# Reconnaissance automatique sans-contact de l'état affectif de la personne par fusion physio-visuelle à partir de vidéos du visage


Travaux de thèse présentés par

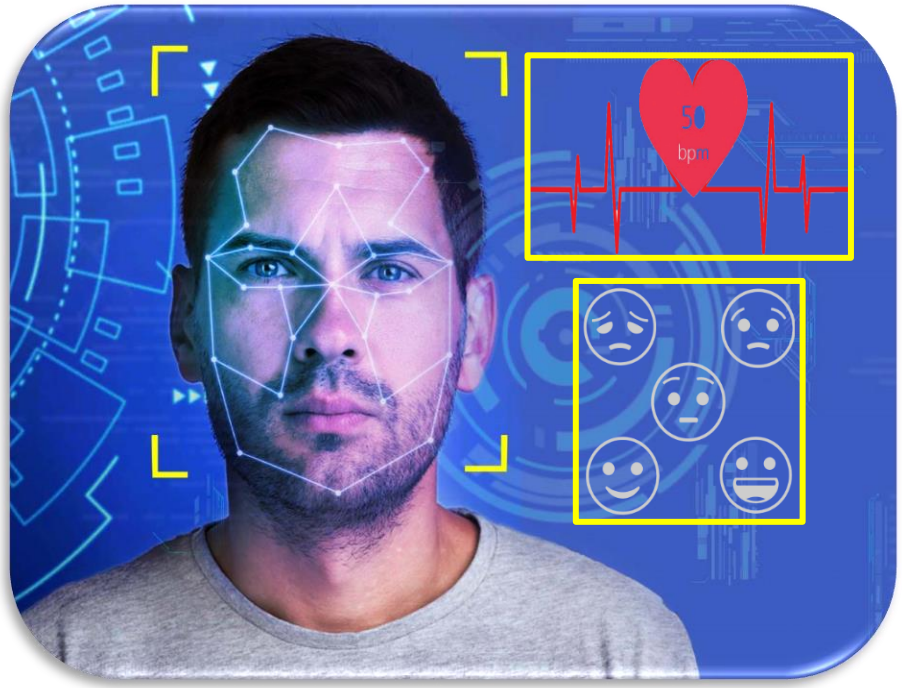
Yassine OUZAR

yassine.ouzar@univ-lorraine.fr

Encadré par :

Pr. Choubeila MAAOUI

Dr. Frédéric BOUSEFSAF

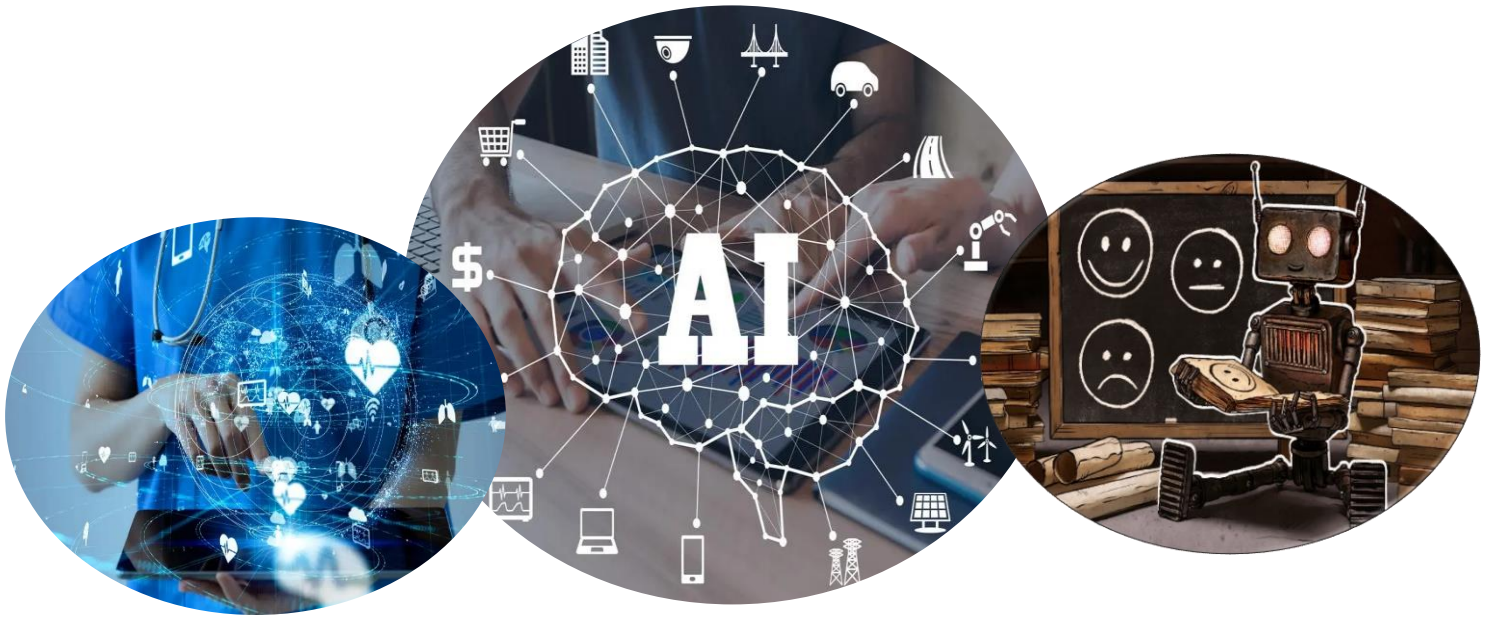


UNIVERSITÉ DE LORRAINE

Laboratoire LCOMS, Equipe ENOSIS, Université de Lorraine.

LCOMS

# Intelligence artificielle



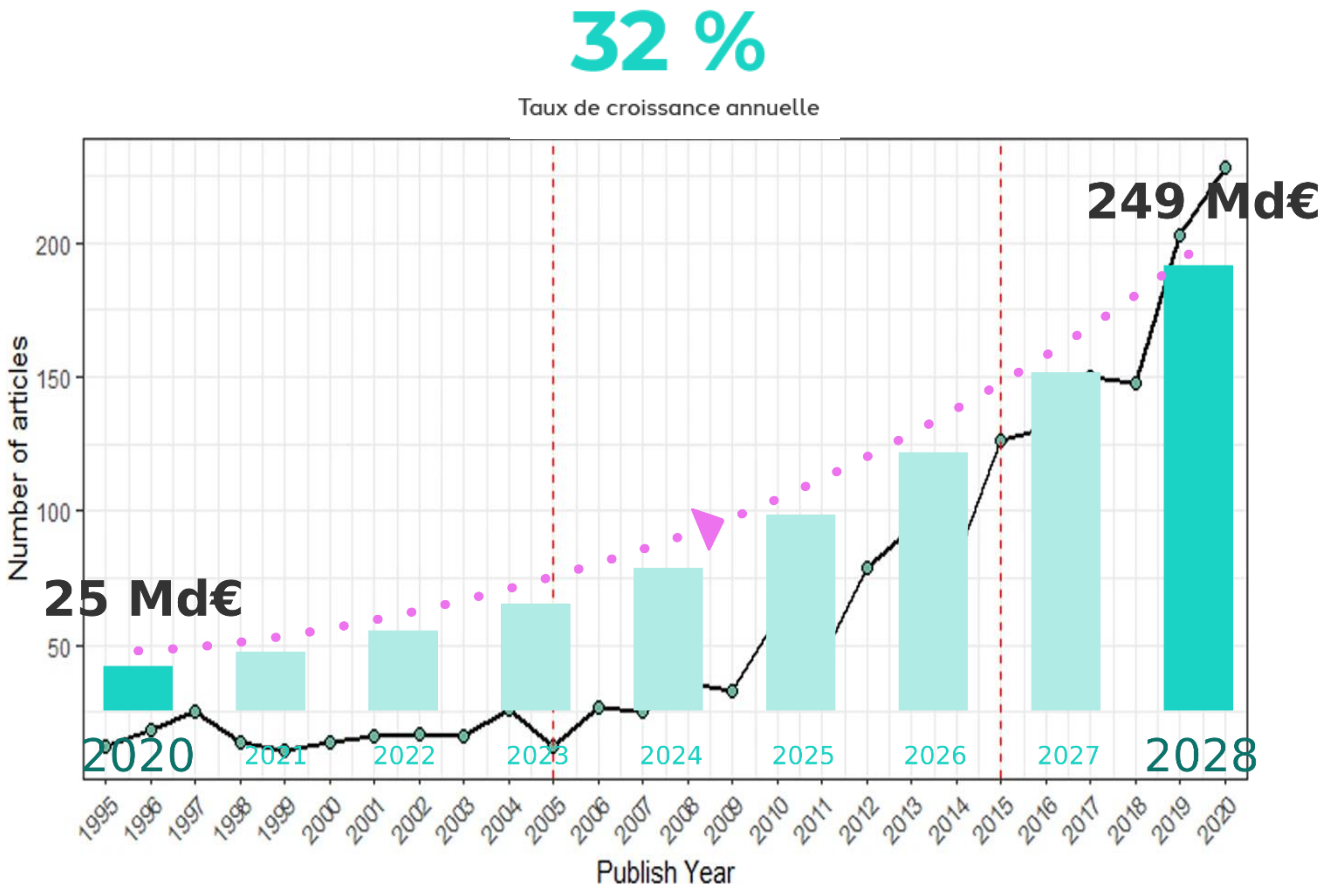

Ingénierie biomédicale

Mesure sans contact des signaux physiologiques

Informatique affective

Reconnaissance multimodale de l'état affectif

Informatique affective

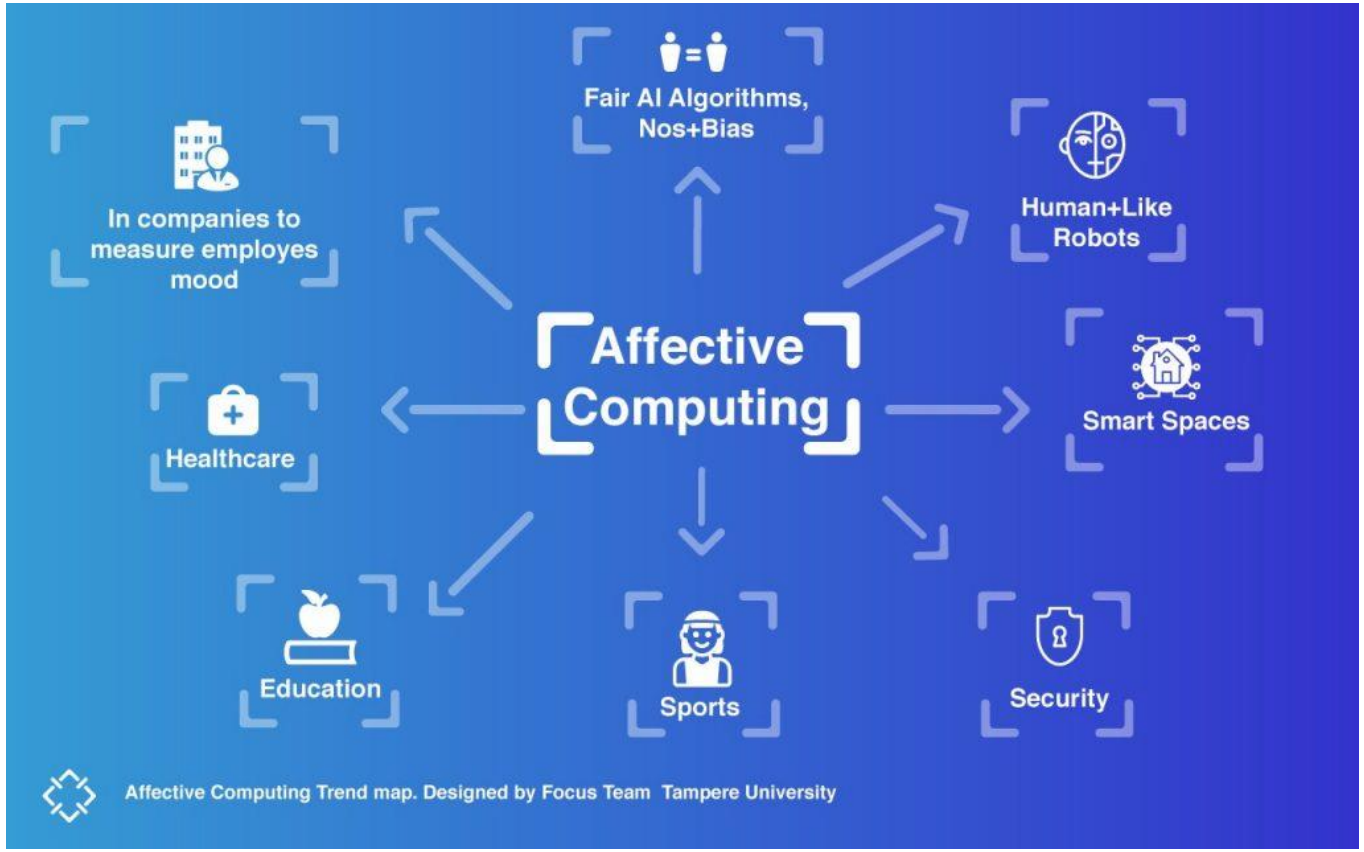



Croissance de la production scientifique

Ho. et al. Affective computing scholarship and the rise of China: a view from 25 years of bibliometric data. Humanit Soc Sci Commun 8, 282 (2021).

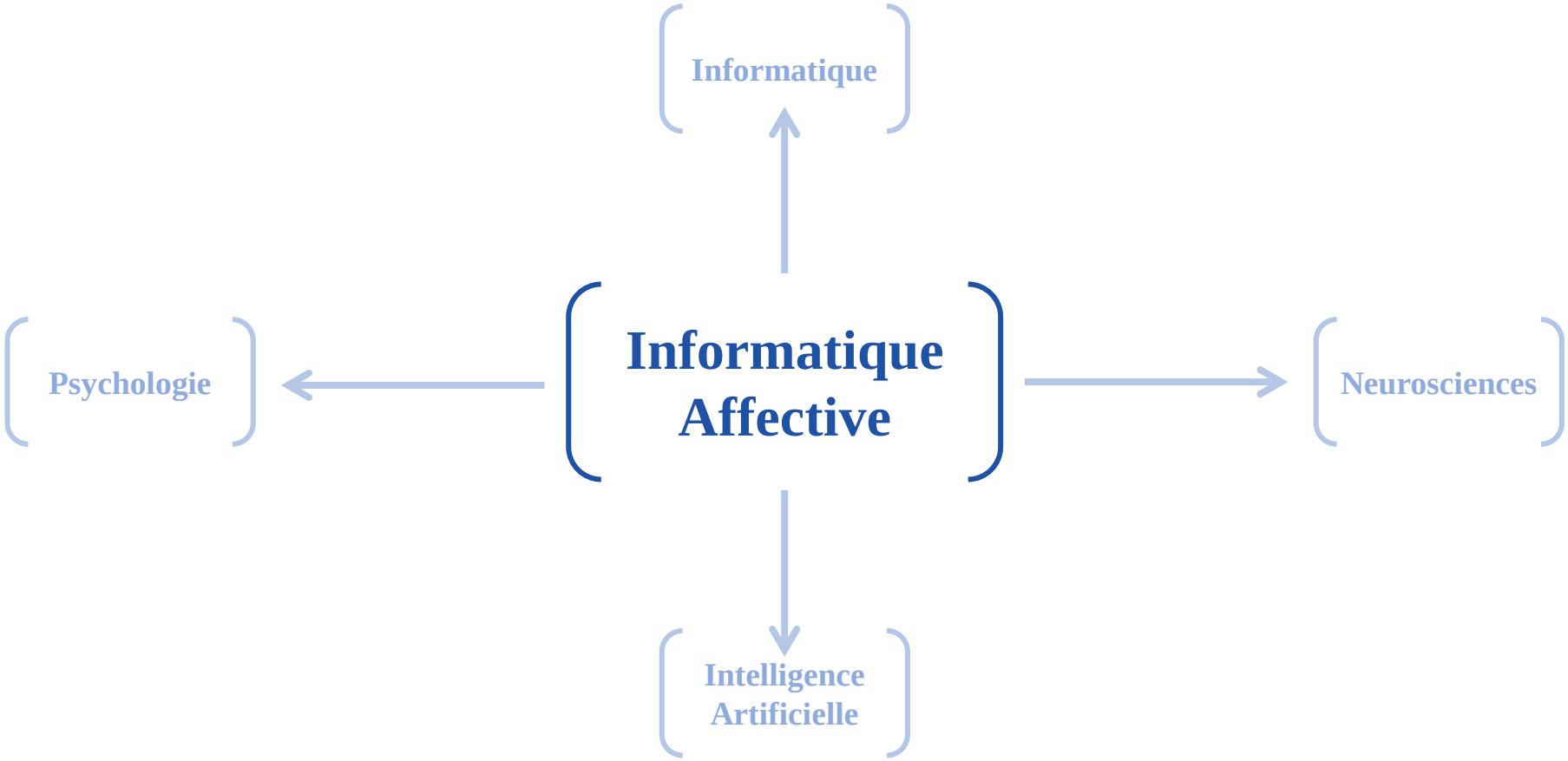
# Applications

Informatique affective

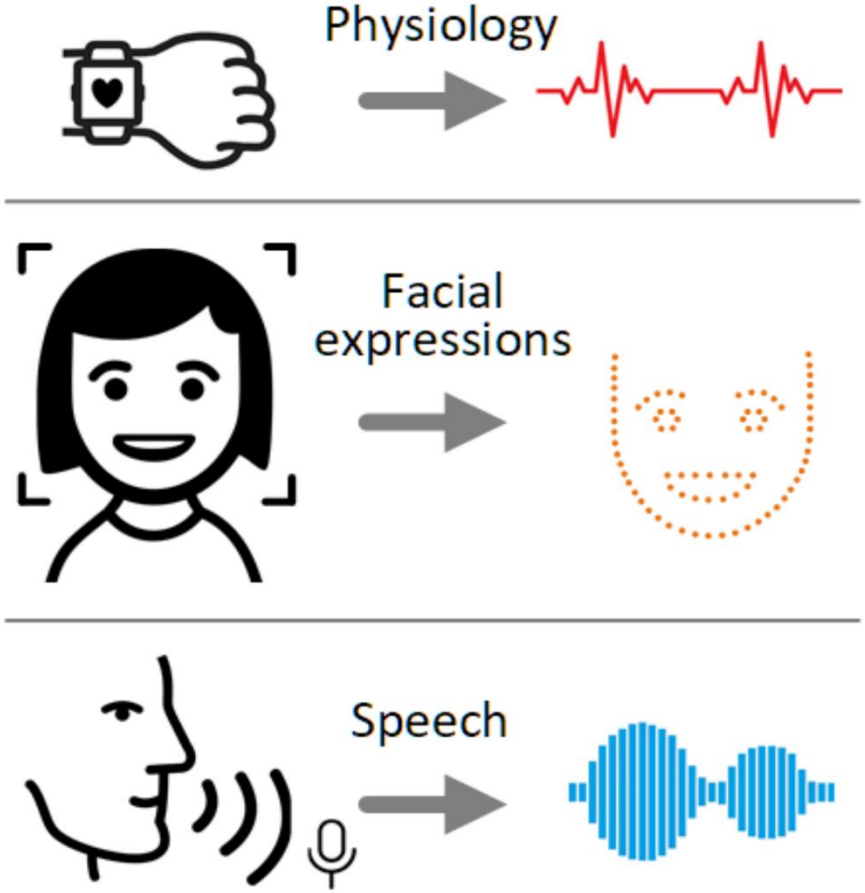


# Informatique affective

- L'informatique affective est un domaine de recherche qui vise à intégrer la compréhension et l'expression des émotions humaines dans les systèmes informatiques.



# Informatique affective

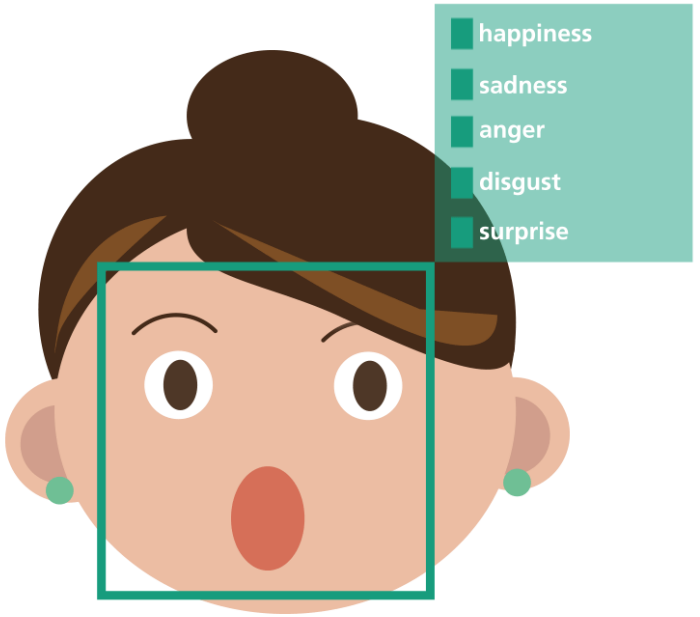


# Les expressions faciales

- **Les expressions faciales sont :**

  - Facilement observables.

  - Facile de collecter un grand ensemble de données.



# Les expressions faciales

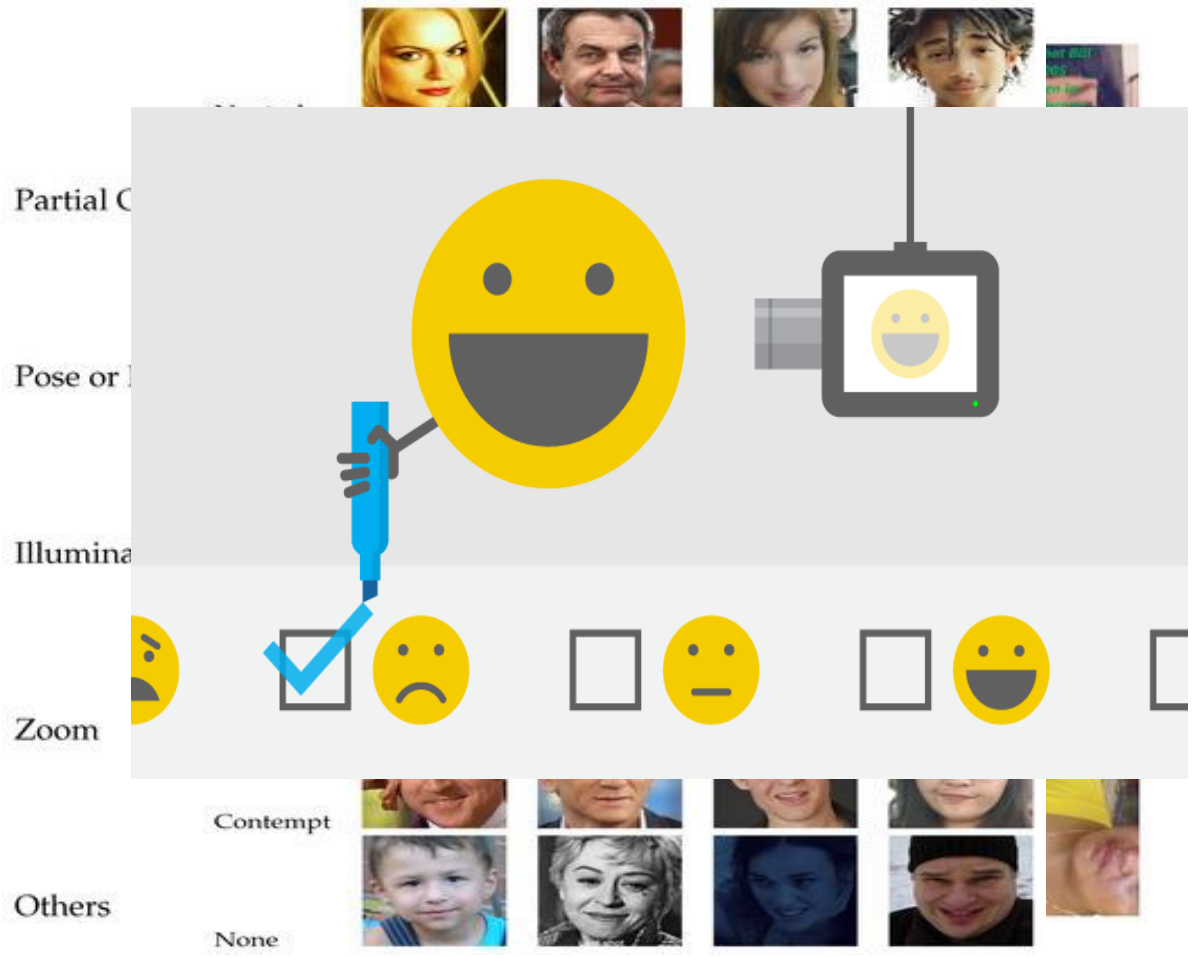
- **Les expressions faciales sont affectées par :**

- Les conditions environnementales.

- Les différences sociales et culturelles.

- La capacité à les contrôler et à les simuler.

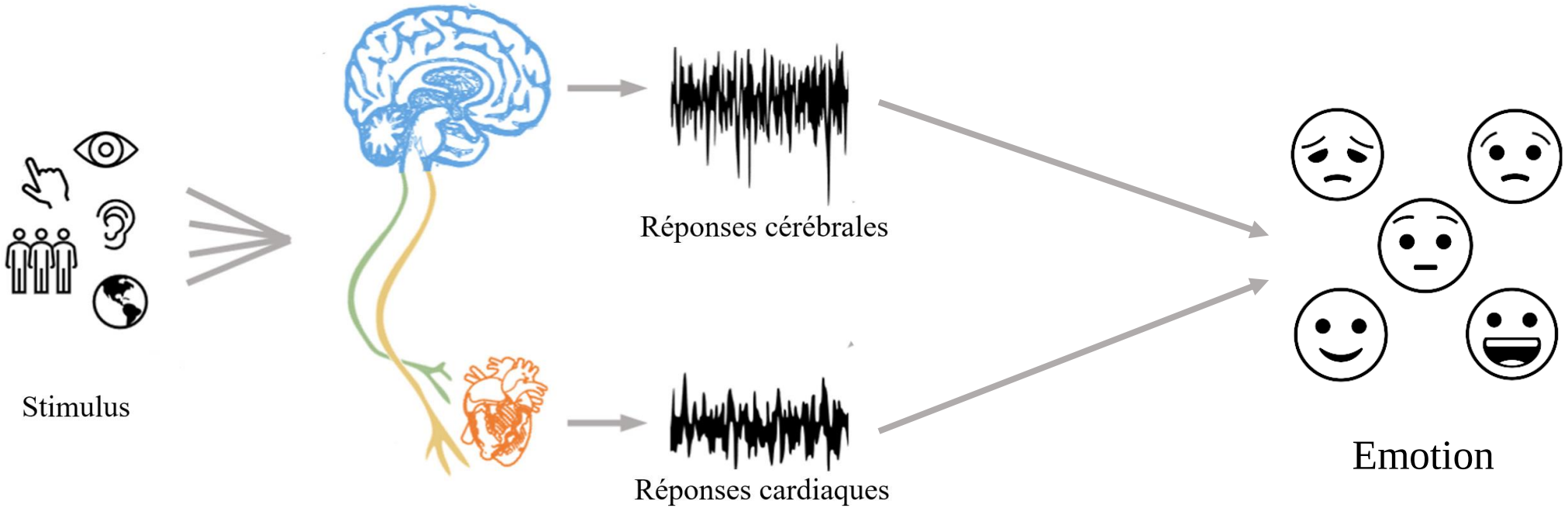
- Ambiguïté et dépendance au contexte.




Li, Shan, and Weihong Deng. "Deep facial expression recognition: A survey." IEEE transactions on affective computing 13.3 (2020).

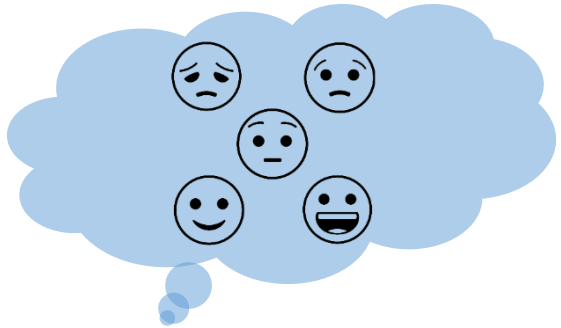
# Les signaux physiologiques

- Les signaux physiologiques sont très pertinents pour l'évaluation de l'état affectif.
- Chaque émotion est caractérisée par une variation physiologique particulière associée à des modifications du système nerveux autonome.

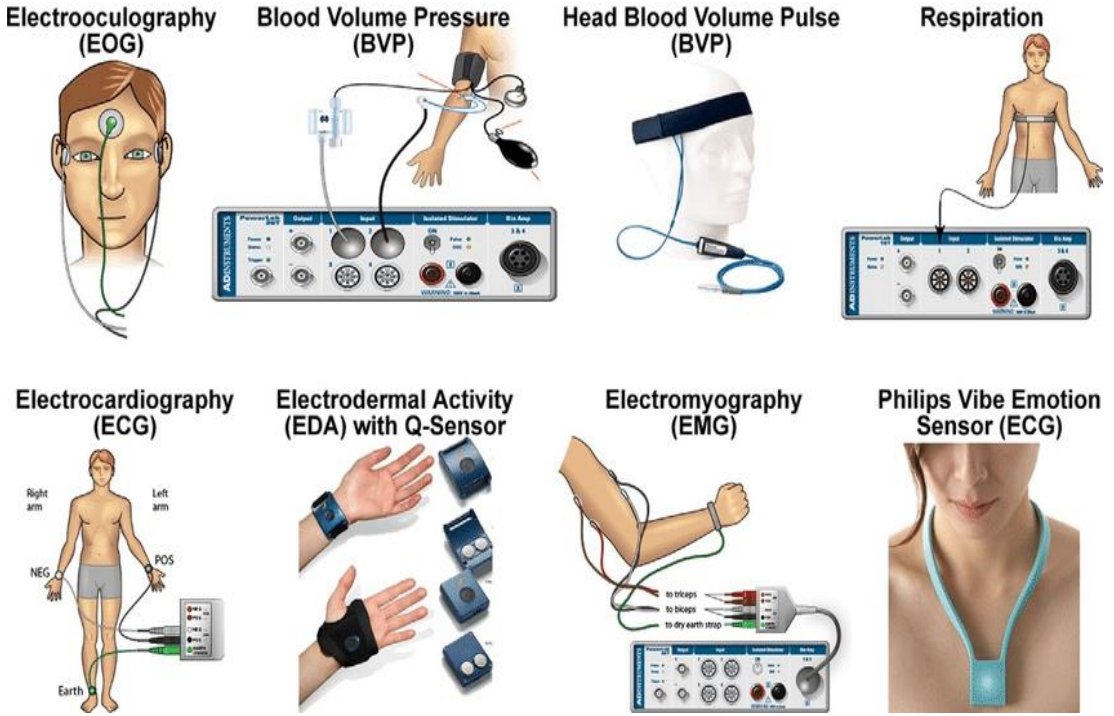



R. W. Levenson, "Emotion and the autonomic nervous system : A prospectus for research on autonomic specificity.," 1988.

# Les signaux physiologiques



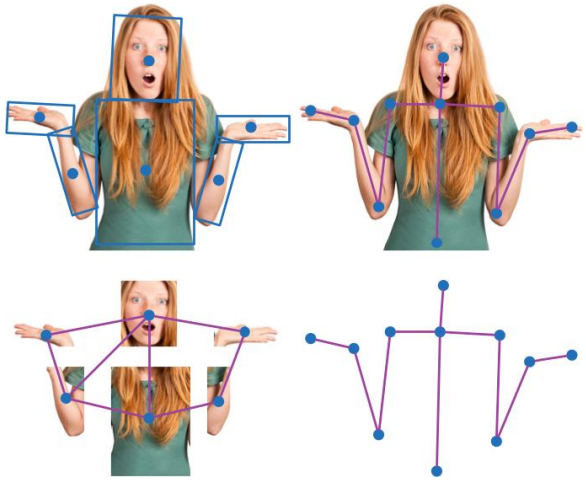
- **Les signaux physiologiques sont :**
  - Incontrôlables.
  - Moins affectés par les différences sociales et culturelles.
  - Riches d'informations complémentaires.
  - Mesures intrusives et/ou contraignantes.
  - Limités à l'utilisation en laboratoire.



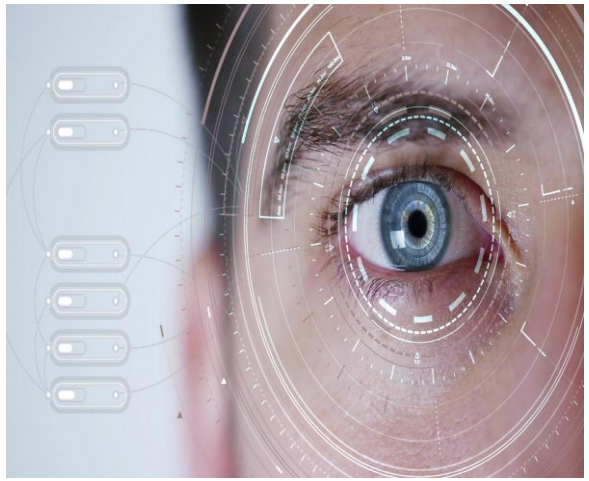

M. Egger et al., "Emotion recognition from physiological signal analysis : A review," The proceedings of AmI, the 2018 European Conference on Ambient Intelligence.

# Autres modalités

## Posture



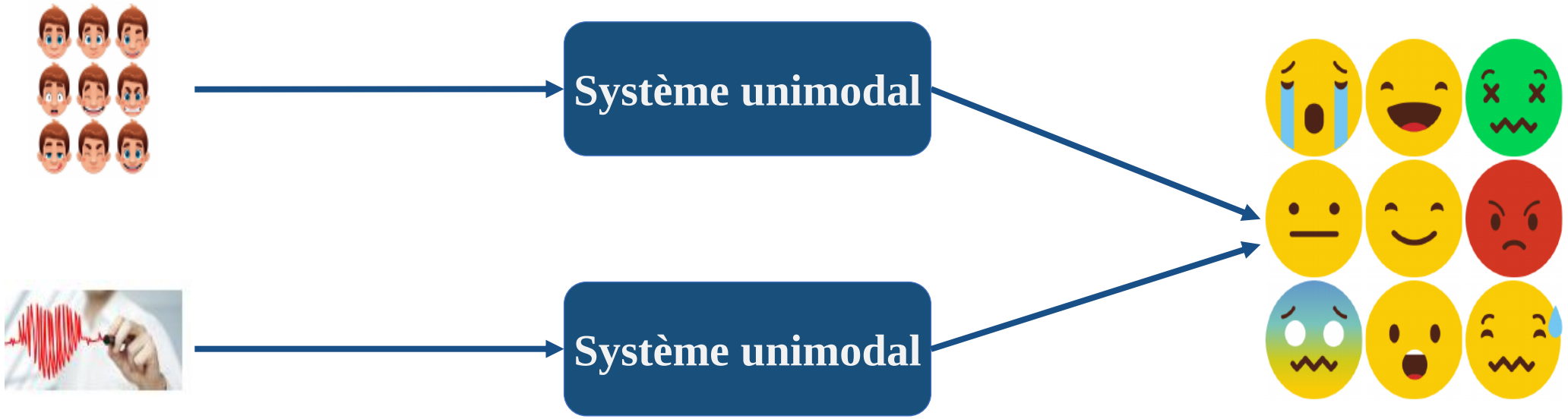
## Regard



## Parole

# Approche unimodale

- Arrivée au point de saturation.
- Impossible de décrire entièrement une émotion.




J. Zhang, et al., "Emotion recognition using multi-modal data and machine learning techniques : A tutorial and review," Inf. Fusion, vol. 59, pp. 103–126, 2020.

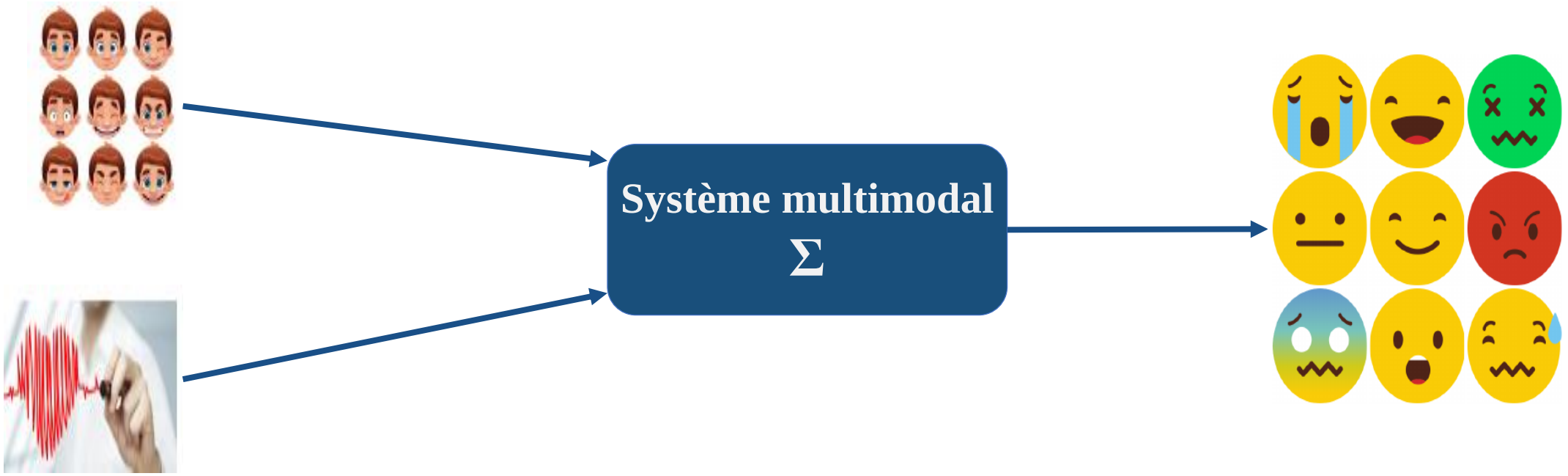
# Approche multimodale

- La fusion de plusieurs modalités est plus représentative et améliore la précision.
- Possède un champ d'applications plus large.


J. Zhang, et al., "Emotion recognition using multi-modal data and machine learning techniques : A tutorial and review," Inf. Fusion, vol. 59, pp. 103–126, 2020.

# Approche multimodale

- La fusion des expressions faciales et des signaux physiologiques permet d'améliorer la précision et de surmonter le problème des expressions contrefaites.



# Positionnement des capteurs et limitations

- Les signaux physiologiques sont habituellement mesurés à l'aide de capteurs en contact


électrodes


Oxymètre


brassard

- Les capteurs en contact ne sont pas utilisables dans toutes les situations


brûlures

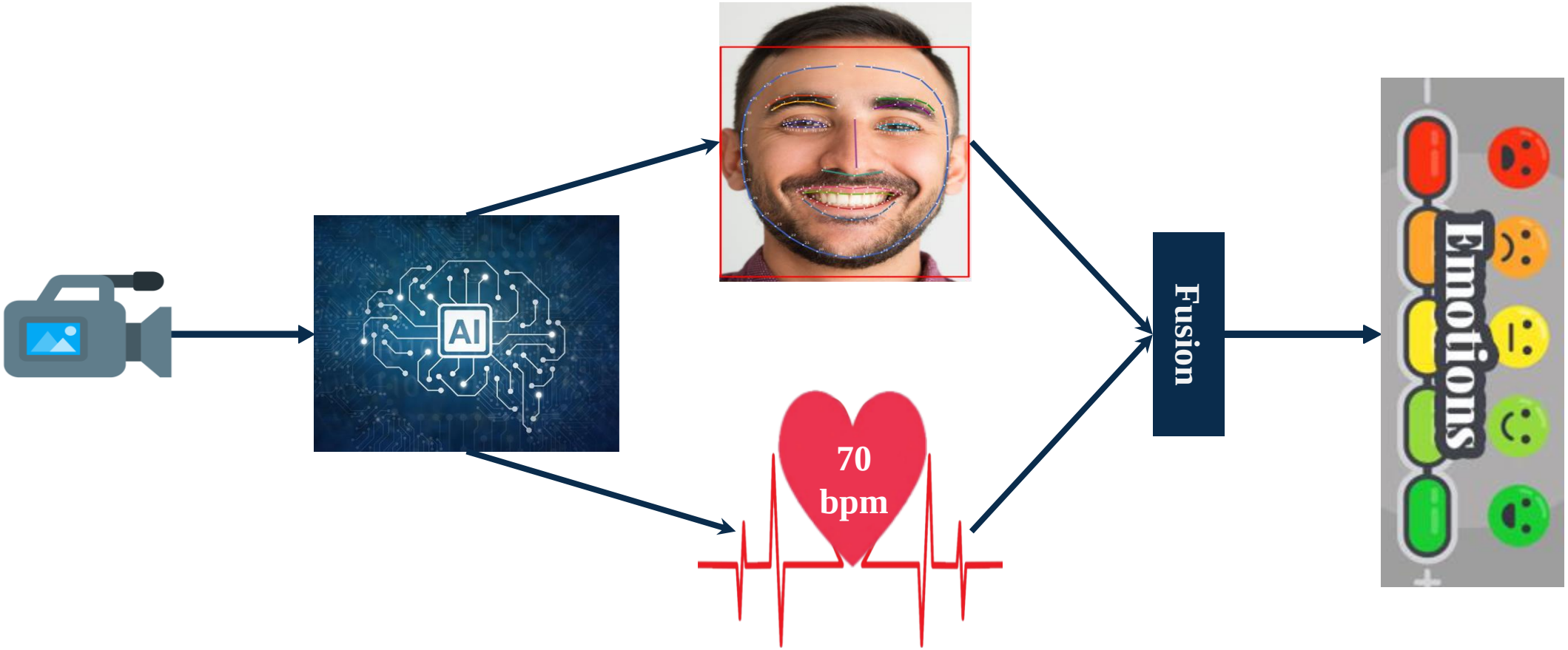
ulcère


maladie contagieuse

- Un expert doit être présent pour assurer la pose des capteurs.

# Fusion physio-visuelle à partir des vidéos du visage

Caractéristiques faciales




Caractéristiques physiologiques

# Capteurs en contact

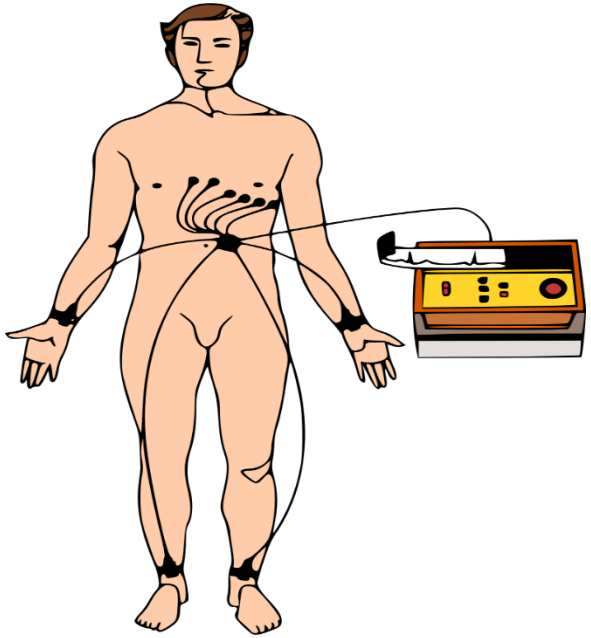
- La fréquence cardiaque constitue l'un des indicateurs essentiels permettant d'évaluer à la fois l'état émotionnel et l'état de santé d'une personne.
- L'électrocardiographie (ECG) et la photopléthysmographie (PPG) sont les principaux moyens permettant de mesurer l'activité cardiaque.

❖ **Avantages :**

- Précis et non invasifs.

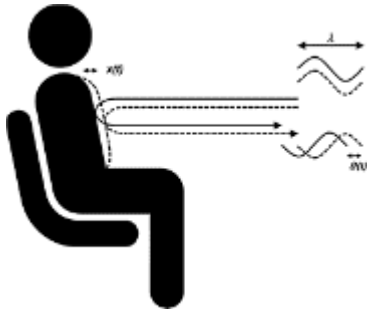
❖ **Inconvénients :**

- Contraignants.
- Psychologiquement stressants.



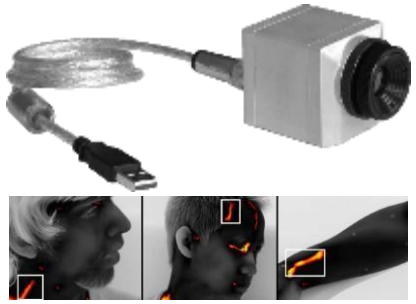
# Capteurs sans contact

- Il existe également différents types de dispositifs sans contact :



Radars Doppler
[Vasu et al., 2010]

Complexes, contraignants et coûteux



Caméras thermiques
[Pavlidis et al., 2007]

Moins précises et coûteuses

Webcam [Poh et al., 2010]

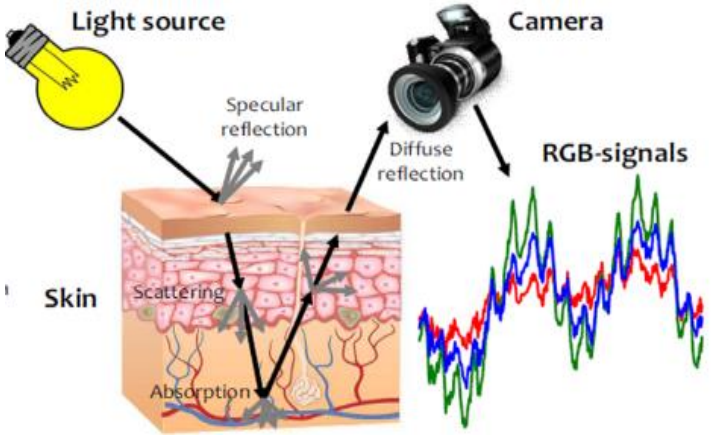
- ► Faible coût
- ► Plus de confort
- ► Champ d'application plus large

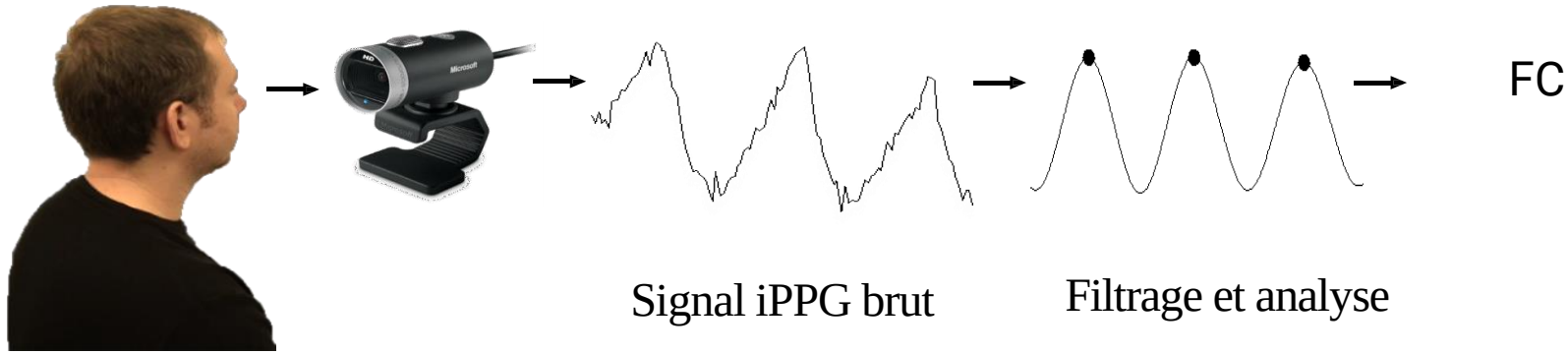
# Photopléthysmographie par imagerie (iPPG)

- Mesure optique des variations périodiques de l'absorption de lumière en fonction du changement de volume sanguin

Ce que nous percevons

Ce que perçoit la caméra



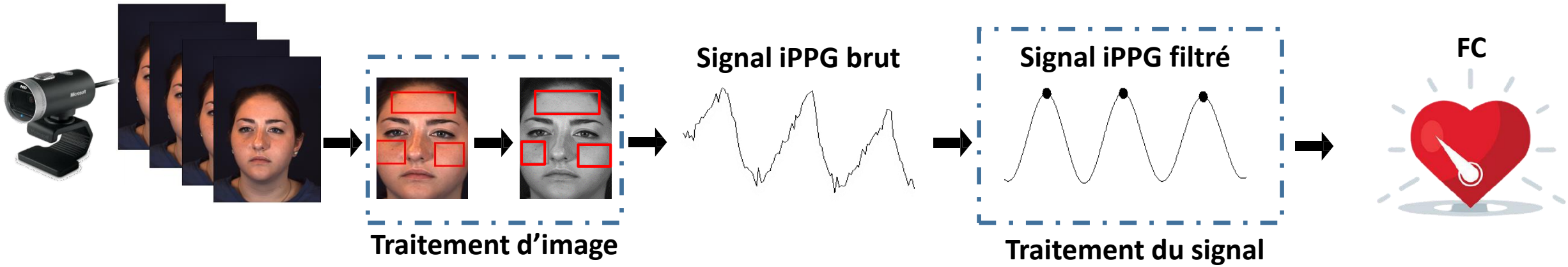




Signal iPPG brut

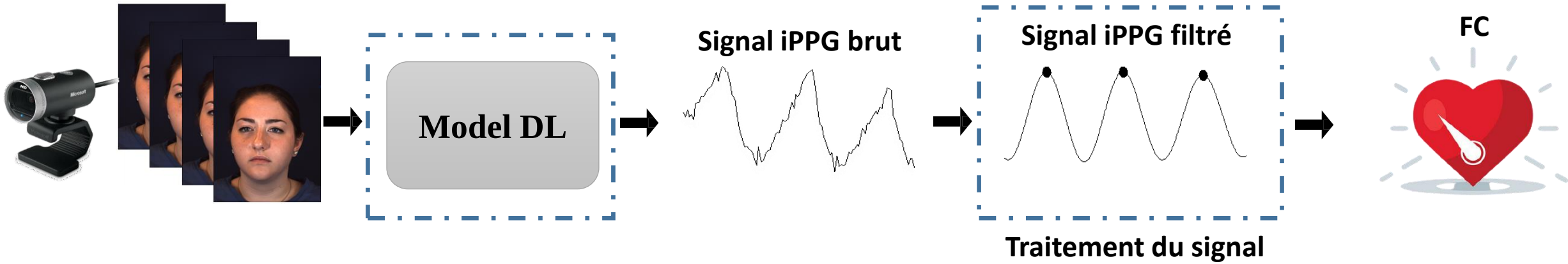
Filtrage et analyse

# Photopléthysmographie par imagerie (iPPG)

- **Méthodes conventionnelles :**



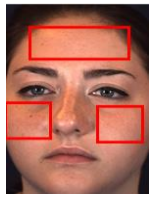

- **Méthodes basées sur deep learning (DL) :**

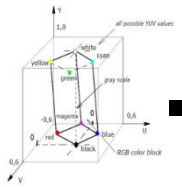


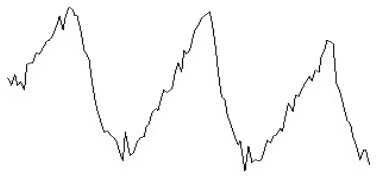
# Méthodes basées sur l'extraction du signal iPPG

- **Méthodes conventionnelles :**

Caméra







Signal iPPG brut

1 2 3

1. **Sélection de la région d'intérêt**
   (dlib [Djeldjli et al., 2021], MediaPipe [Haugg et al., 2023])
2. **Changement d'espace couleur**
   (CHROM [Haan et al., 2013], POS [Wang et al., 2016])
3. **Extraction du signal iPPG**
   Moyenne spatiale [Poh et al., 2010]

Traitement d'image

Signal iPPG filtré

- **Filtrage**
  Filtre passe bande [Verkruysse et al., 2008]
- **Séparation aveugle de sources**
  ICA [Poh et al., 2010]
  PCA [Lewandowska et al., 2011]
- **Estimation de la FC**
  Détection des pics [Wang et al., 2016]
  FFT [Poh et al., 2011]

Traitement du signal

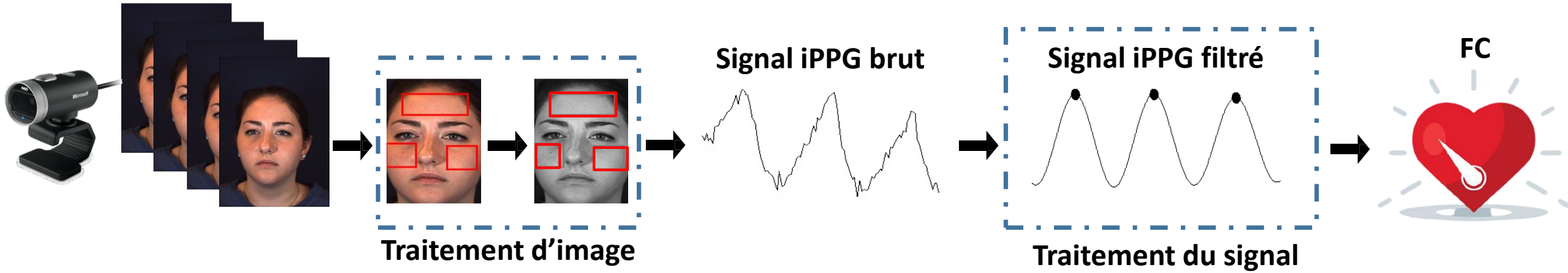
FC

Fréquence Cardiaque

V. Selvaraju et al.,"Continuous monitoring of vital signs using cameras : A systematic review,"Sensors, vol. 22, no. 11, p. 4097, 2022.

# Méthodes basées sur l'extraction du signal iPPG

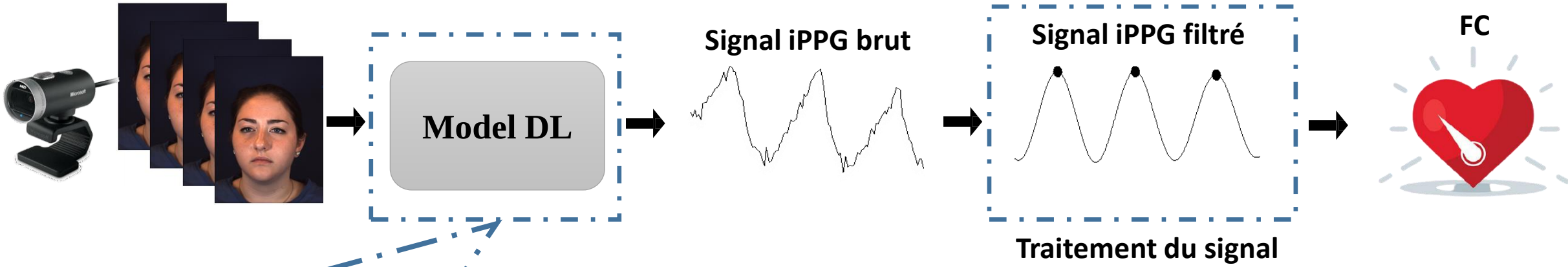
- **Méthodes conventionnelles :**




- Mauvaise capacité de généralisation.
- Nécessite des étapes de pré-traitement et de post-traitement.
- La précision dépend de la qualité du signal iPPG extrait et de la détection des pics.

# Méthodes basées sur l'extraction du signal iPPG

- **Méthodes basées sur deep learning (DL) :**




**CNN 2D + Mécanisme d'attention**
- **DeepPhys** [Chen et al.,2018]

**CNN 2D + CNN 1D**
- **HR-CNN** [Spetlik et al., 2019]

**CNN 3D**
- **PhysNet** [Yu et al.,2019]

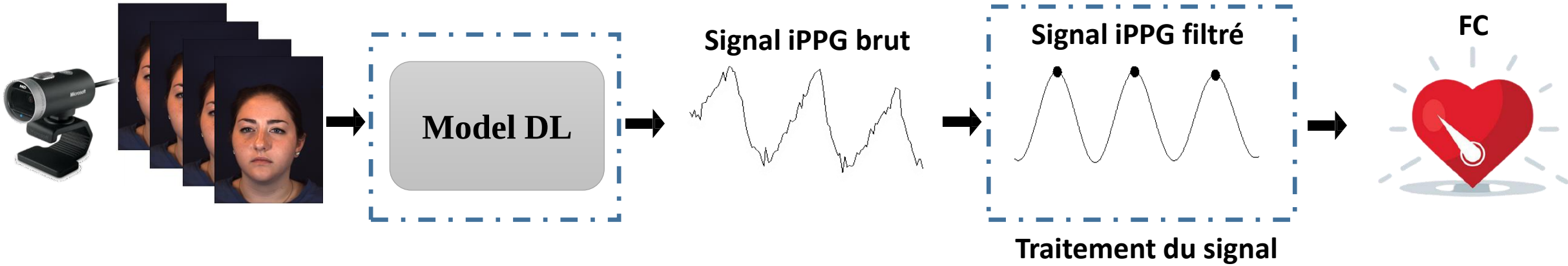
**CNN 3D + RNN**
- **RhythmNet** [Niu et al., 2019]

**NAS**
- **Auto-HR** [Yu et al., 2020]

**GAN**
- **PulseGAN** [Song et al., 2021]

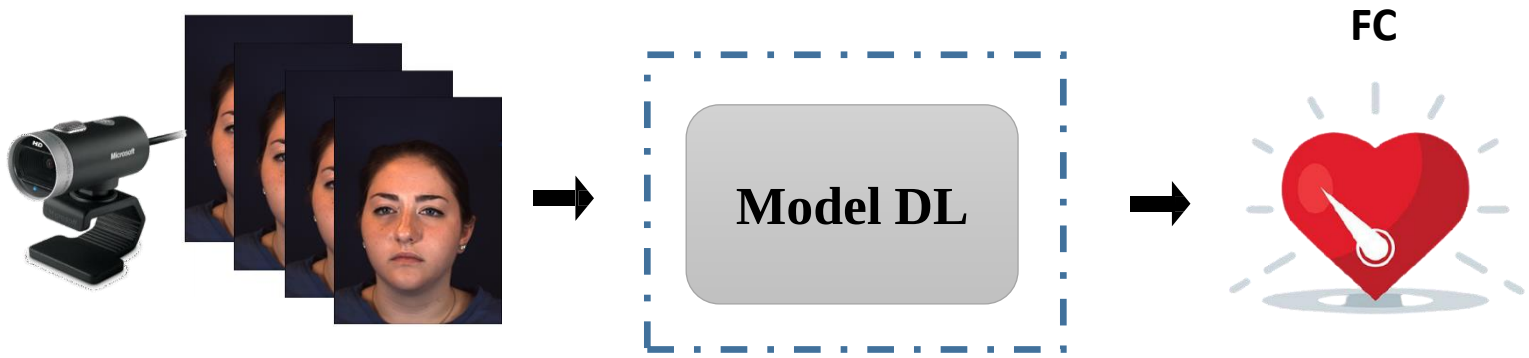
# Méthodes basées sur l'extraction du signal iPPG

- **Méthodes basées sur deep learning (DL) :**




- Bonne capacité de généralisation.
- Nécessite des étapes de pré-traitement et de post-traitement.
- La précision dépend de la qualité du signal iPPG extrait et de la détection du pic principal.

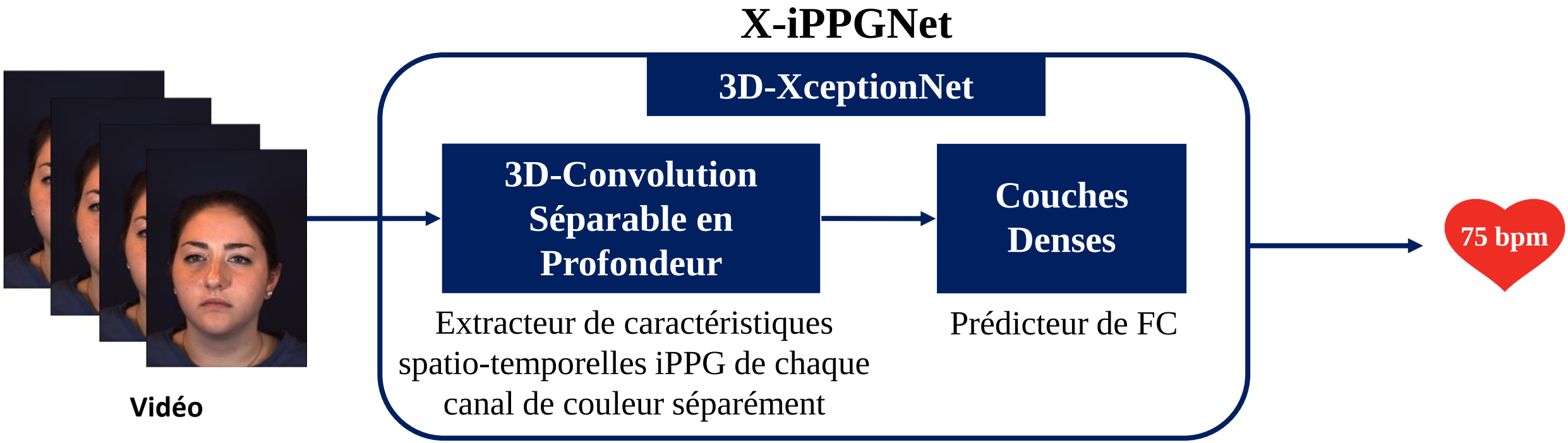
# Estimation de la fréquence cardiaque en une seule étape



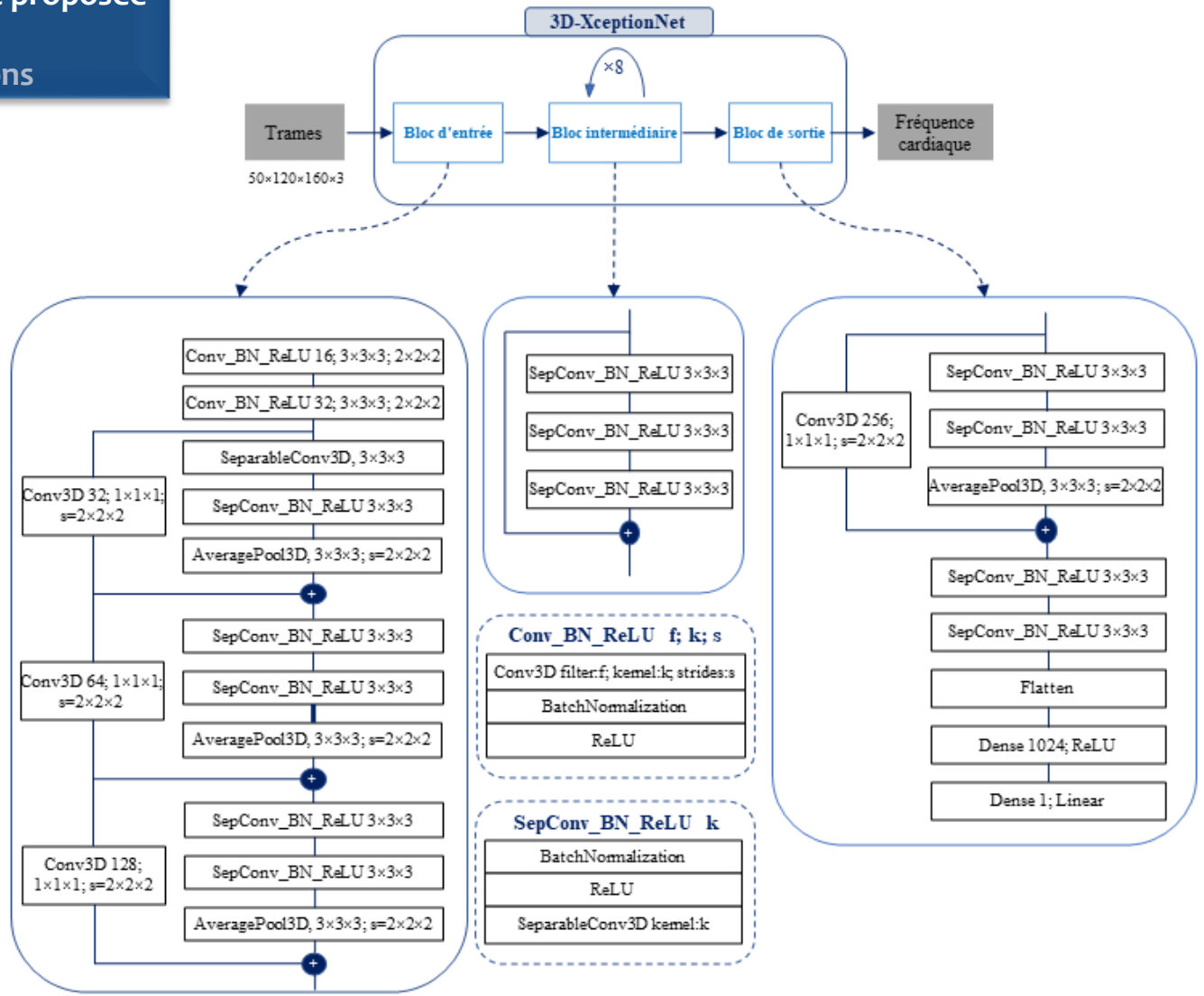

- Estimation directe de la FC sans extraction séparée du signal iPPG.
- Ne nécessite pas des étapes de pré-traitement et de post-traitement.

# Estimation de la fréquence cardiaque en une seule étape

- La prédiction de la fréquence cardiaque (FC) est un problème de régression.
- L'entraînement est entièrement supervisé, chaque vidéo prend une valeur de FC comme étiquette.



# Architecture proposée



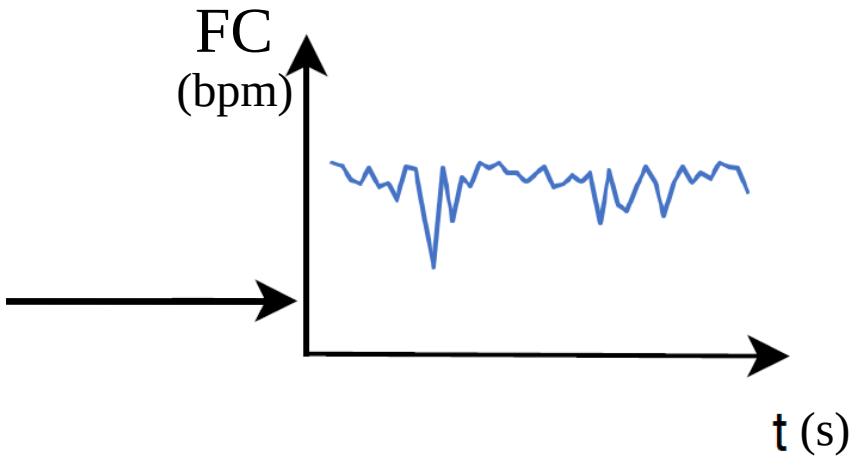

**Convolution séparable en profondeur :**

- Extraction des caractéristiques spatiales et temporelles de chaque canal de couleur séparément.
- Plus économique en termes de temps de calcul et de consommation de mémoire.

**Connexions résiduelles :**

- Réduire l'impact de fuite du gradient.

# Base de données BP4D+



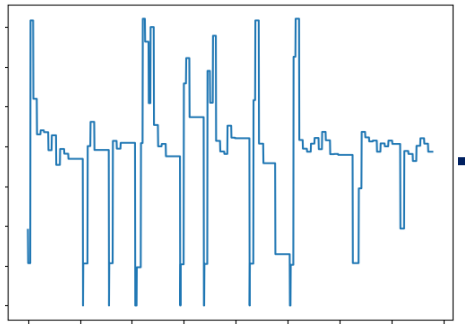

- 140 participants.
- 1400 vidéos RGB (30 s à 1 mn).
- 25 fps, 1040x1392 pixels.

- La fréquence cardiaque (FC) est recueillie par un capteur en contact fonctionnant à 1kHz.

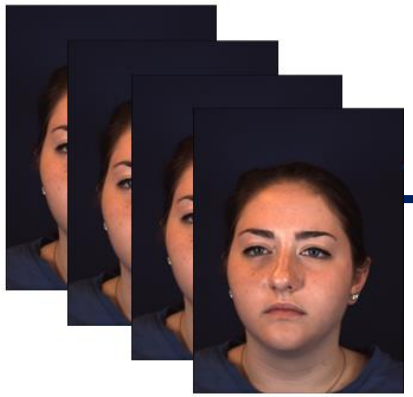
Z. Zhang, et al., Multimodal spontaneous emotion corpus for human behavior analysis. In CVPR, 2016.

# Préparation et nettoyage de données

**Base de données**

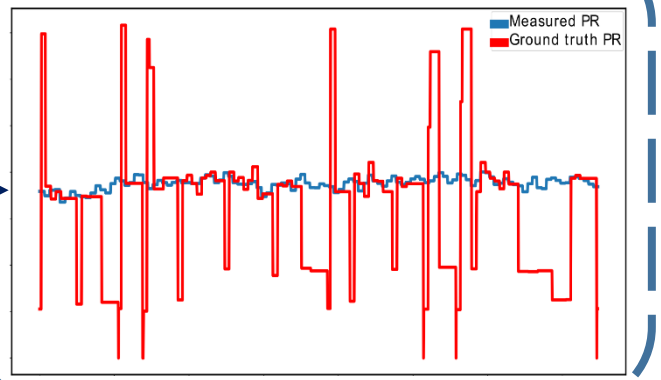


**FC de vérité terrain**

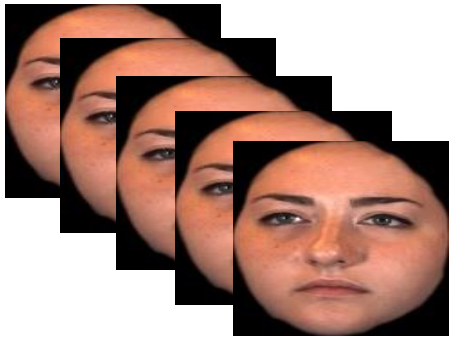


**Vidéo**

**Filtrage et nettoyage**




**Segmentation et découpage du visage** [Nirkin et al. 2017]



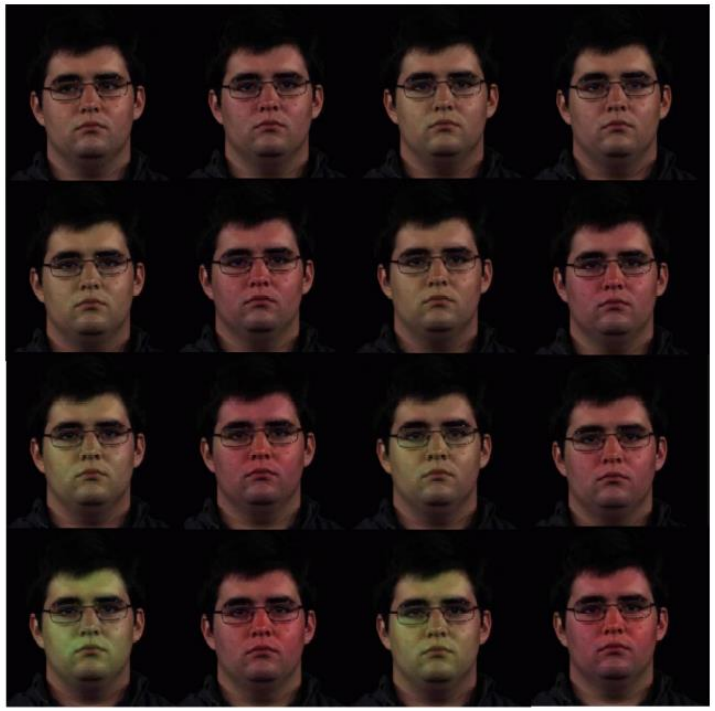
- Entrainer le modèle correctement
- Augmenter le SNR

Y. Nirkin, et al., On face segmentation, face swapping, and face perception, CoRR abs/1704.06729 (2017).

# Augmentation de données

- Accroître le volume de données d'apprentissage et améliorer la robustesse du modèle.

**Transformations géométriques**



**Amplification vidéo**

E. M. Nowara, et al., "Combining magnification and measurement for non-contact cardiac monitoring," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

# Evaluation de la capacité de généralisation

### MMSE-HR [Zhang et al. 2016]

| Méthode | SD (bpm) | RMSE (bpm) | r |
|---|---|---|---|
| Li2014 | 20.02 | 19.95 | 0.37 |
| CHROM | 14.08 | 13.97 | 0.55 |
| PhysNet | 12.76 | 13.25 | 0.44 |
| SAMC | 12.24 | 11.37 | 0.71 |
| RhythmNet | 6.98 | 7.33 | 0.78 |
| AutoHR | 5.71 | 5.87 | 0.89 |
| **X-iPPGNet** | **5,32** | **5.34** | **0.89** |

### UBFC-rPPG [Bobbia et al. 2019]

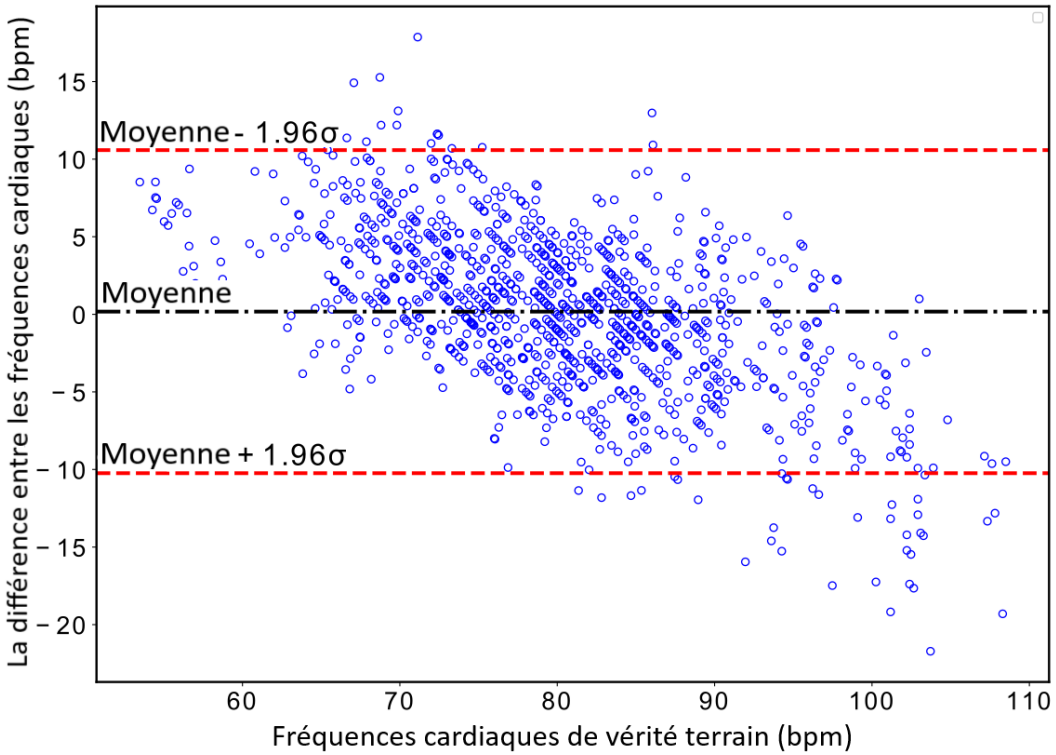
| Méthode | MAE (bpm) | RMSE (bpm) | r |
|---|---|---|---|
| CHROM | 10,6 | 20,3 | - |
| ICA | 8,43 | 18,8 | - |
| 3D-CNN | 5.45 | 8.64 | - |
| Meta-rPPG | 5,97 | 7,42 | 0,53 |
| PRNet | 5.29 | 7.24 | - |
| POS | **4.12** | 10.5 | - |
| **X-iPPGNet** | 4.99 | **6.26** | **0,67** |

### MAHNOB-HCI [Soleymani et al. 2012]

| Méthode | MAE (bpm) | RMSE (bpm) | r |
|---|---|---|---|
| CHROM | 13,49 | 22,36 | 0,21 |
| SAMC | 4,96 | 6,23 | 0,83 |
| HR-CNN | 7,25 | 9,24 | 0,51 |
| rPPGNet | 5,51 | 7,82 | 0,78 |
| PhysNet | 5,96 | 7,88 | 0,76 |
| PulseGan | 4,15 | 6,53 | 0,71 |
| AutoHR | 3,78 | 5,10 | 0,86 |
| RhythmNet | - | 3,99 | 0,87 |
| **X-iPPGNet** | **3,17** | **3,93** | **0,88** |

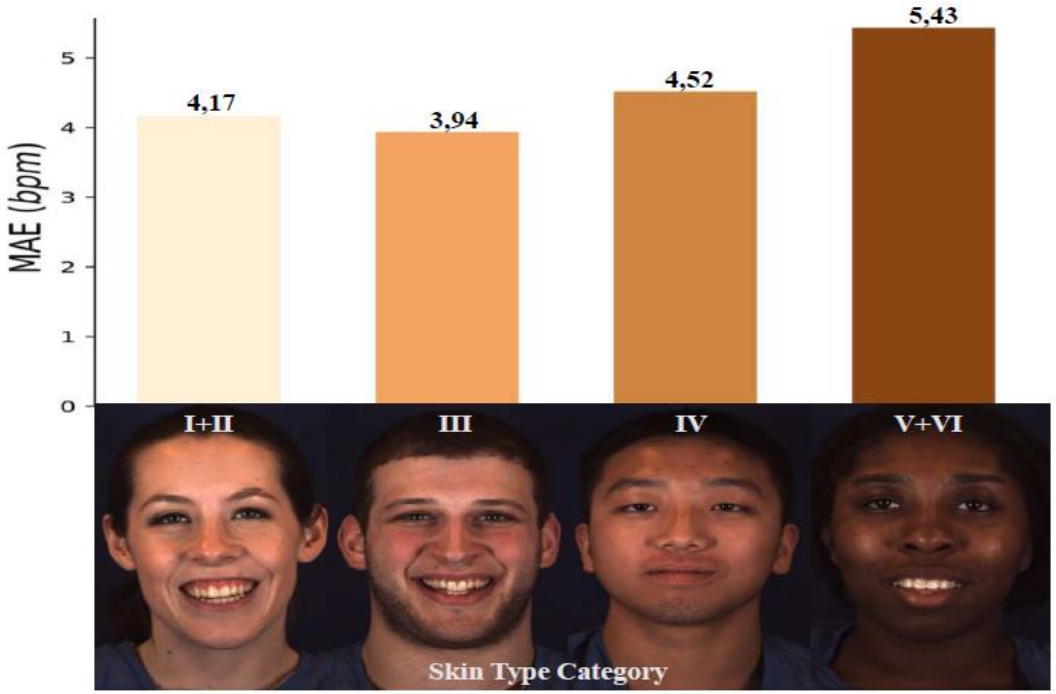
# Impact de la distribution de la fréquence cardiaque

- La distribution est plus concentrée entre les limites de concordances.
- Quelques valeurs aberrantes pour les hautes fréquences.
- Une tendance négative marquée.



# Impact de types de peau

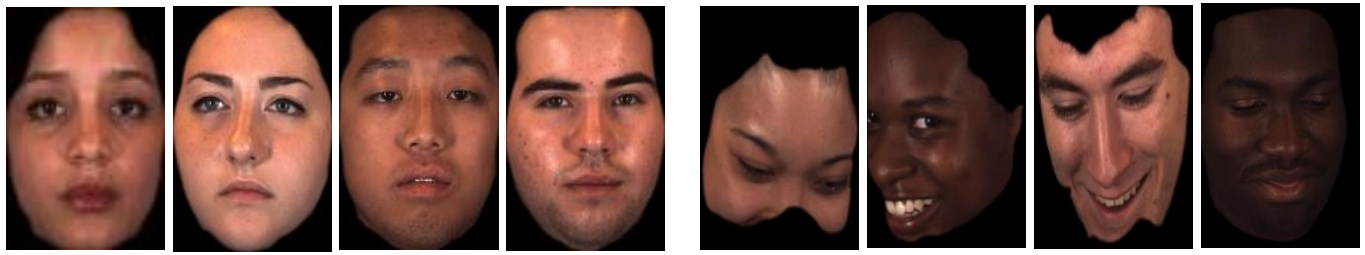
- Les types de peau moins représentés dans la base d'entraînement présentent de mauvaises performances.




| Les types de peau Fitzpatrick | I+II | III | IV | V+VI |
|---|---|---|---|---|
| MAE (bpm) | 4.17 | 3.94 | 4.52 | 5.43 |
| RMSE (bpm) | 5.31 | 5.18 | 5.76 | 6.82 |
| $r$ | 0.87 | 0.81 | 0.84 | 0.40 |

# Impact de mouvement de la tête

- Une faible dégradation de performances entre les vidéos où les têtes sont stables et celles qui présentent des mouvements significatifs.



| Conditions de mouvement de la tête | Stables | Mouvements significatifs |
|---|---|---|
| MAE (bpm) | 3,88 | 4,44 |
| RMSE (bpm) | 4,91 | 5,74 |
| r | 0,86 | 0,82 |

# Taille de la fenêtre de temps Vs Précision Vs Temps de calcul

- Le temps de calcul augmente avec l'augmentation de la taille de fenêtre de temps.

- À l'exception de la fenêtre de 1 seconde qui ne couvre pas la plage des basses fréquences, lorsque la fenêtre de temps augmente, l'erreur de prédiction augmente également.

| Taille de fenêtre de temps | 1 s | 2 s | 3 s | 4 s | 6 s |
|---|---|---|---|---|---|
| MAE (bpm) | 10,21 | **4,10** | 6,41 | 7,75 | 8,13 |
| RMSE (bpm) | 12,89 | **5,32** | 7,98 | 9,77 | 10,03 |
| Temps de calcul (ms) | **120** | 140 | 160 | 180 | 220 |

# Conclusions

- Prédiction de la fréquence cardiaque directement à partir des vidéos du visage sans extraction séparée du signal iPPG et ou des étapes de pré-traitement ou de post-traitement.

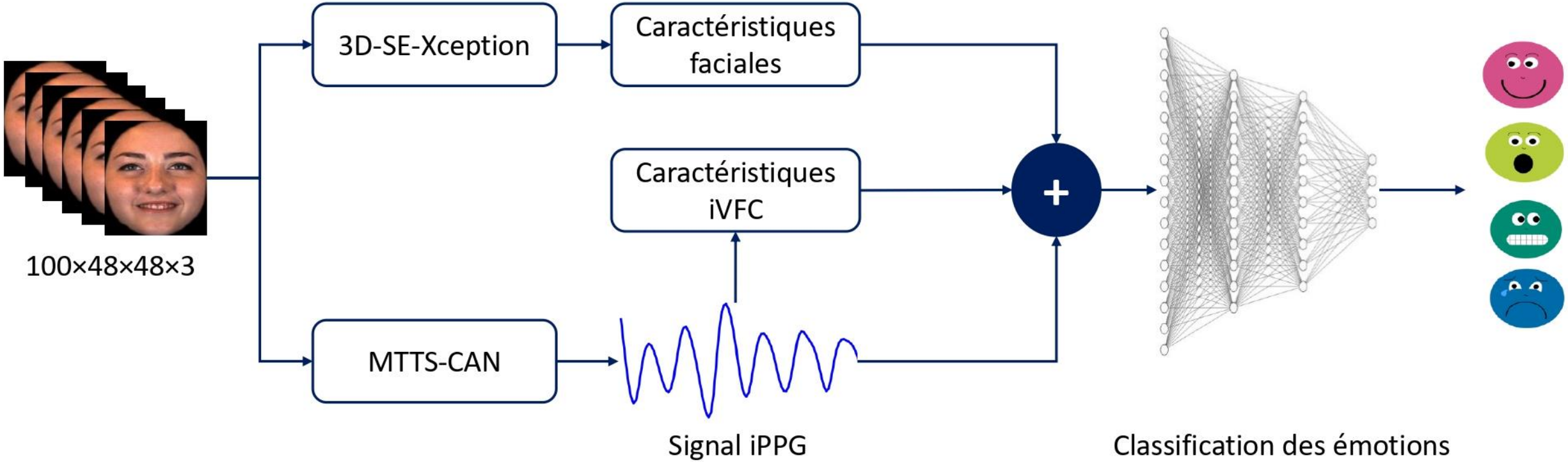
- Avantages en termes de précision, de temps d'exécution et de besoin en mémoire.

# Perspectives

- Génération de données synthétiques en utilisant des GAN.
- Tester des méthodes avancées d'augmentation de données.
- Utilisation des techniques d'apprentissage auto-supervisé ou semi-supervisé.
- Mesure d'autres signaux physiologiques tels que la pression artérielle et la saturation en oxygène.

# Approche proposée

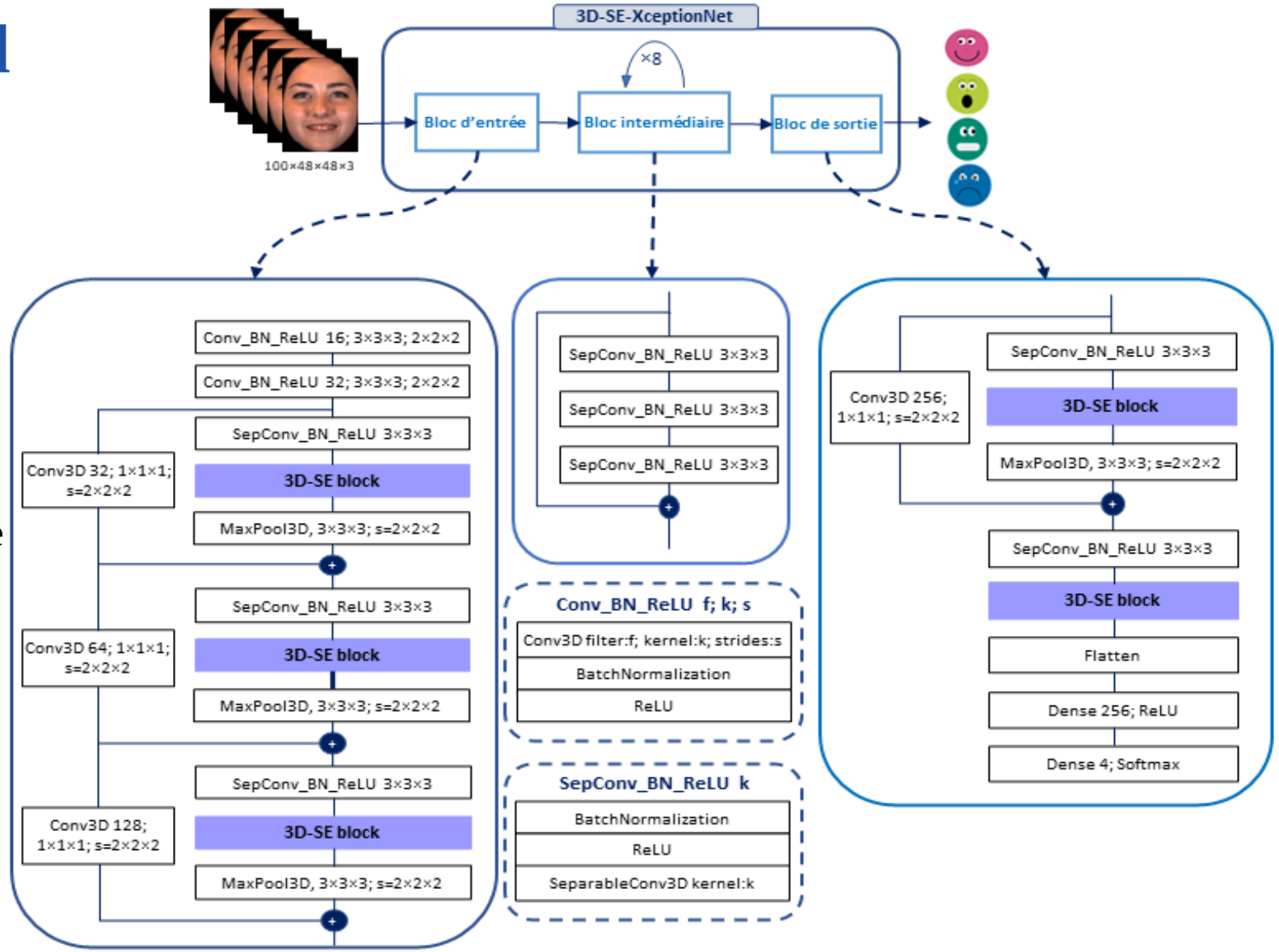
- Reconnaissance des émotions par fusion physio-visuelle à partir des vidéos du visage.



# Pipeline visuel

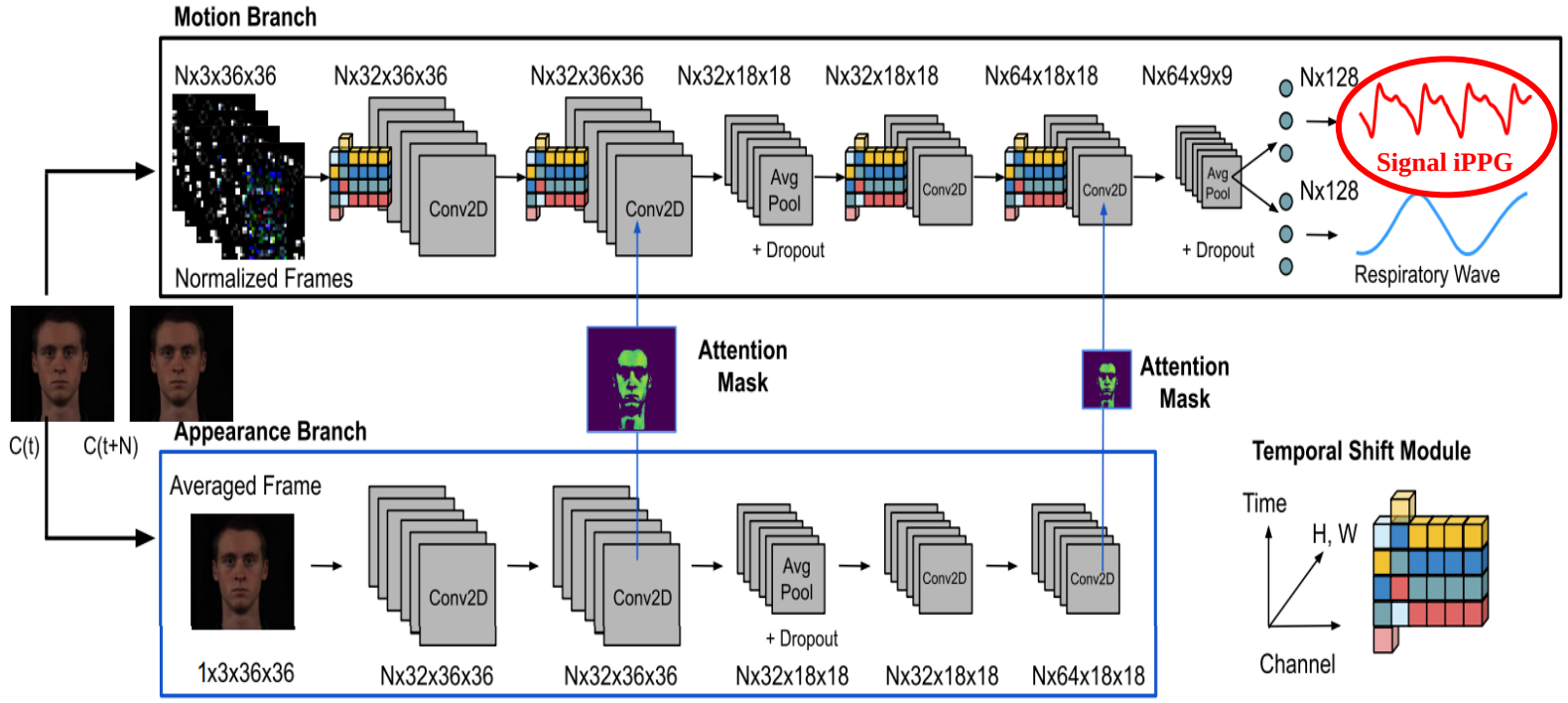
- Combinaison de l'architecture Xception avec le module Squeeze-and-Exitation (SE).

- SE améliore la capacité du réseau à extraire des caractéristiques pertinentes et à réduire le bruit ou les informations redondantes.



# Pipeline physiologique

- MTTS-CAN est basé sur un module de décalage temporel, supervisé par un mécanisme d'attention.



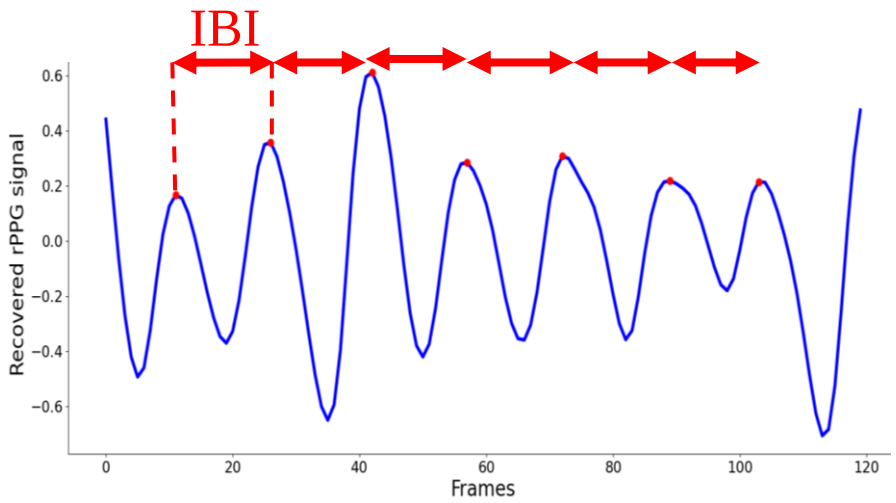

Liu, X., et al.. (2020). Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement. ArXiv, abs/2006.03790.

# Pipeline physiologique

- **6 caractéristiques de la variabilité cardiaque ont été extraites :**

**Domaine temporel :** meanFC, stdFC, RMSSD

➔ FC = 60/IBI ; RMSSD = $\sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(IBI_{i+1} - IBI_i)^2}$

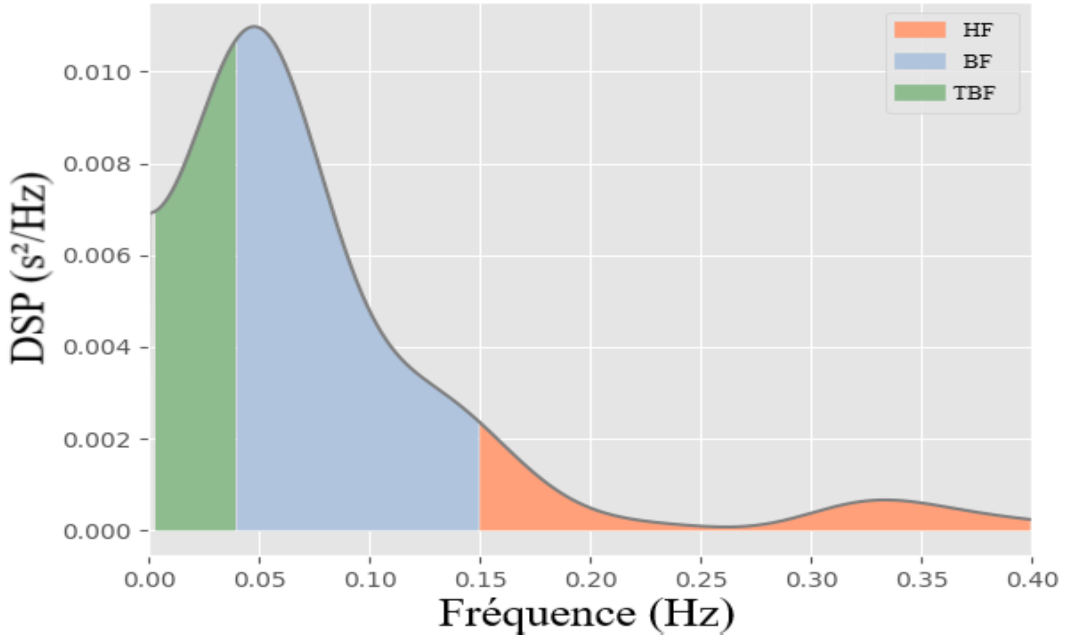


# Pipeline physiologique

- **6 caractéristiques de la variabilité cardiaque ont été extraites :**

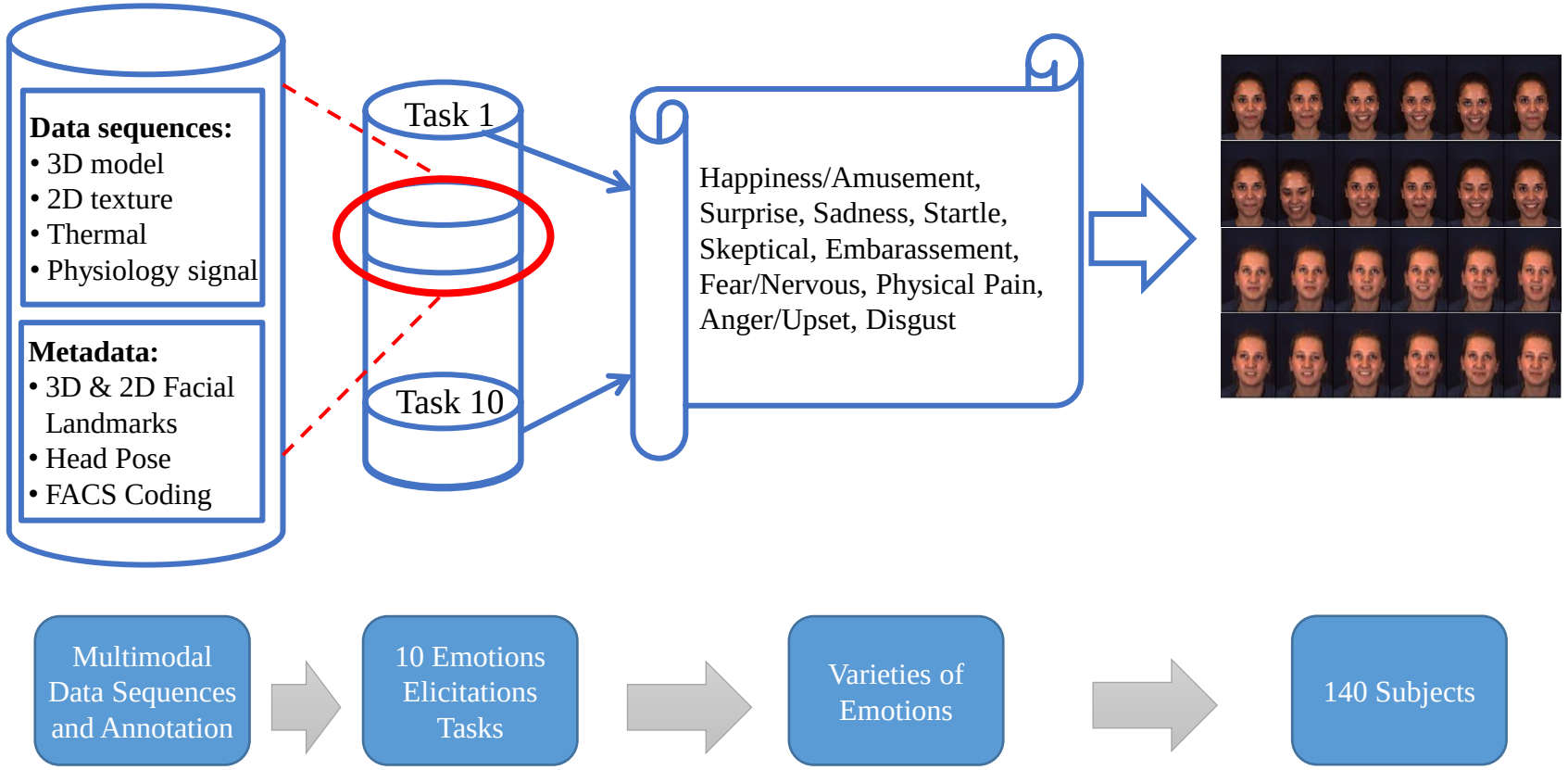
**Domaine fréquentiel :** BF, HF, BF/HF

→ En utilisant la densité spectrale de puissance des séries IBI



# Base de données BP4D+

- Annotation catégorielle.
- Seulement 4 émotions (joie, peur, douleur, embarras) sont annotées.

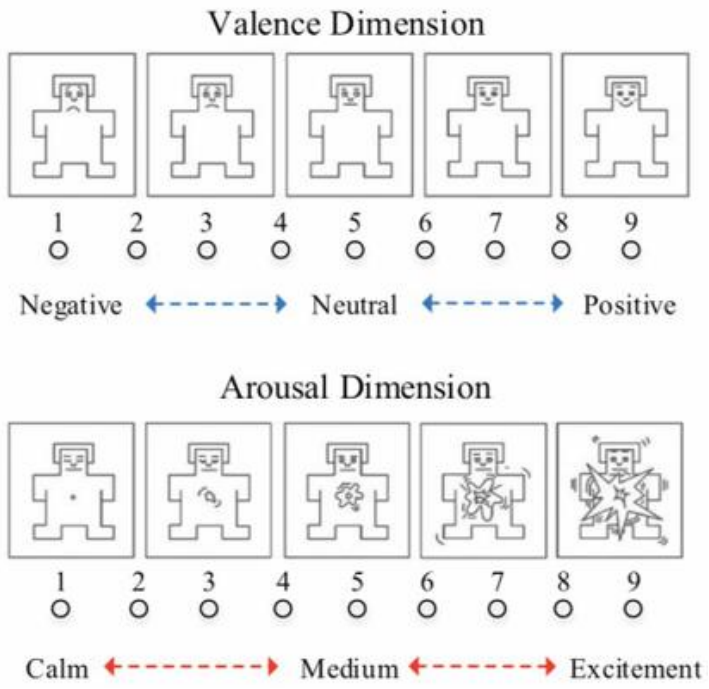



Z. Zhang, et al., Multimodal spontaneous emotion corpus for human behavior analysis. In CVPR, 2016.

# Base de données MAHNOB-HCI

## Enregistrement

## Annotation




M. Soleymani, et al., "A Multimodal Database for Affect Recognition and Implicit Tagging," in IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 42-55, Jan.-March 2012.

# Reconnaissance des émotions à partir des expressions faciales

- Les résultats sont obtenus par une validation croisée 3-fold.

**Résultats sur BP4D+**

| Méthode | Précision (%) |
|---|---|
| 3D-DenseNet | 37,91 |
| 3D-InceptionNet | 42,48 |
| 3D-ResNet | 44,44 |
| 3D-VGGNet | 49,02 |
| 3D-XceptionNet | **53,59** |
| 3D-SE-VGGNet | 57,49 |
| **3D-SE-XceptionNet** | **63,40** |

**Résultats sur MAHNOB-HCI**

| Méthode | Valence (%) | Arousal (%) |
|---|---|---|
| Huang et al. 2016 | 50,57 | 53,64 |
| Wang et al. 2014 | 51,01 | 64,45 |
| Zhong et al. 2017 | 54,06 | 56,47 |
| Koelstra et al. 2013 | 64 | 67,5 |
| Fang et al. 2021 | **67,97** | 66,73 |
| **3D-SE-XceptionNet** | **66** | **74** |

# Reconnaissance des émotions à partir des signaux physiologiques

- La taille et la qualité des signaux iPPG a un impact sur les performances.
- Les résultats dépendent de la base de données utilisées.

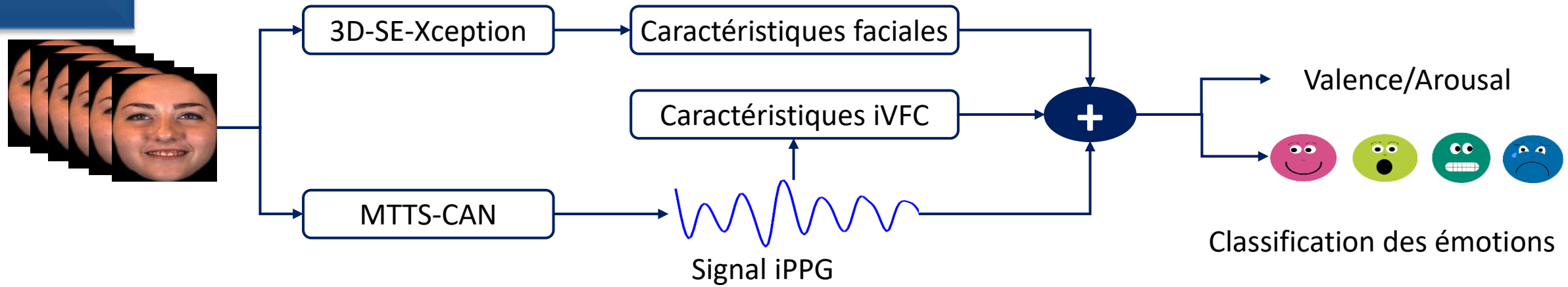
**Résultats sur BP4D+**

| Caractéristiques | Précision (%) |
|---|---|
| **iPPG** | **55,33** |
| iVFC | 53,59 |

**Résultats sur MAHNOB-HCI**

| Caractéristiques | Valence (%) | Arousal (%) |
|---|---|---|
| iPPG | 71 | 67,5 |
| **iVFC** | **78** | **71** |

# Reconnaissance multimodale des émotions




## Résultats sur BP4D+

| Caractéristiques | Précision (%) |
|---|---|
| Caractéristiques faciales | 63,40 |
| iPPG | 55,33 |
| iVFC | 53,59 |
| **Caractéristiques faciales + iPPG** | **71,90** |
| Caractéristiques faciales + iVFC | 70,59 |

## Résultats sur MAHNOB-HCI

| Caractéristiques | Valence (%) | Arousal (%) |
|---|---|---|
| Caractéristiques faciales | 66 | 74 |
| iPPG | 71 | 67,5 |
| iVFC | 78 | 71 |
| Caractéristiques faciales + iPPG | 79 | 76 |
| **Caractéristiques faciales + iVFC** | **86** | **81** |

# Reconnaissance multimodale des émotions

- **Matrice de confusion sur BP4D+**

**EF**

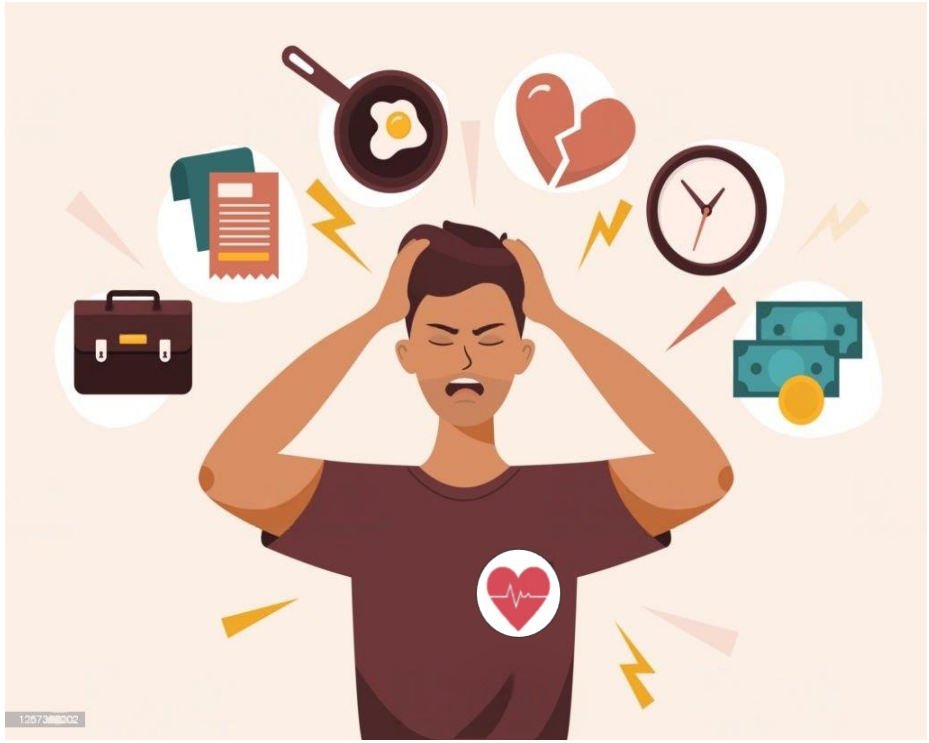

**EF + iPPG**


- La joie et la douleur sont les émotions les plus reconnues.
- La peur est confondue avec la joie et la douleur.

- La fusion des EF et les signaux physiologiques a amélioré la précision des émotions mal identifiées.

# Stress

- Stress ≠ émotion,
- Stress : une réponse physiologique et comportementale à un évènement ou une situation menaçante.



| Emotions associées au stress | Emotions associées au non stress |
|---|---|
| Anxiété | Joie |
| Peur | Amour |
| Colère | Admiration |
| Frustration | Sérénité |

# Approche proposée

VGG16

Caractéristiques faciales

Pipeline visuel

Caractéristiques iVFC

MTTS-CAN

Signal iPPG

Pipeline physiologique

Fusion

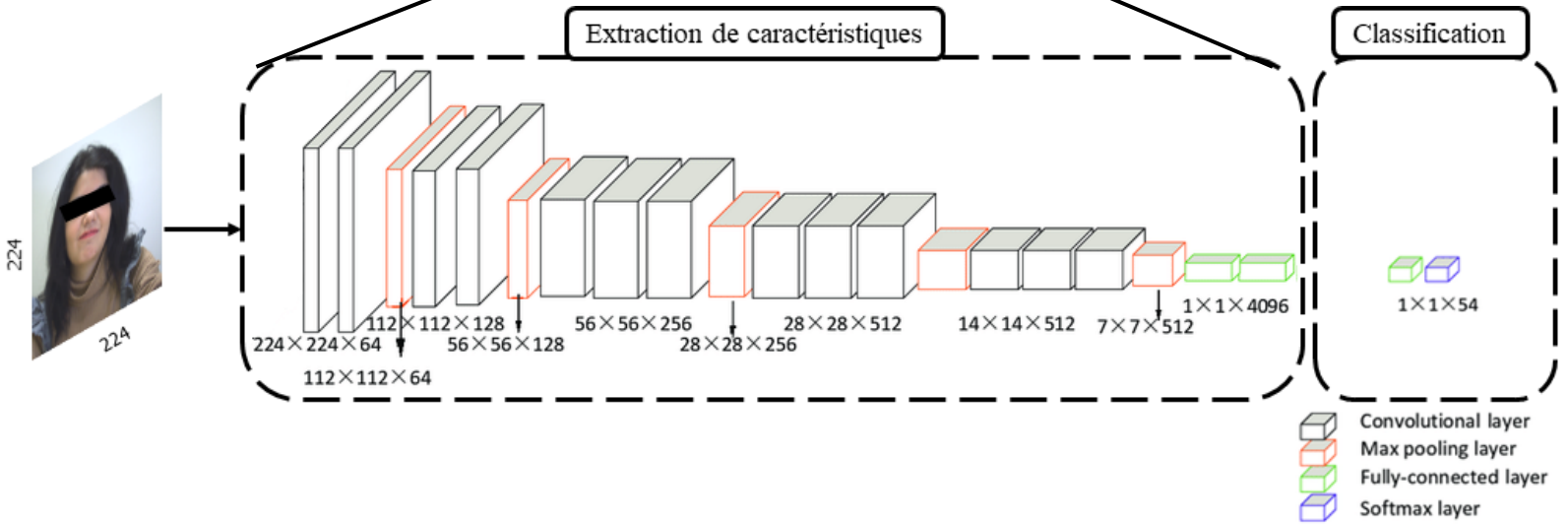
Classification de l'état de stress

Stress

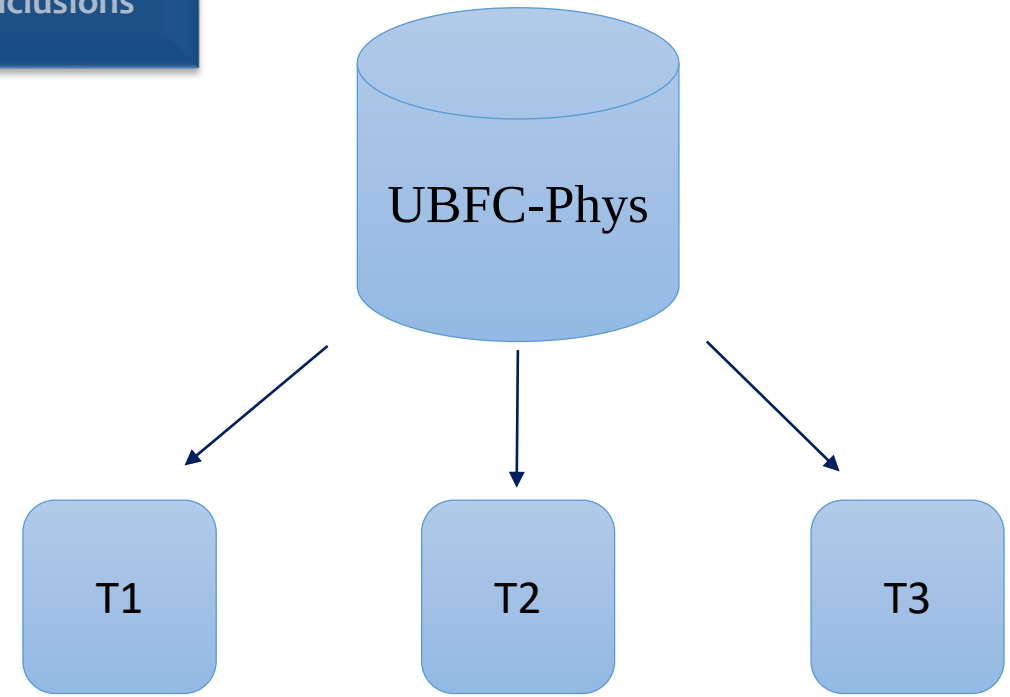
Non-stress

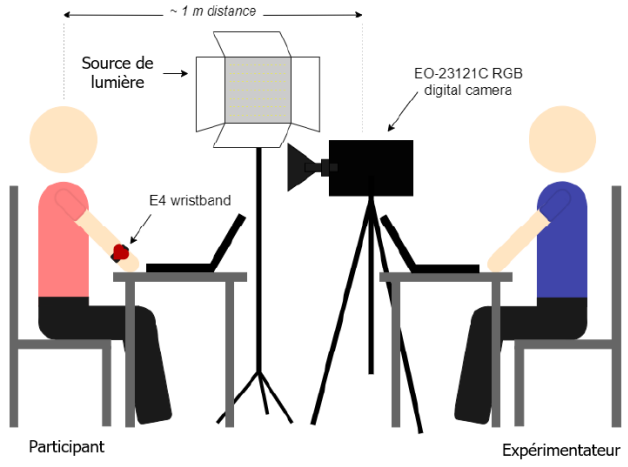
# Pipeline visuel

VGG16 → Caractéristiques faciales → GlobalMaxPooling1D → Couche LSTM → Couche Dense (2) → Stress / Non stress



# Base de données [UBFC-Phys]




- 56 participants (46 femmes et 10 hommes, âgés de 19 à 38 ans)

Séquences vidéo

Signaux physiologiques

- Activité électrodermale EDA
- Signaux PPG

R. Meziati Sabour, et al., "UBFC-Phys: A Multimodal Database For Psychophysiological Studies Of Social Stress", IEEE Transactions on Affective Computing, 2021.

# Reconnaissance du stress à partir des signaux physiologiques

- Résultats obtenus par validation croisée 7-fold en utilisant des algorithmes de ML.

| Caractéristiques | Classifieur | Précision (%) |
|---|---|---|
| iVFC | SVM RBF Kernel | 58,58 |
| | SVM Poly Kernel | 57,61 |
| | NB | 56,92 |
| | RF | 58,58 |
| | kNN | **72,22** |
| VFC | SVM RBF Kernel | 72,74 |
| | SVM Poly Kernel | 74,55 |
| | NB | **78,16** |
| | RF | 58,58 |
| | kNN | 73,64 |

| Caractéristiques | Classifieur | Précision (%) |
|---|---|---|
| iPPG | SVM RBF Kernel | 57,81 |
| | SVM Poly Kernel | 57,81 |
| | NB | 61,82 |
| | RF | **62,40** |
| | kNN | 59,96 |
| PPG | SVM RBF Kernel | 69,72 |
| | SVM Poly Kernel | 58,58 |
| | NB | **72,61** |
| | RF | 66,96 |
| | kNN | 44,22 |

# Reconnaissance multimodale du stress

- Les expressions faciales produisent de meilleurs résultats que les signaux physiologiques.
- La fusion des caractéristiques faciales et des signaux iVFC permet d'obtenir la meilleure précision.

| Caractéristiques | Précision (%) |
|---|---|
| Caractéristiques faciales | 82,48 |
| Caractéristiques faciales + iPPG | 83,12 |
| **Caractéristiques faciales + iVFC** | **91,07** |

# Conclusions

- Développement de la première fusion physio-visuelle en utilisant une seule source d'entrée (vidéos du visage).
- La fusion des caractéristiques faciales avec des signaux physiologiques a amélioré de manière significative la précision.
- La fusion physio-visuelle à partir des vidéos du visage permet à la fois de surmonter le problème des émotions contrefaites et également améliorer la précision en recueillant des informations complémentaires sur l'état affectif de la personne.

# Perspectives

- Effectuer une analyse approfondie pour examiner et explorer les différents facteurs impactant les performances.
- Tester la fusion de caractéristiques à plusieurs niveaux.
- Comparer les performances avec d'autres architectures de l'état de l'art et d'autres modèles de transfer learning.
- Exploiter d'autres modalités telles que le regard, la posture.

# Conclusions générales

- Fusion physio-visuelle par une approche sans contact et mono-capteur.
- Avantages en termes de coût, d'accessibilité et de confort.
- Amélioration de la précision tout en évitant la falsification des émotions.

# Perspectives générales

## Court terme

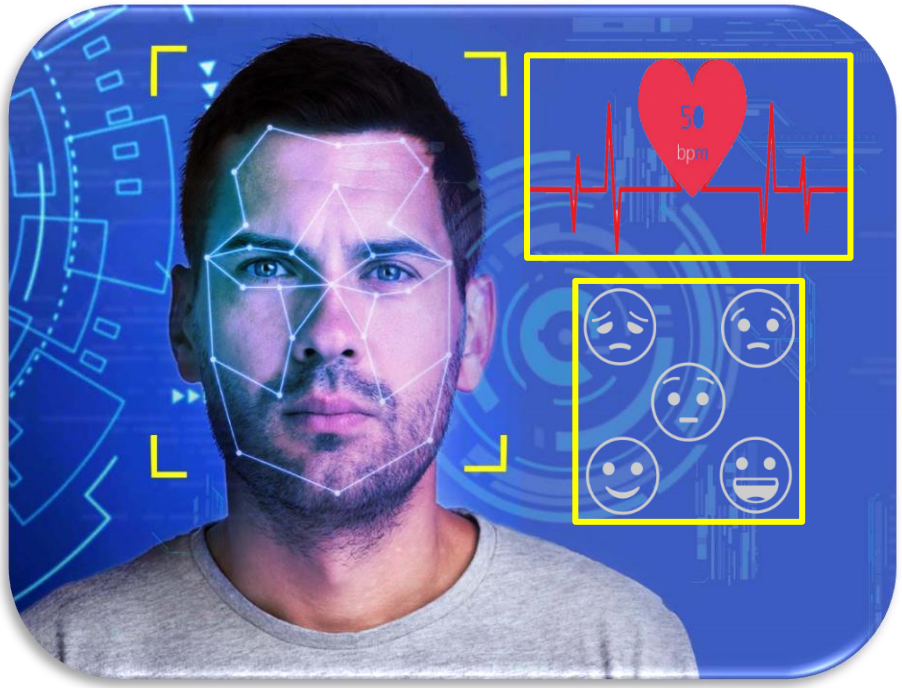
- Réaliser une étude des performances « par ablation ».

## Moyen terme

- Optimiser l'architecture proposées afin de la déployer dans des applications à ressources limitées.

## Long terme

- Développer un système affectif complet exploitant toutes les modalités contenues dans la vidéo.




# MERCI DE VOTRE ATTENTION

yassine.ouzar@univ-lorraine.fr