages audiovisuels des sujets déprimés.

$\longrightarrow$ **Modèle de normalité.**

- Un modèle de normalité est une approche basée le calcul d'une distance d'anomalie entre l'encodage d'un échantillon de test et une représentation apprise à partir des encodages audiovisuels des échantillons sans anomalie [1].

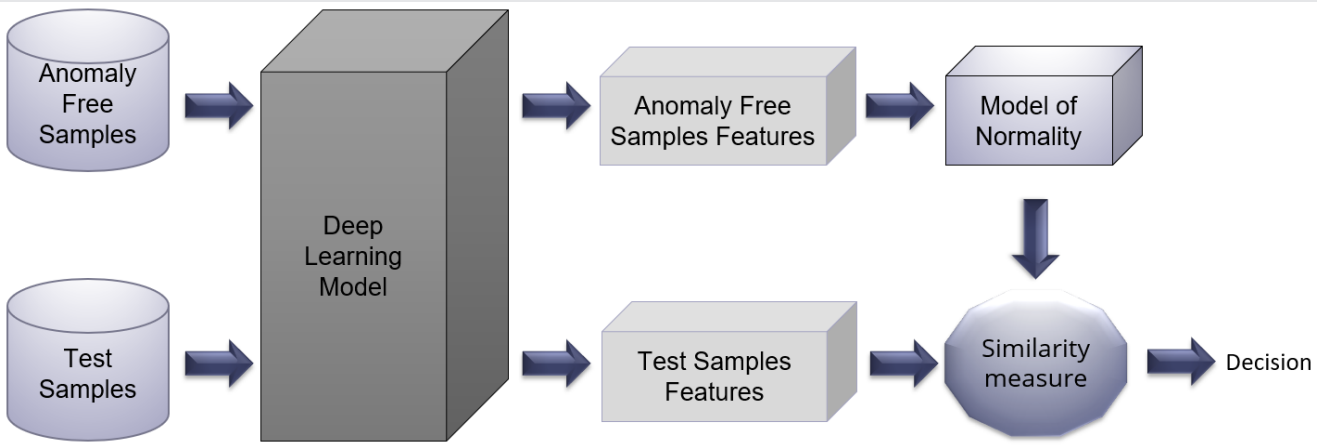
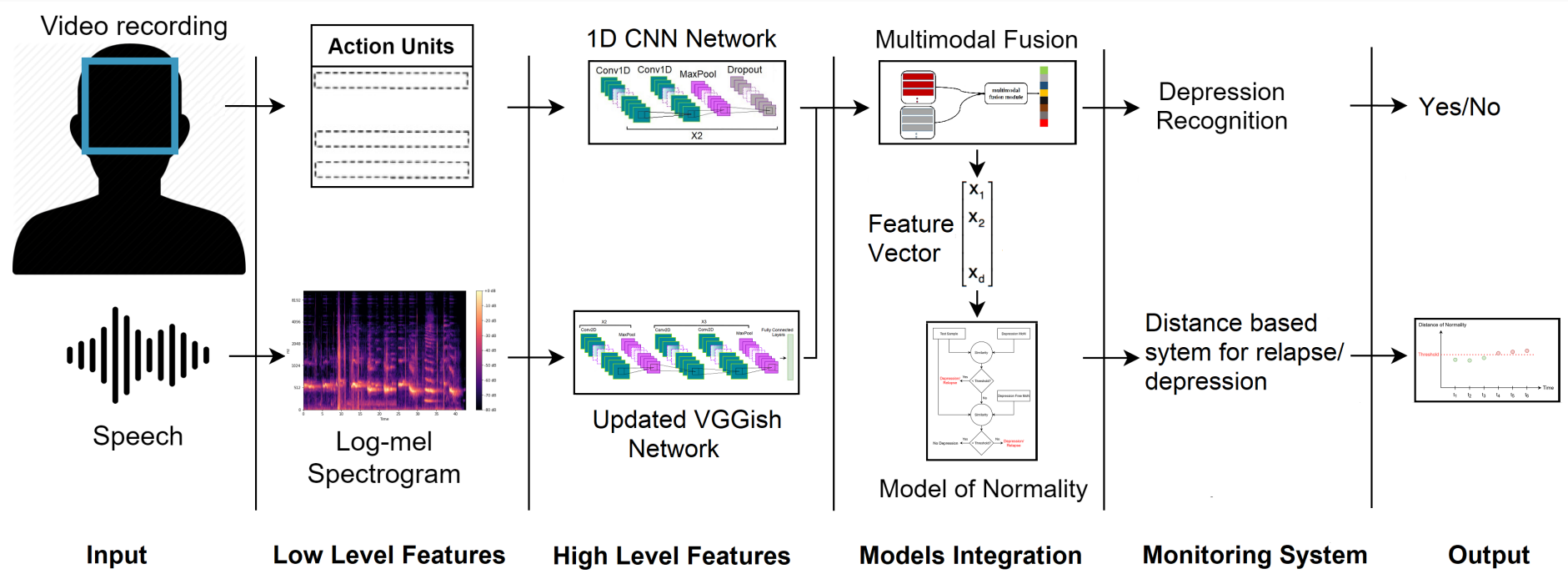


Diagramme d'un modèle de normalité **de base**.

Système de diagnostic assisté par ordinateur multimodal utilisant des vidéos biomédicales pour la prédiction des rechutes de dépression



Notre approche multimodale basée sur un modèle de normalité pour la reconnaissance de la dépression et la prédiction des rechutes de dépression. Tout d'abord, les enregistrements vidéo et la parole sont acquis et traités. Deuxièmement, les caractéristiques de bas et de haut niveau sont extraites. Troisièmement, une étape de fusion multimodale est effectuée, suivie d'une classification pour la reconnaissance de la dépression. Les caractéristiques extraites du module de fusion sont utilisées pour définir un modèle de normalité pour la reconnaissance de la dépression.

Nos deux permiers modèles de normalité MoN1 et MoN2

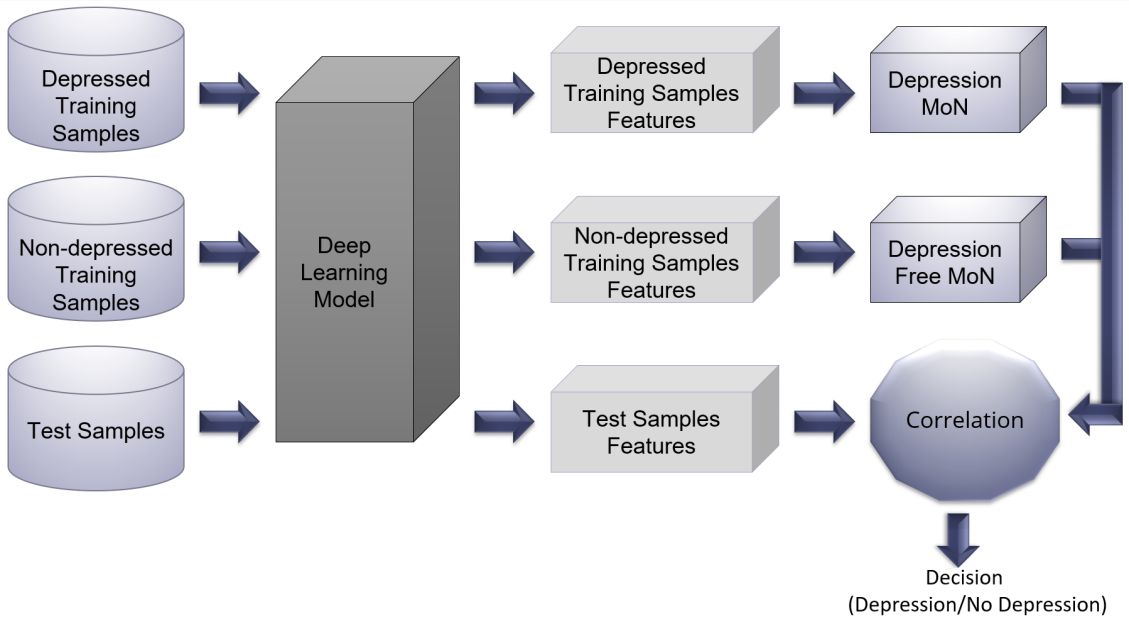
Algorithm 2 Our distance-based anomaly detection algorithm for depression relapse prediction. Only the samples labeled as minimal depression, and severe depression are considered. f_ϕ is the transformation function of the deep learning model that transfers the audio-visual inputs into its deep representation.

$(\mathbf{ND}, \mathbf{D}) \leftarrow$ Obtain the training set of non-depression samples (ND) and depression samples (D)
$\mathbf{S_{ND}} \leftarrow f_\phi(\mathbf{ND})$ ▷ Obtain deep feature representation non-depression samples
$\mathbf{S_D} \leftarrow f_\phi(\mathbf{D})$ ▷ Obtain deep feature representation depression samples
$\mathbf{M_{ND}} \leftarrow \frac{1}{|\mathbf{S_{ND}}|} \sum_{x_i \in \mathbf{S_{ND}}} x_i$ ▷ Deep representation of non-depression class
$\mathbf{M_D} \leftarrow \frac{1}{|\mathbf{S_D}|} \sum_{x_i \in \mathbf{S_D}} x_i$ ▷ Deep representation of depression class
$\mathbf{T} \leftarrow$ Obtain the set of test samples

for *each sample* t *in* $\mathbf{T}$ **do**
  if $Corr(t, \mathbf{M_{ND}}) > Corr(t, \mathbf{M_D})$ **then**
    $y_t \leftarrow$ non-depression
  else
    $y_t \leftarrow$ depression
  end
end

$$Corr(X, M_D) = \frac{\text{cov}(X, M_D)}{\sigma_X \sigma_{M_D}}$$

Nos deux permiers modèles de normalité MoN1 et MoN2



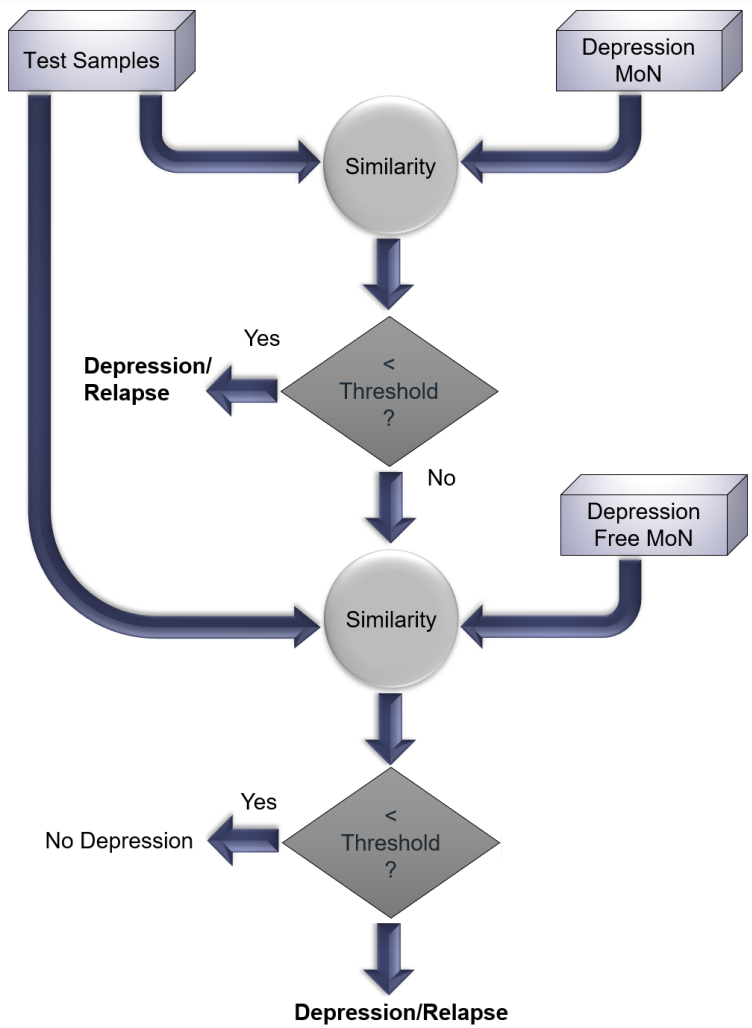
MoN1

- les échantillons avec des scores PhQ $<$ 11 sont utilisés pour construire le Depression Free MoN
- les échantillons avec des scores PhQ $>$ 11 sont utilisés pour construire le MoN de la dépression.

MoN2

- seuls les échantillons avec des scores PHQ compris entre 0 et 4 sont utilisés pour créer un modèle sans dépression
- les échantillons avec des scores supérieurs ou égaux à 20 sont utilisés pour créer le modèle de dépression.

Notre troisième modèle de normalité MoN3



Processus de vérification en deux étapes utilisé dans le MoN3.

Algorithm 3 Our Model Of Normality MoN3 algorithm for depression relapse prediction. Only the samples labeled as minimal depression, and severe depression are considered in the creation of the MoN. f_ϕ is the transformation function of the deep learning model that transfers the audio-visual inputs into their deep representation. f_{th} is the function that sets the thresholds following table II.

$(\mathrm{ND}, \mathrm{D}) \leftarrow$ Obtain the training set of non-depression samples (ND) and depression samples (D)
$\mathrm{S_{ND}} \leftarrow f_\phi(\mathrm{ND})$ ▷ Obtain deep feature representation non-depression samples
$\mathrm{S_D} \leftarrow f_\phi(\mathrm{D})$ ▷ Obtain deep feature representation depression samples
$\mathrm{M_{ND}} \leftarrow \frac{1}{|\mathrm{S_{ND}}|} \sum_{x_i \in \mathrm{S_{ND}}} x_i$ ▷ Model of normality of non-depression class
$\mathrm{M_D} \leftarrow \frac{1}{|\mathrm{S_D}|} \sum_{x_i \in \mathrm{S_D}} x_i$ ▷ Model of normality of depression class
$\mathrm{D_{ND}} \leftarrow Dist(\mathrm{M_{ND}}, \mathrm{S_{ND}})$ ▷ Calculate distances between samples and MoN
$\mathrm{D_D} \leftarrow Dist(\mathrm{M_D}, \mathrm{S_D})$ ▷ Calculate distances between samples and MoN
$(\tau_\mathrm{D}, \tau_\mathrm{ND}) \leftarrow f_{th}(\mathrm{D_D}, \mathrm{D_{ND}})$ ▷ Setting thresholds
$\mathrm{T} \leftarrow$ Obtain the set of test samples

for *each sample* t *in* T **do**
  if $dist(t, \mathrm{M_D}) < \tau_\mathrm{D}$ **then**
    $y_t \leftarrow$ depression
  else
    if $dist(t, \mathrm{M_{ND}}) < \tau_\mathrm{ND}$ **then**
      $y_t \leftarrow$ non-depression
    else
      $y_t \leftarrow$ depression
    end
  end
end

Notre troisième modèle de normalité MoN3

MoN3

- Seuls les échantillons avec des scores PHQ compris entre 0 et 4 sont utilisés pour créer un modèle sans dépression
- Les échantillons avec des scores supérieurs ou égaux à 20 sont utilisés pour créer le modèle de dépression.

Les seuils

Quatre seuils sont définis en calculant les distances euclidiennes entre le modèle de normalité et les échantillons utilisés pour le créer.

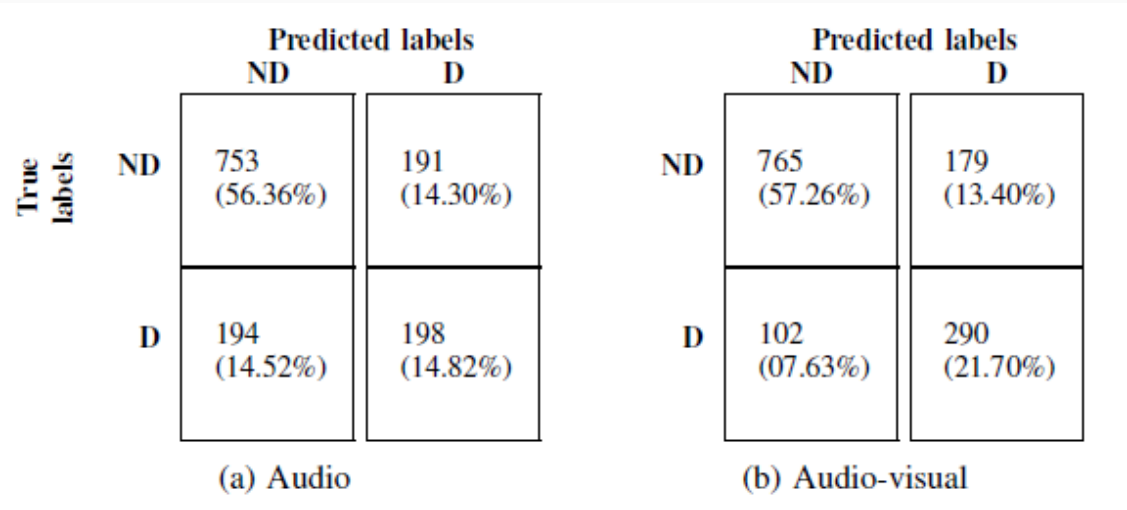
Seuil	D = liste des distances
T11	$Max(D)$
T12	$Max(D) - std(D)$
T13	$Mean(D)$
T14	$Mean(D) + std(D)$

Les seuils sont calculés pour les deux modèles de normalité du test MoN3.

Résultats expérimentaux

Résultats obtenus par le modèle d'apprentissage proposé pour la détection de la dépression. Deux classes: Dépression (D) and Non-Dépression (ND).

Réseau	Acc(%)	Précision		Rappel		F1-score		
		D	ND	D	ND	D	ND	AVG
Audio	71.18	0.509	0.795	0.505	0.798	0.507	0.796	0.652
Audio/Visuel	78.97	0.618	0.882	0.740	0.811	0.674	0.845	0.759



Matrices de confusion des différentes expériences. (A) représente la matrice de confusion du premier modèle basé uniquement sur les données audio. (B) représente la matrice de confusion du modèle basé sur les données audiovisuelles.

Résultats expérimentaux

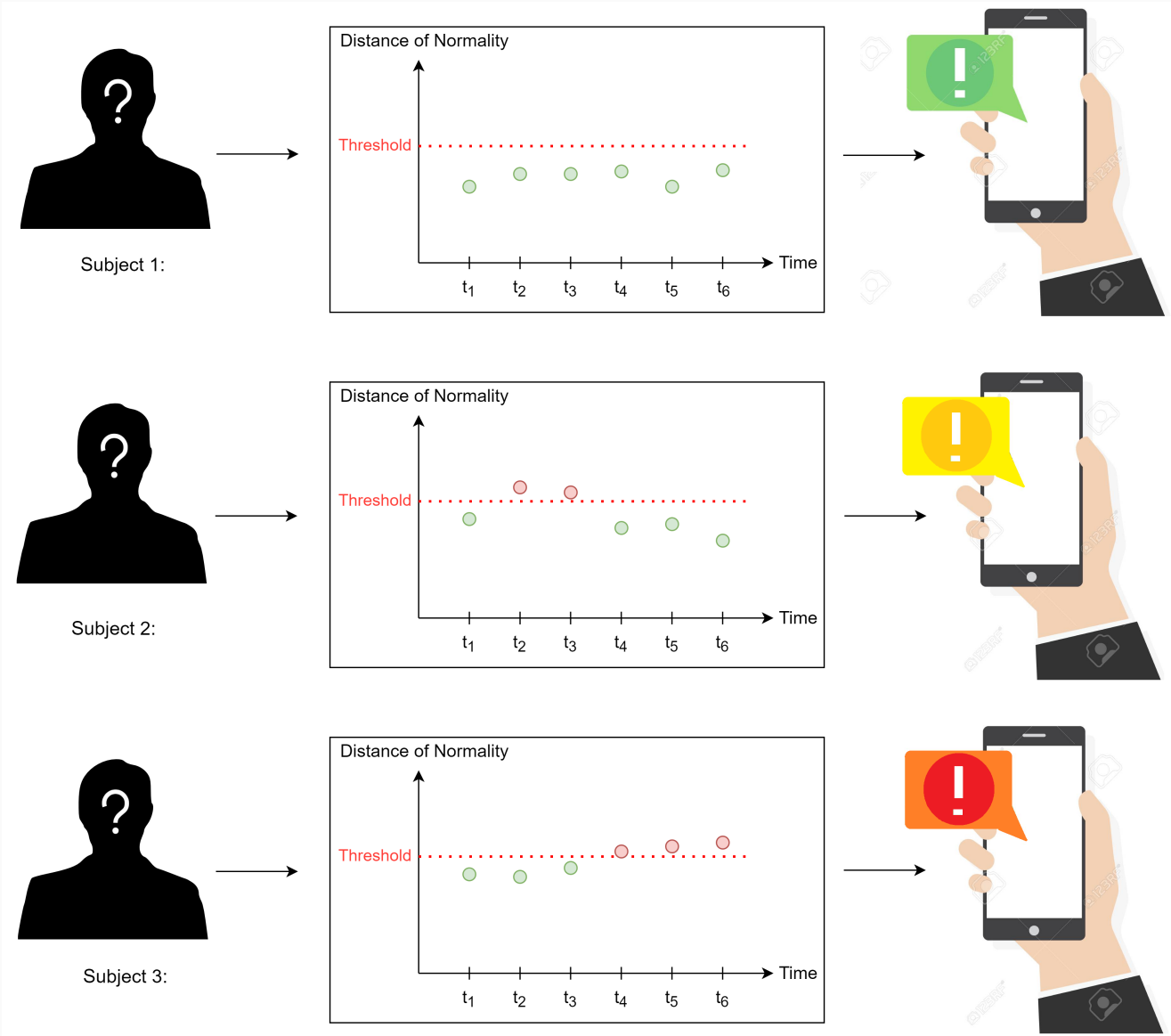
Modèle de Normalité	Seuil	Acc(%)	F1 score
MON1	—	77.84%	75.04%
MON2	—	80.99%	78.28%
MON3	T11	76.83%	73.1%
MON3	T12	78.12%	75.67%
MON3	T13	77.52%	74.22%
MON3	T14	**81.89%**	**79.4%**

La Précision et F1-score obtenus en utilisant le modèle de normalité pour la prédiction des rechutes après une dépression.

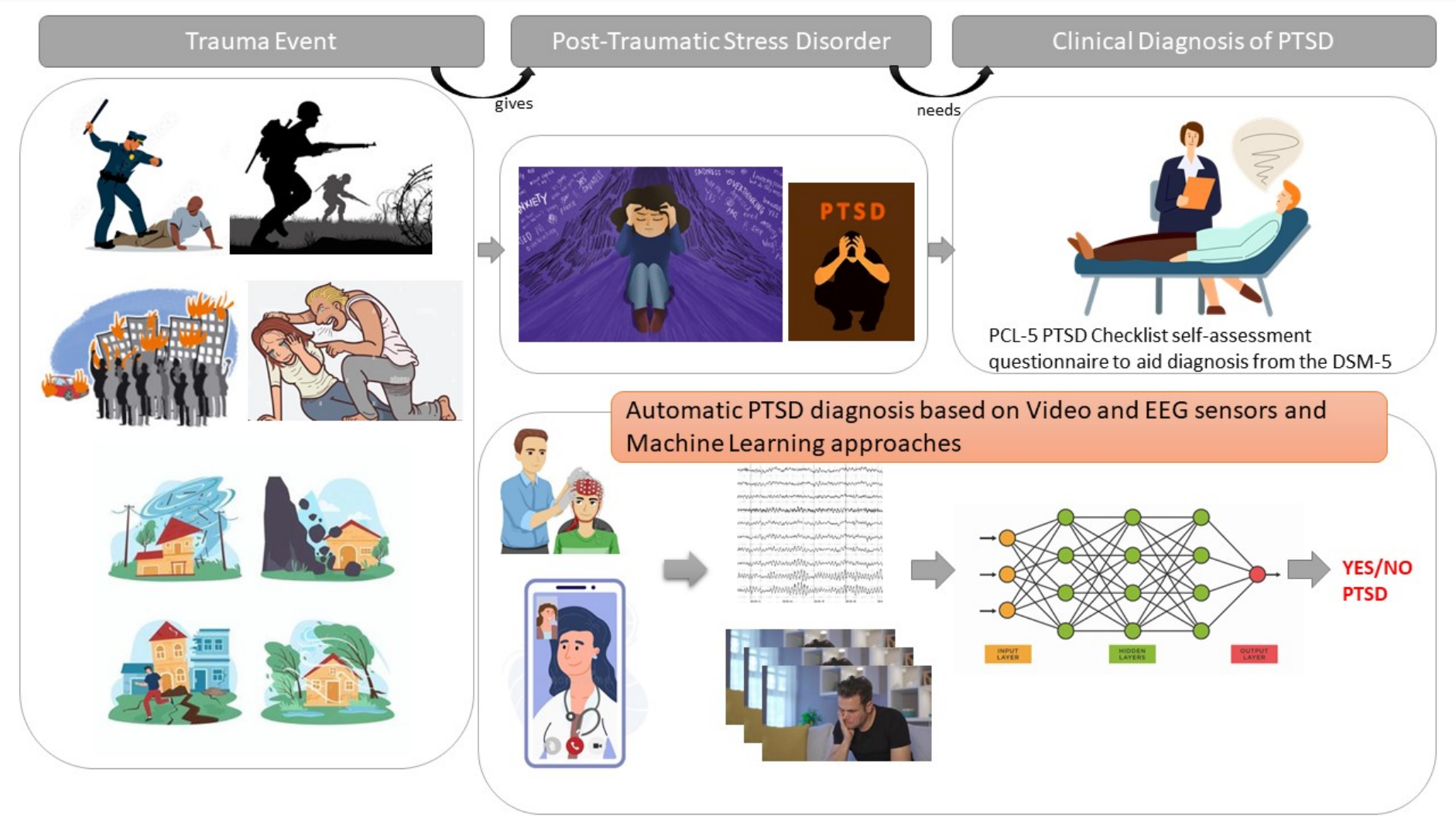
Approche	**Tâche de prédiction**	**Stratégie**	**Modalité**	**Acc(%)**	**F1-score**
DepAudioNet [10]	Depression	Train/Test split	Audio	—	ND=70%, Avg=52%
BLSTM-MIL [22]	Depression	Leave-one-out	Audio	**96.7%**	Avg=85.4%
MFCC-based RNN [19]	Depression	Train/Test split	Audio	76.27%	ND=85%, D=46%
Ensemble 1d-CNN[28]	Depression	Train/Test split	Audio	72%	Avg=63%
EmoAudioNet [17]	Depression	Train/Test split	Audio	74.13%	ND=82%, D=49%
AudVowelConsNet [13]	Depression	Train/Test split	Audio	**86.06%**	ND=90%, D=77%
Our Audio based Network	Depression	Train/Test split	Audio	71.18%	Avg=65.20%
Our Audio based Network	Depression	Leave-one-out	Audio	94.12%	**Avg=86.1%**
Yang et al.[33]	Depression	Train/Test split	Audio / Visual / Text	—	ND=87.7%, D=57.1%
Nasir et al.[14]	Depression	Train/Test split	Audio/Visual	—	**ND=89%**, D=63%
Our Audio-visual based Network	Depression	Train/Test split	Audio/Visual	**78.97%**	ND=84.5%, **D=67.4%**
SiameseNet [**muzammel2021**]	Relapse	Train/test split	Audio/Visual	80.16%	—
Our Model of normality	Depression / Relapse	Train/Test split	Audio/Visual	**81.89%**	**Avg=79.4%**

Comparaison entre nos réseaux basé sur les données audio, les données audiovisuelles et notre modèle de normalité le plus performant avec des méthodes de pointe pour la détection de la dépression et la prédiction des rechutes de dépression en termes de précision et de scores F1.

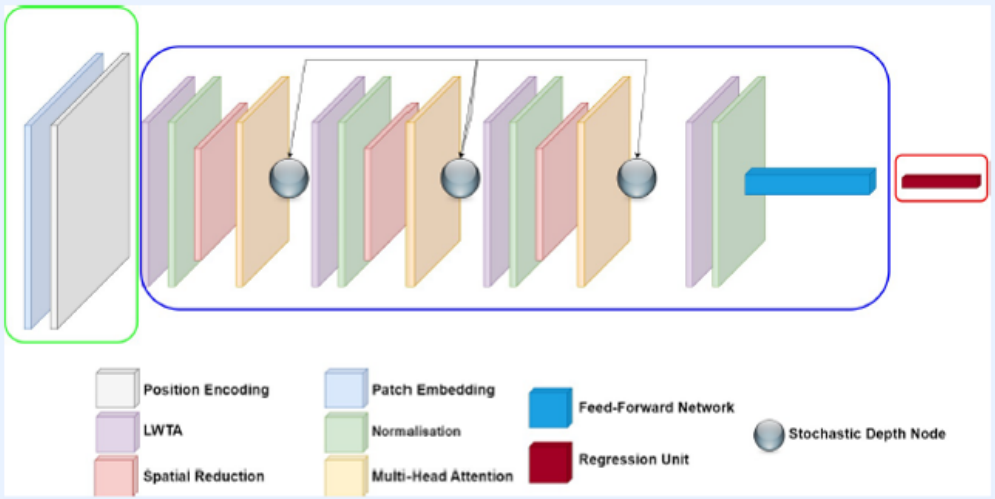
Notre système de surveillance prospective pour la prédiction de la rechute de la dépression: un cas d'étude



Le trouble de stress post-traumatique

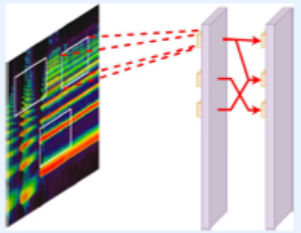


A Novel Stochastic Transformer-based Approach for Post-Traumatic Stress Disorder Detection using Audio recording of Clinical Interviews



Architecture générale de notre Stochastic Transformer.

Locally Connected Layers



Spatial Reduction

$$SR(x) = Norm(Reshape(x, Ri)W) \quad (2)$$

Local-Winner-Take-All layers (LWTA)

$$y = \begin{cases} y_i; & if \quad y_i = max(0, x.W) \\ 0 & otherwise \end{cases} \quad (1)$$

GeLU activation function

$$GELU(x) = x.\Phi(x) = x\frac{1}{2}[1 + erf(\frac{x}{\sqrt{2}})] \quad (3)$$

$\Phi(x)$ is the Cumulative Distribution Function.

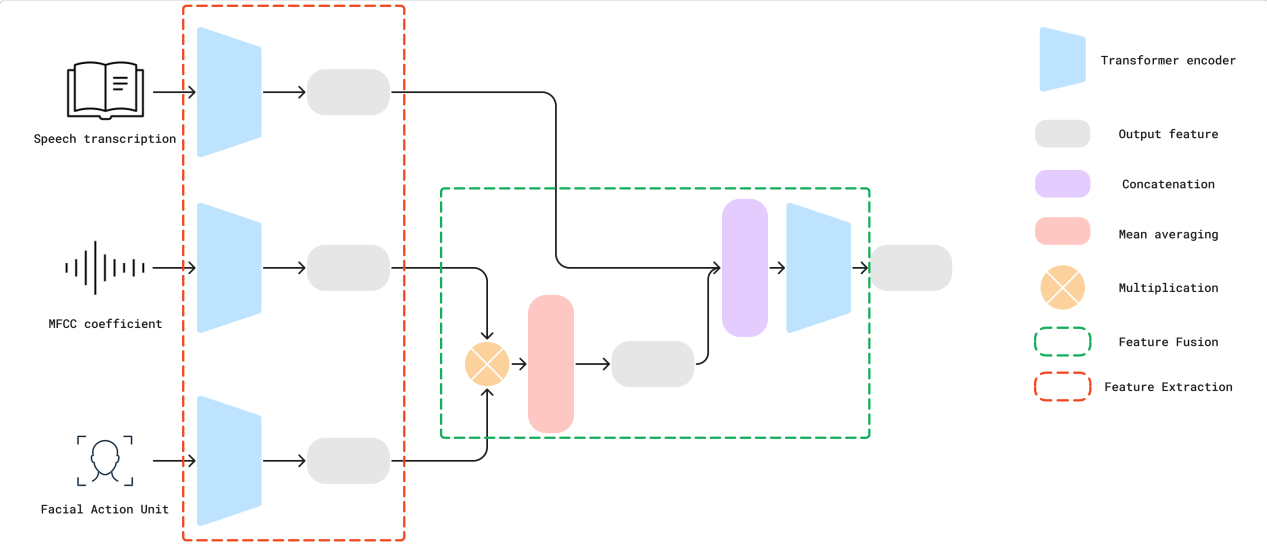
Dia, M., Khodabandelou, G., & **Othmani, A.** (2023). A novel stochastic transformer-based approach for post-traumatic stress disorder detection using audio recording of clinical interviews. In 2023 IEEE 36th International Symposium on Computer-Based

Système d'aide au diagnostic du stress post-traumatique par approche multimodale

- Une approche multimodale pour la détection du TSPT utilisant l'audio, les traits du visage et la transcription textuelle.
- 4 Transformateurs utilisant des composants stochastiques (couche d'apprentissage profond et fonctions d'activation) pour la détection du TSPT par vidéo.
- Un modèle de fusion combinant fusion de caractéristiques et fusion de décisions pour la détection multimodale audiovisuelle et textuelle du TSPT.
- Une approche basée sur l'équilibre pour l'extraction de caractéristiques visuelles sur les AU pour la reconnaissance des schémas du TSPT.

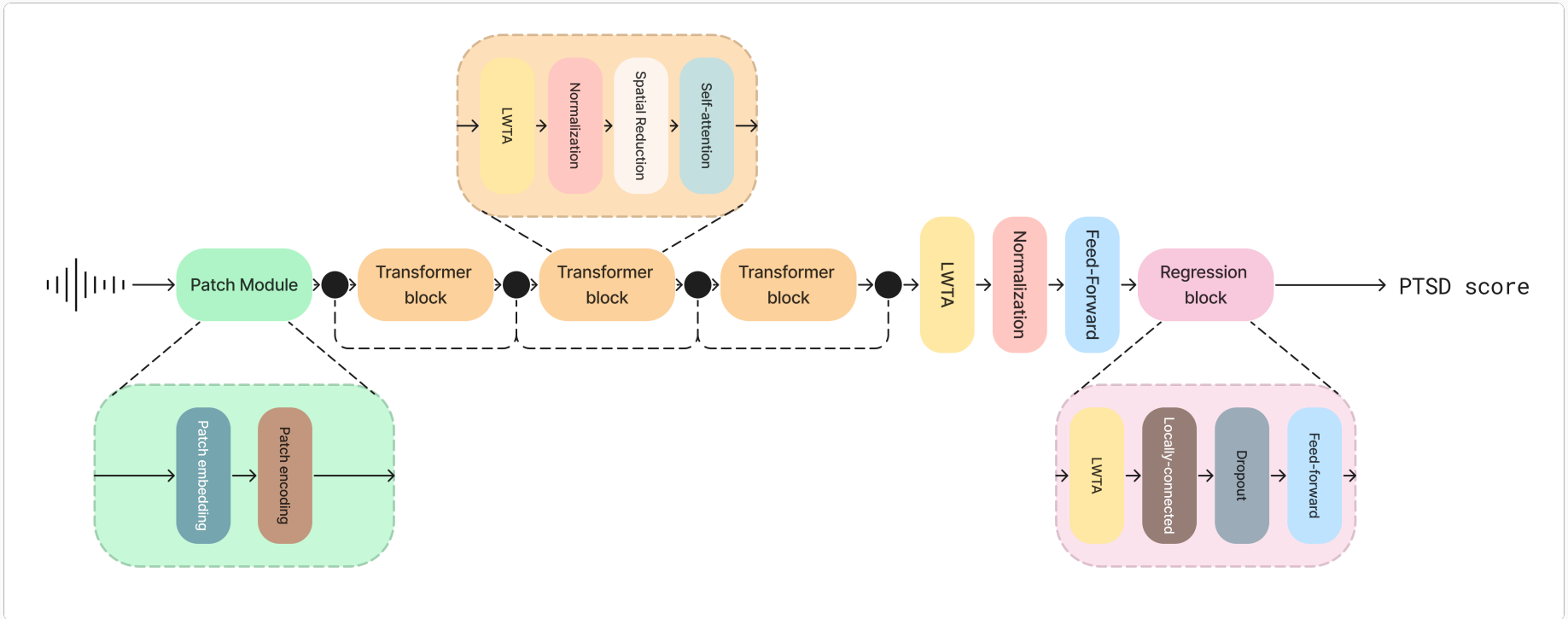
Dia, M., Khodabandelou, G., & Othmani, A. (2024). Paying attention to uncertainty: A stochastic multimodal transformers for post-traumatic stress disorder detection using video. Computer Methods and Programs in Biomedicine, 257, 108439.

Paying attention to uncertainty: A stochastic multimodal transformers for post-traumatic stress disorder detection using video



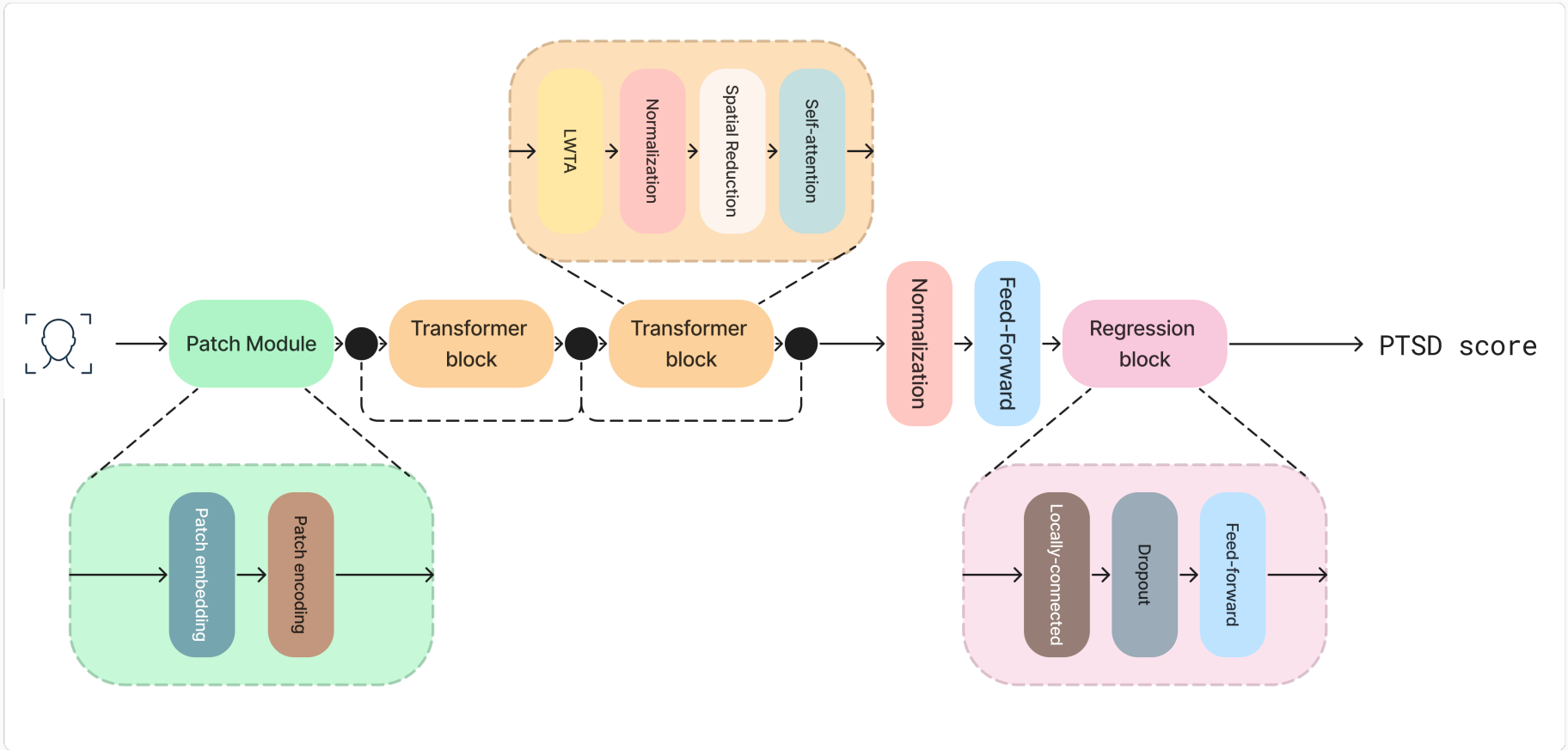
Vue d'ensemble de notre Transformer stochastique multimodal proposé. Les entrées sont les coefficients MFCC, la matrice des unités d'action (AU), et la transcription audio-texte. Chaque modalité est extraite et encodée à l'aide d'un encodeur Transformer dédié. Les caractéristiques ainsi obtenues sont ensuite concaténées et envoyées à un encodeur de fusion, qui combine une couche LSTM et un Transformer. La sortie finale est le score prédit de stress post-traumatique.

Paying attention to uncertainty: A stochastic multimodal transformers for post-traumatic stress disorder detection using video



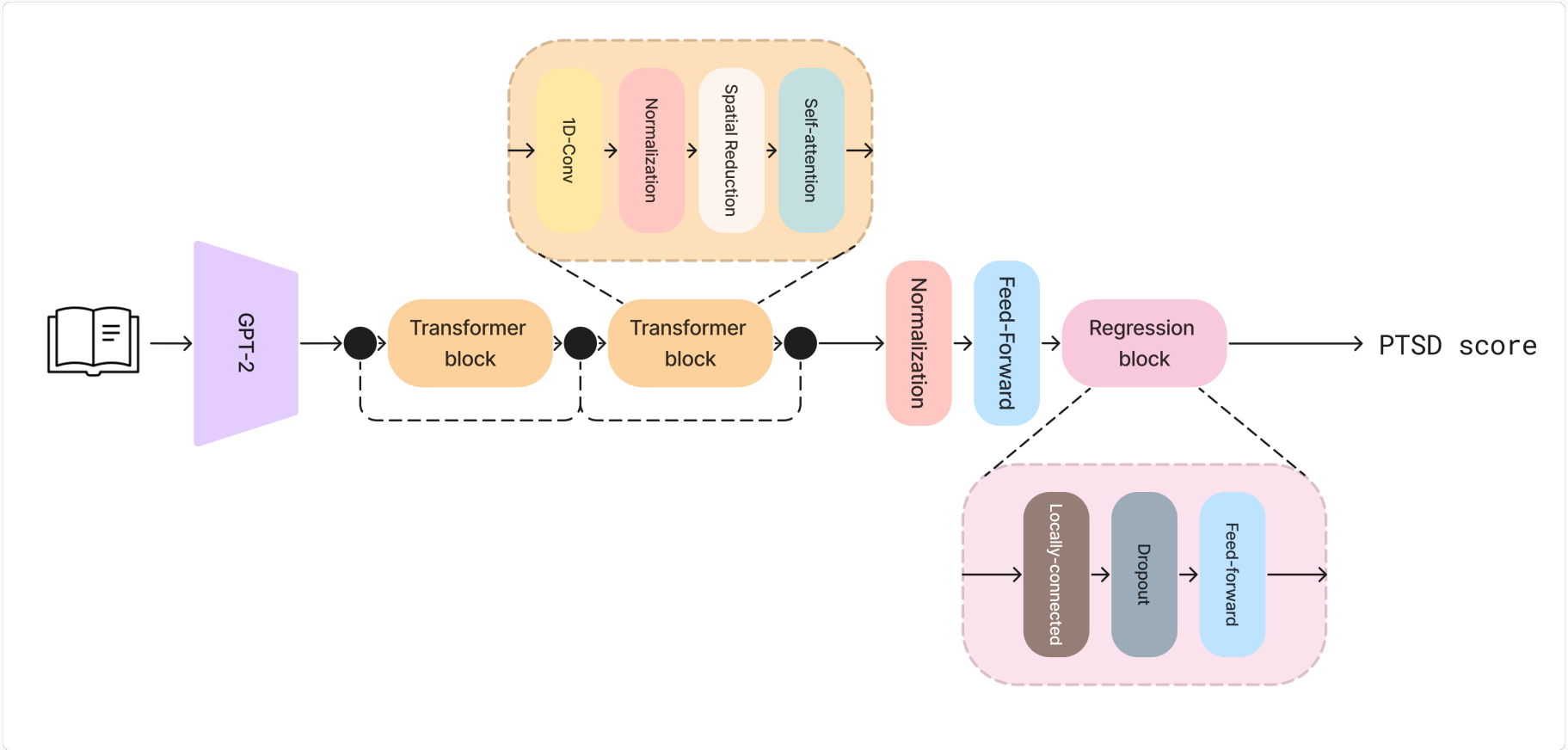
Architecture de notre encodeur audio. Cet encodeur est un transformer stochastique. Il est composé d'un module Patch, de 3 blocs Transformer reliés par des nœuds à profondeur stochastique, de couches LWTA, d'un réseau feed-forward, et d'un bloc de régression.

Paying attention to uncertainty: A stochastic multimodal transformers for post-traumatic stress disorder detection using video



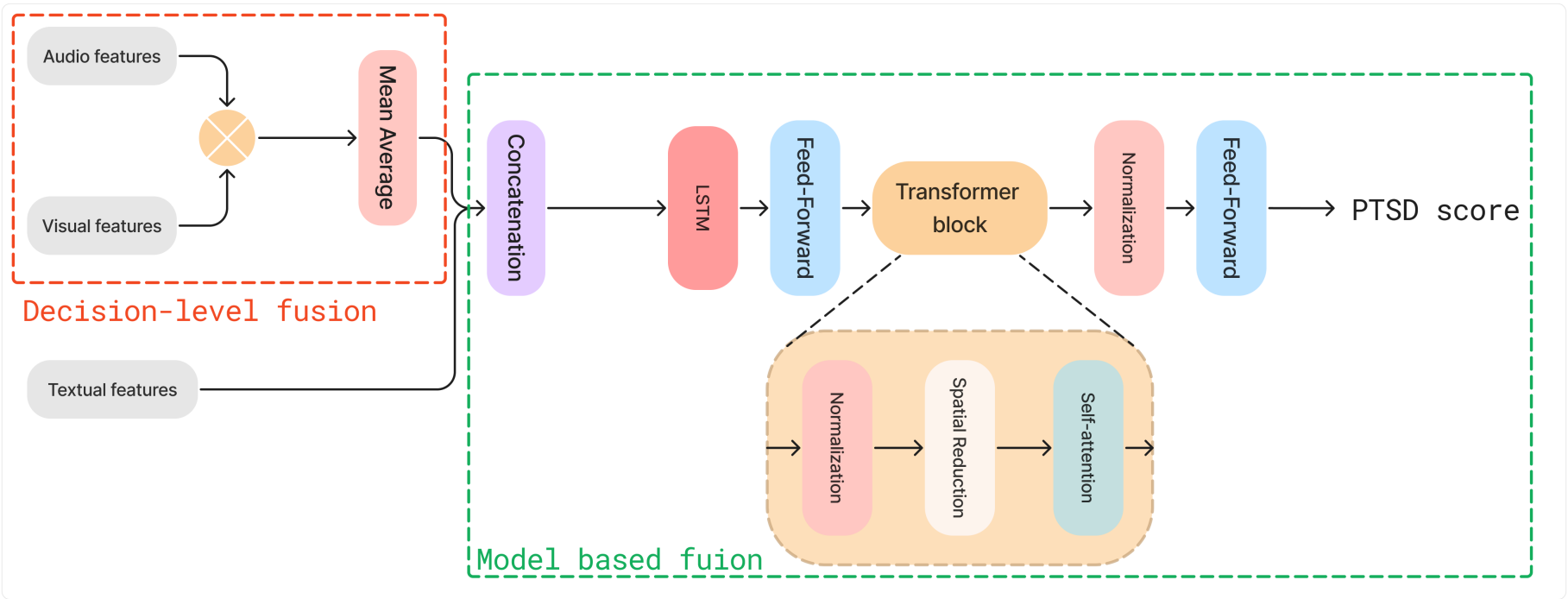
Présentation de notre encodeur basé sur les AU. Ce transformer a été préalablement entraîné. Il s'agit d'un transformer stochastique. Contrairement à l'encodeur audio, celui-ci contient des couches LWTA uniquement sur les blocs de transformer, et non sur le bloc de régression.

Paying attention to uncertainty: A stochastic multimodal transformers for post-traumatic stress disorder detection using video



Architecture de notre encodeur de texte. Ce Transformer joue le rôle d'encodeur. Les caractéristiques sont extraites via GPT2, puis traitées via deux blocs Transformer avec attention temporelle. De plus, cet encodeur ne possède pas de couche LWTA..

Paying attention to uncertainty: A stochastic multimodal transformers for post



Architecture de notre modèle de fusion. La fusion décisionnelle entre les données audio et visuelles est réalisée par une multiplication élément par élément et une opération de moyenne. La fusion des caractéristiques est obtenue par concaténation. Ce tenseur fusionné est ensuite transmis au modèle composé d'un LSTM, de couches entièrement connectées et d'un transformer pour prédire le score de TSPT.

EEG et aide au diagnostic et compréhension des troubles mentaux

EEG pour la detection des troubles mentaux

- Un pipeline de prétraitement qui réduit le bruit et transforme le signal EEG en une représentation temps-fréquence 2D grâce à la transformée en ondelettes.
- Un nouvel encodeur de transformateur convolutif multicanal (MCT) qui extrait efficacement les caractéristiques de chaque canal EEG, sans nécessiter un jeu de données volumineux.
- Une méthode de fusion permettant de fusionner les informations de tous les canaux tout en préservant les informations.
- Une fonction de perte basée sur l'entropie qui améliore la précision des prédictions du modèle.

Dia, M., Khodabandelou, G., Anwar, S. M., & Othmani, A. (2025). Multichannel convolutional transformer for detecting mental disorders using electroancephalogrpahy records. Scientific Reports, 15(1), 15387.

Stratification des troubles neurologiques par analyse automatique d'EEG

- Combinaison de paradigmes EEG actifs (testant spécifiquement la mémoire/apprentissage) et passifs pour une cartographie multidimensionnelle des altérations cérébrales liées à la dépression.
- Précision diagnostique accrue grâce à l'IA, notamment via les tâches actives
- Identification de sous-types cliniques basés sur des dysfonctionnements cognitifs mesurables (ex. : atteinte de la mémoire)
- Biomarqueurs EEG objectifs révélant l'hétérogénéité neurobiologique de la dépression
- Passage d'un diagnostic catégoriel à une médecine personnalisée : les sous-groupes identifiés ouvrent la voie à des thérapies ciblées adaptées aux déficits cognitifs spécifiques de chaque patient.

Yasin, S., Othmani, A., Mohamed, B., Raza, I., & Hussain, S. A. (2024). Depression detection and subgrouping by using the active and passive EEG paradigms. Multimedia Tools and Applications, 1-24.

Autism Spectrum Disorder (ASD) ou Trouble du spectre de l'autisme (TSA)

- Le TSA est défini comme un trouble neurodéveloppemental qui se caractérise par un manque d'interaction sociale et d'intelligence émotionnelle, ainsi que par un comportement répétitif, odieux, stigmatisé et figé.
- Environ 1 enfant sur 44 aux États-Unis est atteint de TSA, tandis qu'une étude mondiale estime la prévalence à environ 1 enfant sur 100.

Motivations

Motivations	Objectives
Autism is not a disease	An Artificial Intelligence based platform for longitudinal monitoring and follow up of ASD children based on an automated approach for ASD's severity level assessment : • AudioVisual Emotion assessment • Behavior Assessment
Autism needs no cure but requires longitudinal monitoring and assistance (growth devel-	

State of the art

Computer vision-based approaches have also emerged as promising tools for the recognition and assessment of Autism Spectrum Disorder (ASD) [**DIA2024105712**, **khor2024autism**, **CanAutism**, **CvInAutism**].

- Tang et al. [**FaceRecoAndVisual**] highlighted the potential of utilizing behavioral cues to automatically identify severe levels of Autism Spectrum Disorder (ASD).
- Nag et al. [**TowardsContinuousSocial**] conducted a study measuring eye gaze patterns and emotion recognition in children with ASD, comparing them to neurotypical controls.
- Li et al. [**twoStageMultiModal**] conducted a study proposing a framework for assessing children's affect states in play therapy settings using multi-modal emotion signals.

Contributions

- A novel approach for Visual Affect Recognition **(VAR) designed for children with Autism Spectrum Disorder (ASD) in a continuous domain**.
- An extension of an existing **video dataset** of children with ASD with a labeling of the intensity of the emotional or affective state in the continuous affective domain.
- A new optimization technique called **Stochastic Average Gradient Augmented with Tracking (SAGAT)** is used to train deep neural networks with less storing memory requirements and high efficiency.

Our proposed approach

- **Objective:** Predict the affect (valence and arousal) of children with autism using a deep learning approach.
- **Model:** Uses a deep Convolutional Neural Network (CNN).
- **Input Data:** Video frames are fed into the CNN for affect prediction.
- **Optimization Algorithm:** Introduces a novel variation of the Stochastic Average Gradient with Averaging (SAGA).
- **Proposed Algorithm:** Named Stochastic Average Gradient Augmented with Tracking (SAGAT).
- **Key Improvements:**
  - Faster Convergence: Reduces the number of iterations needed.
  - Lower Bias: Enhances optimization accuracy.
  - Reduced Memory Storage: Optimizes computational efficiency.

The CNN architecture

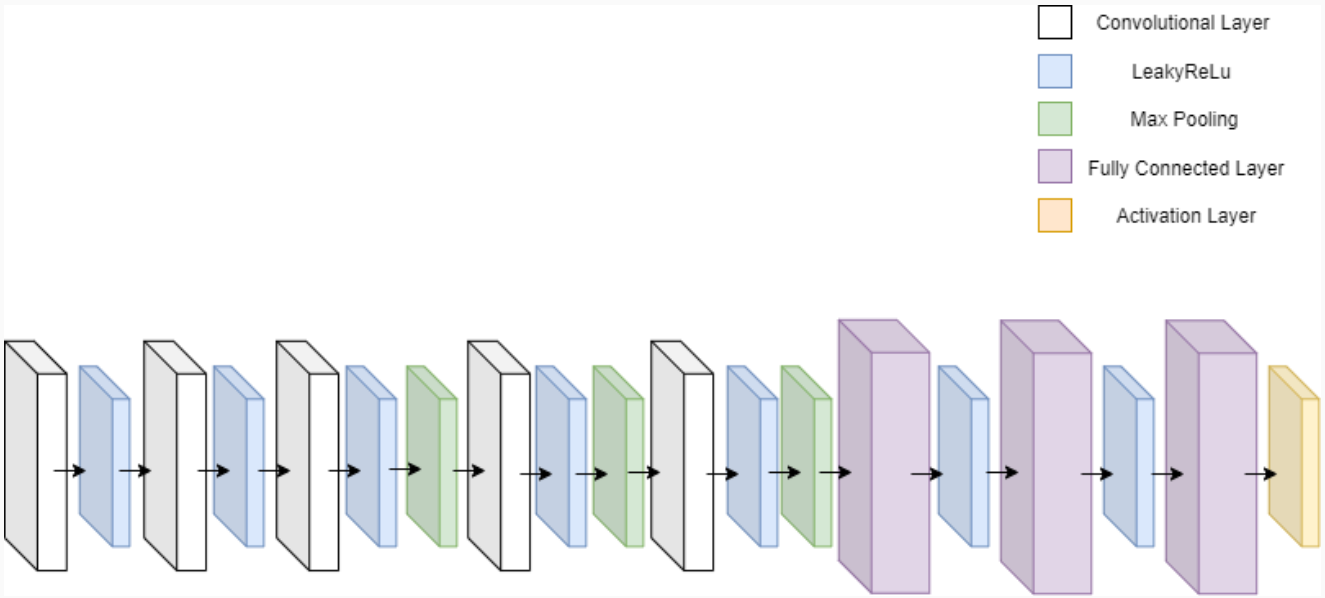


Figure 1: Our proposed CNN model for Affect recognition in children with Autism.

$$F_{LeakyReLu} = \begin{cases} x & \text{if } x > 0 \\ 0.01x & \text{else.} \end{cases} \tag{5}$$

Stochastic Average Gradient with Averaging (SAGA)

Algorithm 1 SAGA Optimization Algorithm

Initialize: $w_0 \in \mathbb{R}^d$, gradient memory $G_i = \nabla f_i(w_0), \forall i$ **for** $k \in \{1, \ldots, T\}$ **do**

  Pick i_k uniformly at random from $\{1, \ldots, n\}$

  $w_{k+1} = w_k - \alpha \left(\nabla f_{i_k}(w_k) - G_{i_k} + \frac{1}{n} \sum_{i=1}^{n} G_i \right)$ **for** $i \in \{1, \ldots, n\}$ **do**

    $i = i_k$

  $G_i = \nabla f_i(w_k)$ **else**

    Do not update G_i

Return w_{k+1}

SAGAT Optimization Algorithm

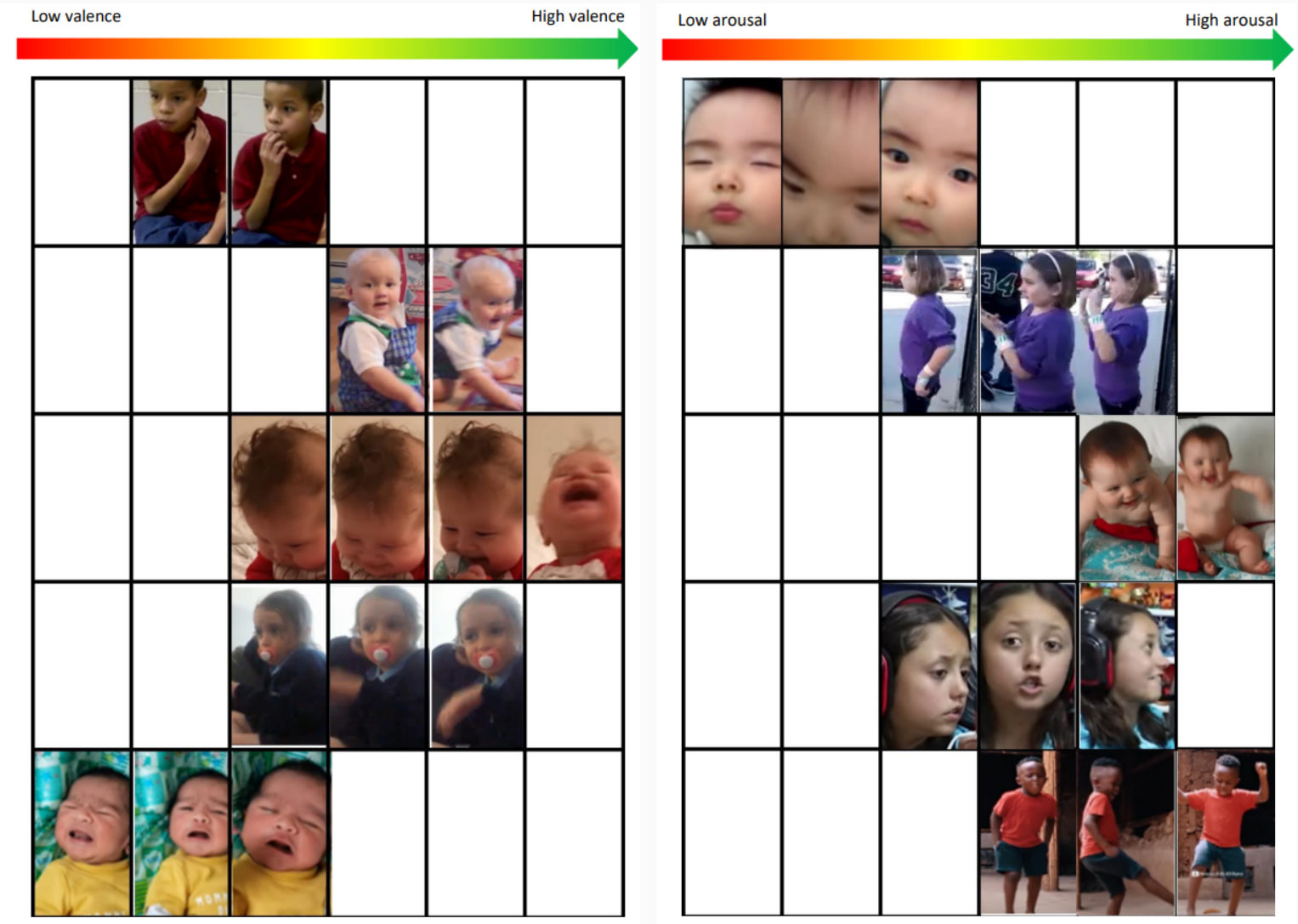
Algorithm 2 Stochastic Average Gradient Augmented with Tracking

Initialize: $w_0 \in \mathbb{R}^d$
Initialize: LastUpdate $= [0, \ldots, 0]$, an array of size n
Initialize: $\forall i \in \{1, \ldots, n\}, G_i = \nabla f_i(w_0)$
for $k \in \{1, \ldots, T\}$ **do**
Pick $I_k \subset \{1, \ldots, n\}$ uniformly at random such that $|I_k| = \text{card}$
$cur\_G_k = \frac{1}{\text{card}} \sum_{i \in I_k} \nabla f_i(w_k)$
$old\_G_k = \frac{1}{\text{card}} \sum_{i \in I_k} G_i$
$w_{k+1} = w_k - \alpha \left(cur\_G_k - old\_G_k + \frac{1}{n} \sum_{i=1}^{n} G_i \right)$
for $i \in \{1, \ldots, n\}$ **do**
$i \in I_k$ **or** $i = \arg\min(\text{LastUpdate})$
$G_i = \nabla f_i(w_k)$
LastUpdate$[i] = k$
Return w_{k+1}

Labeling of the SSBD dataset with valence and arousal values

- Goal: Annotate the SSBD dataset with arousal and valence values.
- Arousal: Represents the subject's level of excitement.
- Valence: Represents the subject's level of emotional positivity.
- Labeling Procedure: Introduced a six-category ordered scale (0 to 5) for both arousal and valence.
- Annotator Training: Each annotator receives a detailed explanation of the labeling protocol before starting.

Labeling of the SSBD dataset with valence and arousal values



Both arousal and valence labeling protocol explaining how to give a correct annotation. All the values must be between 0 and 5. Five distinct annotators were requested to label the videos in accordance with the protocol's instructions. The average of frame annotations from all annotators is computed, followed by a normalization step, which scales the values between [-1,1].

Performances of our proposed approach for affect recognition in children with Autism

Table 1: Performances of our proposed approach in continuous affect levels prediction of children with ASD using SSBD-affect dataset

Type of experience	MSE-Arousal	MSE-Valence
Training and testing on SSBD-affect dataset	0.225	0.174
Pre-training on AffectNet dataset and Fine tuning and testing on SSBD-affect dataset	0.187	0.156

D'autres travaux sur le trouble du spectre de l'autisme et l'intelligence émotionnelle

Reconnaissance continue des émotions des enfants atteints de TSA

- Une approche vidéo pour la reconnaissance continue des affects chez les enfants autistes
- Un nouveau modèle d'apprentissage profond basé sur des transformater pour la reconnaissance des affects à partir d'images
- Les données concernant les personnes neurotypiques ne sont pas pertinentes pour l'étude des enfants autistes
- L'analyse du comportement par stimulation et du visage peut contribuer à l'évaluation de l'autisme

Dia, M., Khodabandelou, G., Sabri, A. Q. M., & Othmani, A. (2024). Video-based continuous affect recognition of children with Autism Spectrum Disorder using deep learning. Biomedical Signal Processing and Control, 89, 105712.

Classification et surveillance de l'autisme à partir des émotions catégorielles et dimensionnelles

- Objectif : Mieux comprendre les émotions des enfants autistes.
- Méthode : Associer des émotions simples (catégories) à des émotions plus nuancées (dimensions continues).
- Technologie :
  Un réseau neuronal (CNN) identifie d'abord une émotion de base.
  Un modèle de régression profonde prédit ensuite l'intensité continue de cette émotion.
- Application : Le système est accessible via une application web pour visualiser les résultats en vidéo.
- Analyse de l'autisme : Le système classe les comportements d'autostimulation (stimming) à partir d'images et de vidéos, en utilisant les émotions et comportements détectés.

Khor, S. W. H., Md Sabri, A. Q., & Othmani, A. (2024). Autism classification and monitoring from predicted categorical and dimensional emotions of video features. Signal, Image and Video Processing, 18(1), 191-198.

Anciens postdoctorants, doctorants et étudiants de master

MERCI POUR VOTRE ATTENTION

References i

[1] Sulaiman A. Aburakhia et al. "A Transfer Learning Framework for Anomaly Detection Using Model of Normality". In: *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (2020), pp. 0055–0061.

[2] A. Dhall et al. "Video and image based emotion recognition challenges in the wild: Emotiw 2015.". In: *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 2015, pp. 423–426.

[3] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, and Marius Popescu. "Local learning with deep and handcrafted features for facial expression recognition". In: *IEEE Access* 7 (2019), pp. 64827–64836.

References ii

[4] Jonathan Gratch et al. "The distress analysis interview corpus of human and computer interviews.". In: *LREC*. 2014, pp. 3123–3128.

[5] Jing Han et al. "Strength modelling for real-world automatic continuous affect recognition from audiovisual signals". In: *Image and Vision Computing* 65 (2017), pp. 76–86.

[6] Lang He et al. "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks". In: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* (2015), pp. 73–80.

[7] Shawn Hershey et al. "CNN Architectures for Large-Scale Audio Classification". In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017).

References iii

[8] Steven CY Hung et al. "Increasingly packing multiple facial-informatics modules in a unified deep-learning model via lifelong learning". In: *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. 2019, pp. 339–343.

[9] M. Lyons et al. "Coding facial expressions with gabor wavelets.". In: *Proceedings Third IEEE international conference on automatic face and gesture recognition*. 1998, pp. 200–205.

[10] Xingchen Ma et al. "Depaudionet: An efficient deep model for audio based depression classification". In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM. 2016, pp. 35–42.

[11] The National Institute of Mental Health. *Depression*. https://www.nimh.nih.gov/health/topics/depression/index.shtml. 2019.

References iv

[12] Alex J Mitchell, Amol Vaze, and Sanjay Rao. "Clinical diagnosis of depression in primary care: a meta-analysis". In: *The Lancet* 374.9690 (2009), pp. 609–619.

[13] Muhammad Muzammel et al. "AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis". In: *Machine Learning with Applications* 2 (2020), p. 100005. ISSN: 2666-8270. DOI: https://doi.org/10.1016/j.mlwa.2020.100005. URL: https://www.sciencedirect.com/science/article/pii/S2666827020300050.

References v

[14] Md Nasir et al. "Multimodal and Multiresolution Depression Detection from Speech and Facial Landmark Features". In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, Oct. 2016. DOI: 10.1145/2988257.2988261.

[15] A. Othmani et al. "Age estimation from faces using deep learning: A comparative analysis". In: *Computer Vision and Image Understanding* (2020), p. 102961.

[16] Alice Othmani et al. "Towards Robust Deep Neural Networks for Affect and Depression Recognition from Speech". In: *arXiv preprint arXiv:1911.00310* (2019).

References vi

[17] Alice Othmani et al. "Towards Robust Deep Neural Networks for Affect and Depression Recognition from Speech". In: *Pattern Recognition. ICPR International Workshops and Challenges*. Springer International Publishing, 2021, pp. 5–19. DOI: 10.1007/978-3-030-68790-8_1.

[18] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[19] Emna Rejaibi et al. "MFCC-based Recurrent Neural Network for Automatic Clinical Depression Recognition and Assessment from Speech". In: *CoRR* abs/1909.07208 (2019). arXiv: 1909.07208. URL: http://arxiv.org/abs/1909.07208.

[20] Emna Rejaibi et al. "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech". In: *arXiv preprint arXiv:1909.07208* (2019).

References vii

[21] Fabien Ringeval et al. "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". In: *In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (2013), pp. 1–8.

[22] Asif Salekin et al. "A Weakly Supervised Learning Framework for Detecting Social Anxiety and Depression". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2.2 (July 2018). DOI: 10.1145/3214284.

[23] Asif Salekin et al. "A weakly supervised learning framework for detecting social anxiety and depression". In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2.2 (2018), p. 81.

References viii

[24] Gabrielle Simcock et al. "Associations between Facial Emotion Recognition and Mental Health in Early Adolescence". In: *International Journal of Environmental Research and Public Health* 17.1 (2020), p. 330.

[25] Güray Tonguç and Betul Ozaydın Ozkara. "Automatic recognition of student emotions from facial expressions during a lecture". In: *Computers & Education* (2020), p. 103797.

[26] Panagiotis Tzarakis et al. "End-to-end multimodal emotion recognition using deep neural networks". In: *IEEE Journal of Selected Topics in Signal Processing* 11.8 (2017), pp. 1301–1309.

References ix

[27] Michel Valstar et al. "Avec 2016: Depression, mood, and emotion recognition workshop and challenge". In: *Proceedings of the 6th international workshop on audio/visual emotion challenge*. 2016, pp. 3–10.

[28] Adrián Vázquez-Romero and Ascensión Gallardo-Antolın. "Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks". In: *Entropy* 22.6 (June 2020), p. 688. DOI: 10.3390/e22060688.

[29] Raviteja Vemulapalli and Aseem Agarwala. "A compact embedding for facial expression similarity". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5683–5692.

References x

[30] Thanh-Hung Vo et al. "Pyramid with super resolution for In-the-Wild facial expression recognition". In: *IEEE Access* 8 (2020), pp. 131988–132001.

[31] Kai Wang et al. "Region attention networks for pose and occlusion robust facial expression recognition". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 4057–4069.

[32] James R Williamson et al. "Detecting depression using vocal, facial and semantic communication cues". In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 2016, pp. 11–18.

[33] Le Yang et al. "Decision Tree Based Depression Classification from Audio Video and Language Information". In: (Nov. 2016). DOI: 10.1145/2988257.2988269.

References xi

[34] Le Yang et al. "Hybrid depression classification and estimation from audio video and text information". In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 2017, pp. 45–51.

[35] Le Yang et al. "Hybrid depression classification and estimation from audio video and text information". In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 2017, pp. 45–51.

[36] Le Yang et al. "Multimodal measurement of depression using deep learning models". In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM. 2017, pp. 53–59.

[37] Guoshen Yu and Jean-Jacques Slotine. "Audio classification from time-frequency texture". In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2009, pp. 1677–1680.

[38] Jiabei Zeng, Shiguang Shan, and Xilin Chen. "Facial expression recognition with inconsistently annotated datasets". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 222–237.

[39] Jingru Zhang and Nanfeng Xiao. "Capsule Network-Based Facial Expression Recognition Method for a Humanoid Robot". In: *Recent Trends in Intelligent Computing, Communication and Devices*. Springer, 2020, pp. 113–121.

[40] Kaipeng Zhang et al. "Joint face detection and alignment using multitask cascaded convolutional networks". In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1499–1503.