Reconnaissance Multimodale du Stress

Réalisé par :
Valentina Di Proietto
Alexandre Mondin
Laurent Nguyen
Vincent Pauwels

Projet Fil Rouge

Encadré par :
Prof. Chloé Clavel
Marc Hucelle

Dirigé par :
Aude Schiavi-Courtiade (AFPA)

MS Big Data 2020-2021
Télécom Paris

Table des matières

1 Contexte et enjeux du projet **6**

2 Définition et objectifs du projet **6**

3 Exigences techniques du protocole **7**

4 État de l'art **8**

4.1 Introduction 8

4.2 Vidéo 10

4.2.1 Détection du visage 10

4.2.2 Direction du regard 11

4.2.3 Rotation de la tête 12

4.2.4 Classification des émotions 13

4.2.5 Analyse dans le cadre vidéo 13

4.3 Audio 13

4.4 Textuel 14

4.5 Modèles 15

4.5.1 Préambule 15

4.5.2 Description détaillée des modèles 16

5 Moyens nécessaires **18**

5.1 Répartition des tâches 18

5.2 Ressources (calcul, logiciel, etc.) 19

6 Données **19**

6.1 Description des données 19
6.2 Qualité des données 20
6.3 Déroulement des entretiens 20
6.4 Annotations 21
6.5 Evaluation des annotations 22
6.6 Annotations finales 25

7 Extraction des caractéristiques **27**

7.1 Caractéristiques vidéos 27
7.2 Caractéristiques audios 29
7.2.1 emobase 29
7.2.2 eGeMAPS 29
7.2.3 Mix eGeMAPS et emobase 30
7.3 Caractéristiques textuelles 31
7.3.1 Pré-traitement des données 31
7.3.2 Représentation TF-IDF 32
7.3.3 Représentation LIWC 32
7.3.4 Plongement lexical - fastText 33
7.3.5 Représentation TF-IDF avec plongement lexical 33

8 Modèles mono-modaux **34**

8.1 Aspects méthodologiques 34
8.1.1 Validation croisée 34
8.1.2 Classification vs. Régression 35
8.1.3 Ajout de caractéristiques 35

8.1.4 Critère de sélection 36
8.1.5 Traitement des données 36
8.2 Architecture des modèles non séquentiels mono-modaux audios et vidéos . 36
8.2.1 Première architecture 37
8.2.2 Deuxième architecture 38
8.3 Modèles mono-modaux vidéo non séquentiels 39
8.3.1 Features engineering 40
8.3.2 Résultats en considérant seulement les features connexes aux AUs 40
8.3.3 Résultats en considérant les features connexes aux AUs et les features spatiales 41
8.3.4 Feature importance 41
8.4 Modèles mono-modaux audios 42
8.4.1 Cas 1 : utilisation de l'ensemble des 8 séquences de l'entretien . . . 42
8.4.2 Cas 2 : utilisation des 5 séquences pour lesquelles une réponse orale est attendue 43
8.5 Modèles mono-modaux textuels 44
8.5.1 Prédiction du stress par séquence 44
8.5.2 Prédiction du stress global par entretien 44
8.6 Modèle LSTM Vidéo 45
8.6.1 Architecture 45
8.6.2 Résultats 46

9 Modèles multi-modaux 47

9.1 Architecture 47
9.1.1 *Early fusion* 47
9.1.2 *Late fusion* 48

9.2 Résultats 49
9.2.1 Stress par séquences 49
9.2.2 Stress global 50

10 Conclusion **51**

10.1 Résultats 51
10.2 Problématiques rencontrées 51
10.3 Améliorations possibles 51

Remerciements

Nous tenons à remercier Aude Schiavi-Courtiade et Jean-Michel Gubert pour leur suivi et leur investissement dans la mise en place du protocole et la récolte des données.

Nous tenons également à remercier Chloé Clavel et Marc Hucelle pour leur aide technique et méthodologique qui aura été très précieuse tout au long de ce projet.

1 Contexte et enjeux du projet

L'Agence nationale pour la Formation Professionnelle des Adultes (AFPA) [1] est un organisme français de formation professionnelle, au service des régions, de l'État, des branches professionnelles et des entreprises. L'AFPA propose des formations professionnelles qualifiantes, sanctionnées par un titre professionnel du ministère du Travail.

Le projet des Ressources Humaines ELOCE a pour objectif l'implémentation d'une application permettant l'accompagnement d'un candidat à une formation AFPA, de la phase de prise d'information à la phase de recrutement. Au cours de cette phase de recrutement, le candidat est positionné sur des formations à des métiers, proposées par l'AFPA. Ces formations "classiques" peuvent être complétées par des formations en *soft skills* qui sont déclinés selon les quatre axes suivants : travail en équipe, adaptabilité, gestion du stress et réflexivité.

Le projet proposé par l'AFPA est la mise en place d'une application innovante qui vise à assister les potentiels candidats durant leur parcours de recrutement ELOCE. Plusieurs équipes interviennent sur le projet.

Ainsi, au cours des premières semaines du projet, nous avons circonscrit le périmètre de notre intervention en décidant de nous concentrer sur l'identification des besoins en formations *soft skills*, en particulier sur la gestion du stress. Nous avons alors conçu un projet d'intelligence artificielle pour détecter le niveau de stress d'un candidat à partir d'une mise en situation filmée. A partir de la vidéo de cette mise en situation, nous pourrons extraire des attributs vidéos, audios et textuels. Ces derniers seront utilisés pour entraîner des algorithmes permettant de mesurer le niveau de stress du candidat.

2 Définition et objectifs du projet

Évaluer la gestion du stress d'un candidat passe par l'identification du niveau de stress de celui-ci. Une telle prédiction est possible par l'utilisation d'algorithmes de machine learning de classification supervisé. L'algorithme apprend à reconnaître le stress chez une personne en combinant plusieurs données obtenues de la façon suivante : on soumet un premier questionnaire à choix multiples (QCM) au candidat, on lui propose ensuite un exercice filmé face caméra (interaction homme-machine). Un dernier QCM est soumis au candidat (Figure 1). On combine alors les features (audio/vidéo/texte) observées pendant l'exercice. Enfin, on prédit le niveau de stress du candidat grâce à un algorithme entraîné sur des annotations d'un expert de l'AFPA concernant le stress des participants.

1. `https://fr.wikipedia.org/wiki/Agence_nationale_pour_la_formation_professionnelle_des_adultes`

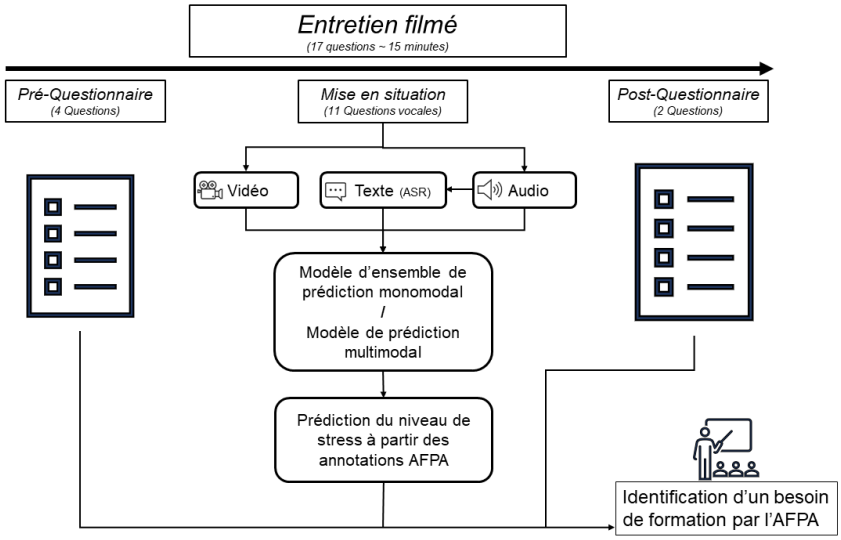


Figure 1 – Fonctionnement du système

Les deux questionnaires écrits (Pré & Post Questionnaires) contiennent des questions directement liées au stress du candidat. Cette mesure, qui révèle des informations importantes, peut être biaisée lorsqu'elle est considérée toute seule—par exemple—par la volonté de la personne de dissimuler certains aspects de sa personnalité.

Par conséquent, nous avons décidé de ne pas prendre en compte les informations des questionnaires pour entraîner les algorithmes. Nous nous baserons sur l'analyse des features vidéos, audios et textuelles obtenues en observant la personne pendant le déroulement de l'exercice. Le niveau de stress peut être aussi mesuré à l'aide de capteurs portables de données physiologiques. Cependant, de tels capteurs ne seront pas accessibles lors de la mise en situation. Notre approche se concentre donc sur le niveau de stress "apparent" du candidat. De ce fait, notre méthode est potentiellement applicable à d'autres applications ne nécessitant pas de capteurs physiologiques. La production de vidéo est aujourd'hui une pratique répandue et aisément accessible à beaucoup d'organisations. Par conséquent la quantité de données disponibles fait entrevoir l'analyse des vidéos comme une manière directe et aisément applicable d'identifier le niveau de stress.

3 Exigences techniques du protocole

Dans le cadre de l'implémentation de notre algorithme multimodale, nous avons défini des besoins concernant les vidéos des entretiens :

— présence de questions qui nécessitent la participation orale du candidat ;

— homogénéité des tests soumis aux candidats (plus il sera variable dans sa forme, moins nous pourrons comparer les réactions des candidats) ;

— enregistrements vidéos et audios de qualité suffisante durant toutes les phases de l'entretien. Les webcams standards sont généralement suffisantes. Le candidat devra se placer face caméra ;

— annotations humaines et manuelles par un expert des niveaux de stress (cas supervisé) durant les phases de l'entretien en vue d'apprendre le modèle ;

— réponses des utilisateurs aux questions des première et dernière phases de l'entretien (format csv, xml, txt, ...).

4 État de l'art

4.1 Introduction

Notre lecture de l'état de l'art s'est portée sur la détection automatique des émotions à travers plusieurs modalités (audio, vidéo, textuelle) au cours d'entretiens ou de présentations, et en particulier du stress lorsque cela était disponible. Les articles lus traitant du stress sont pour l'instant limités mais c'est bien sur ceux-ci que nous nous concentrerons par la suite pour affiner notre connaissance de l'état de l'art en la matière. Notre travail se base principalement sur l'impression des émotions perçues dans la mesure où nous observons uniquement le comportement apparent des candidats.

La littérature répond en général à deux types de besoins actuels : la volonté de l'industrie du recrutement de (pré)sélectionner des candidats par interaction digitale et celle, plus philanthrope, de mettre à disposition des candidats et orateurs des pistes d'améliorations sur l'impression qu'ils génèrent à travers leur comportement apparent. Ce dernier aspect est d'autant plus crucial que la vidéo est un support digital de plus en plus utilisé dans la mise en avant de ses qualités de communication.

C'est dans le cadre de ces deux problématiques que [Hemamou et al., 2019] propose un nouveau réseau de neurones profond (**HireNet**) qui vise à prédire l'employabilité de candidats par une analyse multimodale. Il se distingue des autres parutions scientifiques par la taille importante des données d'entretien utilisées (plus de 7000 entretiens) et leur nature (entretiens authentiques).

À partir de données annotées d'entretien d'embauche, [Finnerty et al., 2016] s'intéresse à la possibilité de prédire le stress physiologique (plutôt qu'apparent) d'une personne en utilisant l'activité électrodermale (EDA : electrodermal activity) ; cela s'est avéré non concluant, contrairement aux features apparentes non verbales (audios et visuelles). Cet article, le seul de notre corpus qui s'intéresse directement au stress, a par ailleurs mis en avant une corrélation négative entre l'employabilité et le niveau de stress

perçu, ce qui était intuitivement attendu.

Le jeu de données de cet article a été aussi utilisé dans [Nguyen, 2015], qui traite de l'extraction automatique des caractéristiques du comportement (verbales et non-verbales) pour l'entraînement de modèles multimodaux. Deux études supervisées ont été menées : une pour la prédiction des impressions d'employabilité par régression à partir des entretiens d'embauche (sur les vidéos complètes ainsi que de courts extraits) et l'autre pour la prédiction de comportements sociaux (compétences, personnalité, employabilité) à partir de vidéos de la plateforme YouTube.

Le papier [Escalante et al., 2020] s'est aussi basé sur des vidéos disponibles sur la plateforme YouTube pour explorer les aspects d'explicabilité et d'interprétabilité des modèles d'apprentissage statistique permettant de prédire l'employabilité ainsi que les traits de personnalité apparents (la personnalité intrinsèque étant difficile à déterminer) d'une personne via les 5 traits habituellement étudiés dans la littérature (les "*Big Five*" :*Openness*, *Conscientiousness*, *Extroversion*, *Agreableness*, *Neuroticism*). Il résume les résultats d'une coopétiiton (coopération/compétition) sur des données de clips de vidéos Youtube dont le but est de mettre à disposition des personnes à la recherche d'emploi les ressorts des décisions des interviewers (et donc de mettre en exergue les potentiels biais) afin d'améliorer leur chance d'être invité à un entretien professionnel. Les données mises à disposition des "coopétiteurs" avaient fait l'objet de traitements et d'analyses préliminaires par les organisateurs et étaient fournies par modalité : textuelle d'un côté et sensorielle de l'autre (audio et vidéo).

Concernant les biais, [Acharyya et al., 2020] traite en particulier de ceux liés aux attributs ethniques ou de genre dans la notation (*rating*) de conférences en ligne (TED *talks*). Les auteurs proposent un modèle mathématique fondé sur la théorie des modèles causaux, la *Counterfactual Fairness* et les *Neural Language models* pour prédire de manière équitable et juste la qualité des présentations orales afin d'assurer une diversité des intervenants TED et des notations (*ratings*) objectives et débiaisées pour les spectateurs. Les données utilisées pour cette étude ont été tirées du site TED et sont constituées de 3 parties pour chaque présentation : le nombre de vues, la transcription de la présentation ainsi que la notation (*rating*) selon 14 labels.

D'autres études ont aussi pris comme données des conférences en ligne du site TED, par exemple [Wu and Qu, 2020] pour proposer un outil d'analyse du contenu multimodal (verbal et non-verbal) selon trois niveaux de vue (*Projection, Comparison and Video views*).

Enfin, nous nous sommes aussi penchés sur des comportements visuels très particuliers, traités dans [Mora and Odobez, 2016], tel que la prédiction de la direction du regard en fonction du visage complet fourni en entrée. Cette tâche contient deux directions : l'estimation du vecteur du regard (en 3D) est généralement utilisée dans le domaine de la sécurité automobile tandis que celle de la position du regard (en 2D) per-

met d'utiliser le point du regard pour contrôler un curseur dans le cadre de l'interaction homme-machine.

4.2 Vidéo

Les candidats à l'entretien vidéo de l'AFPA se tiennent face caméra. Ainsi, on s'intéresse principalement aux émotions du visage. Dans le domaine de la reconnaissance des émotions faciales, la plupart des approches se divisent en 2 étapes : en premier lieu il convient de détecter le visage et les points d'intérêts (bouche, yeux, sourcils...), puis, on utilise le plus souvent des réseaux de neurones à convolution (CNN) pour détecter les émotions qui découlent de l'analyse du visage.

4.2.1 Détection du visage

Pour détecter les points d'intérêts, le FACS (*Facial Action Coding System*) est un système développé par Eckman et Friesen pour l'analyse des micro-expressions sur des visages humains et déterminant l'émotion des sujets analysés. Il est basé sur l'idée que chaque émotion peut être associée à des schémas musculaires faciaux différents et qu'en analysant les régions où ces muscles sont activés, nous pouvons déterminer l'émotion de l'individu. La localisation du visage dans une image doit être indépendante de toute occlusion dans la scène, des variations des conditions d'éclairage et doit tolérer les changements de pose du visage.

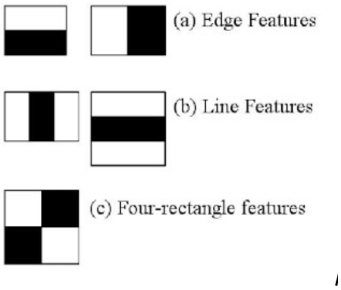
— **Classifieurs en cascade de Haar** : Rapide, mais relativement imprécis. Il peut être difficile de régler les paramètres.
— **HOG + SVM linéaire** : Généralement plus précis que les classifieurs en cascade de Haar, mais souvent plus lent.
— **CNN** : Plus précis et robustes que les algorithmes précédents lorsqu'ils sont correctement entrainés. Peuvent être très lents selon la profondeur et la complexité du modèle. Peut être accéléré en utilisant un GPU.

Pour relever ce défi, une des méthodes les plus utilisées pour la détection du visage et de ses sous-parties (yeux, nez, bouche..) est le *Haar Cascade Classifier*[2].

La détection d'objets à l'aide de classifieurs en cascade basés sur les caractéristiques Haar est une méthode efficace de détection d'objets proposée dans [Viola and Jones, 2001]. L'algorithme se décompose en 4 étapes :

2. `https://docs.opencv.org/3.4/db/d28/tutorial_cascade_classifier.html`

— **Calculer les features de Haar** : Chaque feature est une valeur unique obtenue en soustrayant la somme des pixels sous le rectangle blanc de la somme des pixels sous le rectangle noir.



— **Calculer l'intégrale des images pour réduire le temps de calcul des features de Haar** [3].

— **Entraîner un classifieur (Adaboost) pour déterminer les meilleures features pour la détection d'un visage** : L'inconvénient de cette étape est qu'elle requiert un nombre d'images positives et négatives très élevé. Heureusement, la librairie OpenCV possède des modèles déjà entraînés pour cette tâche.

— **Utiliser le classifieur en cascade** : Plutôt que de vérifier pour chaque fenêtre dans l'image si c'est un visage ou non à l'aide de toutes les features obtenues, on va grouper les features. Si la fenêtre ne reconnaît rien d'un visage avec le premier groupe de features, on passe à la fenêtre suivante sans considérer les autres features. Sinon, on vérifie le second groupe de features. Lorsque la fenêtre reconnaît un visage dans chaque groupe de features, cela correspond à la zone du visage.

  Étant donné que la majorité de l'image ne correspond pas à la zone du visage, cette décomposition en groupe de features permet de réduire considérablement le temps de calcul. De ce fait un des aspects importants de l'entraînement du modèle est la minimisation des faux négatifs.

4.2.2 Direction du regard

Pour l'estimation de la direction du regard, les méthodes sont très variées et la recherche est loin d'être arrêtée. Le papier [Kar and Corcoran, 2017] offre un panorama des différentes techniques. L'angle de déviation semble être la principale métrique d'évaluation, bien que cet article mette en avant le manque d'homogénéité des métriques utilisées. De plus, les systèmes d'interactions homme - machine basés sur le regard, donc relativement similaires à notre projet, sont capables d'opérer à grande vitesse, ce qui conduit à leur mise en œuvre dans une variété de plateformes et d'applications.

3. `https://datasciencechalktalk.com/2019/07/16/haar-cascade-integral-image`

Un des problèmes de l'estimation du regard provient de la connaissance des caractéristiques techniques liées à la webcam, qui aident grandement à la prédiction. La plupart des webcams sont limitées à de petites fréquences d'images et il est difficile de trouver les spécifications réelles, comme la taille des capteurs et les distances focales, pour de nombreuses webcams. Cependant, nous pourrions avoir accès à ces spécifications exactes de l'AFPA dans le cadre de notre application.

Le papier [Mora and Odobez, 2016] s'intéresse à la position du regard ainsi qu'à la variation de position de la tête, deux tâches intrinsèquement liées. Il propose d'utiliser des caméras RGB-D pour capturer la profondeur de l'image. Ils indiquent que ces capteurs ont permis aux chercheurs de traiter des problèmes connus pour être très difficiles lorsqu'ils sont basés sur la seule vision standard, comme l'estimation de la pose du corps ou la reconnaissance des expressions faciales. Grâce à des cartes de profondeur (D), ces capteurs fournissent des mesures explicites et fiables de la forme de la scène, par opposition à la forme implicite des informations intégrées dans les données RGB. Ils notent qu'il est toujours difficile et coûteux de déduire des informations de forme à partir du domaine visuel (RGB) seul.

Bien que nous n'ayons probablement pas accès à de tels capteurs avec l'AFPA, l'article décrit les deux principales approches en prédiction du regard :

— **Geometric Based Method** : Cette méthode repose sur la détection de caractéristiques locales qui sont converties en paramètres du regard. La plupart des méthodes nécessitent une séance d'étalonnage pour recueillir des échantillons annotés du regard. Ceux-ci sont utilisés pour déterminer des paramètres spécifiques à l'utilisateur décrivant la géométrie du globe oculaire ou une cartographie directe au point d'observation.

— **Appearance Based Method** : Cette méthode consiste à utiliser l'image de l'oeil entier pour déterminer les paramètres du regard. Cette approche permet d'éviter le suivi fastidieux des caractéristiques locales, ce qui offre la possibilité de détecter le regard même à faible résolution.

4.2.3 Rotation de la tête

La direction du regard est très liée à la position de la tête. Ainsi, dans l'article [Mora and Odobez, 2016] les chercheurs développent leur propre méthode à l'aide d'une caméra RGB-D.

Par ailleurs il existe une méthode associant les frameworks OpenCV et DLib permettant d'estimer la position de la tête. La méthode s'appuie principalement sur deux features à extraire. La première feature est un ensemble de coordonnées 2D (les coins des yeux, le bout du nez, les coins de la bouche, etc.) pouvant être extraites à l'aide

de l'outil Dlib's *facial landmark detector*. La seconde feature est l'ensemble de coordonnées 3D correspondant aux coordonnées 2D. L'article [Mora and Odobez, 2016] utilise une caméra RGB-D pour avoir accès à ces données, mais elles peuvent cependant être inférées à l'aide d'outils de la librairie OpenCV.

Une fois ces features extraites, il est nécessaire de connaître les paramètres de la caméra (distance focale, centre optique dans l'image et paramètres de distorsion radiale). On peut contourner ce problème en approchant le centre optique par le centre de l'image et en approximant la distance focale par la largeur de l'image en pixels. Enfin, on suppose que la distorsion radiale n'existe pas. Dans le cadre de notre projet, nous allons nous entretenir avec l'AFPA pour récupérer ces informations. Nous ferons les approximations citées le cas échéant. Enfin, la librairie OpenCV implémente des algorithmes d'estimation de la position de la tête reposant sur ces features.

4.2.4 Classification des émotions

Les classifieurs les plus utilisés dans la classification d'image sont les CNN (*Convolutional Neural Networks*) et les SVM (*Support Vector Machine*). Cependant la plupart des travaux récents ont montrés que les CNN sont souvent plus performants ques les SVM dans la classification d'image ([Hasan et al., 2019]).

4.2.5 Analyse dans le cadre vidéo

Les démarches précédentes sont applicables à une image. Cependant, notre travail se base sur des vidéos d'interviews. Une vidéo n'étant qu'une suite d'image à un frame-rate défini, l'approche la plus utilisée est la suivante :

— Analyser la vidéo image par image. On pourra lisser le frame-rate des vidéos si nécessaire ;

— Réduire la dimension (ex : filtre de niveau de gris) ;

— Identifier le visage et ses sous-parties ;

— Appliquer un algorithme de prédiction des émotions (le stress dans notre cas) sur le visage identifié.

4.3 Audio

Les deux principaux groupes de features audios utilisés sont la prosodie et le Mel-Frequency Cepstral Coefficients.

Prosodie Il s'agit de l'ensemble des traits oraux d'une expression verbale d'un locuteur, traduisant la musicalité de sa voix et de ses énoncés, et qui rend les émotions et les intentions plus intelligibles à ses interlocuteurs [Nguyen, 2015]. Le volume de la voix, le timbre ou « coloration », et le débit vocal déterminent différentes composantes de la prosodie : intonation et ton, accentuation et accent marqués par des modulations et inflexions prosodiques, et rythme (vitesse d'élocution, caractérisée par la pause silencieuse et la pause nourrie, et le tempo). [4] Les indicateurs sont généralement extraits sur la base de Hidden Markov Models (HMM) qui segmentent les signaux audios en des régions paroles/non paroles et voix/non voix.

Mel-Frequency Cepstral Coefficients (MFCC) MFCC est un ensemble de caractéristiques acoustiques utilisées dans la reconnaissance d'émotions dans la parole [Finnerty et al., 2016]. Ils constituent des coefficients cepstraux [5] calculés par une transformée en cosinus discrète appliquée au spectre de puissance d'un signal. Les bandes de fréquence de ce spectre sont espacées logarithmiquement selon l'échelle de Mel. [6]

4.4 Textuel

Dans l'ensemble des articles scientifiques lus, les transcriptions des entretiens et présentations sont utilisées pour générer des features de granularité différente.

Classification du texte en mode de présentation Dans [Wu and Qu, 2020] on s'intéresse simplement aux techniques de présentation verbales en classant chaque segment de vidéo de 1 minute selon 3 modes : narration, exposition et argumentation. Aucun traitement sémantique de la transcription n'est effectué.

Extraction de métriques globales du texte Des métriques caractérisant le nombre de mots total, le nombre de mots par phrase, le nombre de mots supérieurs à 6 lettres etc. ont été extraites en tant que features dans [Nguyen, 2015].

Classification du texte en catégories psychologiques Dans [Nguyen, 2015], le texte est aussi représenté par le pourcentage de catégories psychologiques évoquées dans

4. `https://fr.wikipedia.org/wiki/Prosodie`

5. Le cepstre d'un signal $x(t)$ est une transformation de ce signal du domaine temporel vers un autre domaine analogue au domaine temporel. Le cepstre a tout d'abord été défini en 1963 comme étant le résultat de la transformée de Fourier appliquée au logarithme naturel de la transformée de Fourier du signal dont la phase est ignorée.

6. `https://fr.wikipedia.org/wiki/Cepstre`

la transcription en fonction d'un mapping des mots vers des catégories psychologiques selon le dictionnaire LIWC (*Linguistic Inquiry Word Count*).

Plongement syntaxique (*Word embedding*) Il s'agit de la représentation la plus complexe de la modalité textuelle. Dans [Escalante et al., 2020], les données textuelles correspondaient aux transcriptions des vidéos et avaient été pré-processées par une régression Ridge prenant en entrée les transcriptions après plongement lexical (*word embedding*) selon deux modèles différents (*bag-of-words model* et *skip-thought vectors model*). Le *bag-of-words model* utilise un plongement lexical qui représente les transcriptions comme des vecteurs de dimension 5000, c'est-à-dire les occurrences des 5000 mots hors mots vides (*stop words*) les plus fréquents dans les transcriptions. Le *skip-thought vectors model* utilise quant à lui un plongement lexical qui représente les phrases des transcriptions comme des *skip-thought* vecteurs moyens de dimension 4800. La représentation en tant que phrase ajoute une difficulté supplémentaire (séquence de longueur indéterminée) et a utilisé un réseau de neurones pré-entraîné pour extraire les *skip-thought* vecteurs des transcriptions.

En ce qui concerne [Acharyya et al., 2020], les transcriptions ont été obtenues en utilisant l'implémentation doc2vec de la librairie Gensim (Python) avec des vecteurs de dimension 200.

4.5 Modèles

4.5.1 Préambule

Dans la littérature sur la détection multimodale des émotions, plusieurs modèles sont utilisés. Une première distinction peut être faite entre les modèles avec un output binaire (par exemple *candidat employable* ou *candidat pas employable*) et les modèles qui donnent un score (par exemple un nombre de 1 à 5 qui mesure l'employabilité ou un trait de la personnalité).

Dans les cas où les données disponibles ne sont pas nombreuses, ([Nguyen, 2015]), des algorithmes de régression (régression linéaire simple, ou avec une régularisation de type Ridge ou Lasso) sont utilisés. Les mêmes algorithmes sont aussi appliqués en cas de données plus nombreuses ([Escalante et al., 2020]), et même si les performances sont inférieures aux performances des algorithmes plus complexes, ils ont donné d'intéressants résultats en termes d'explicabilité. De même, un bon résultat en termes d'explicabilité, avec un très bon score en termes de performances, est obtenu en combinant des classifieurs *Extreme Learning Machine* avec des Forêts Aléatoires ([Kaya et al., 2017]), toujours sur les mêmes données très nombreuses. Une Forêt Aléatoire toute seule est utilisée aussi lorsqu'il y a moins de données, toujours avec de bonnes performances (deuxième

partie de [Nguyen, 2015]).

Enfin, dans l'analyse de vidéos, surtout si la longueur est conséquente, le facteur temporel doit être pris en compte : par exemple, un mouvement de la tête peut avoir différentes significations en fonction de quand il a lieu pendant la vidéo. Un réseau de neurones profond efficace et qui tient compte de l'aspect séquentiel du problème est développé dans [Hemamou et al., 2019] et est entraîné sur un grand nombre de données.

4.5.2 Description détaillée des modèles

Dans ce paragraphe nous allons analyser en détails les modèles utilisés dans la littérature.

Dans [Nguyen, 2015], l'auteur analyse 62 vidéos d'entretiens d'embauche, qui ont été annotées manuellement. La variable à prédire est l'employabilité (*hireability*) du candidat. Celle-ci est décrite par cinq caractéristiques, notées généralement sur une échelle variant de 1 à 5 : communication, persuasion, conscience, résistance au stress, décision d'embauche. Les algorithmes utilisés pour prédire ces variables sont la régression linéaire, la régression Ridge et la Forêt Aléatoire (*Random Forest*) qui ont été appliquées selon trois configurations différentes : à toutes les features, aux features réduites grâce à l' ACP (Analyse en Composantes Principales), aux features réduites grâce à la méthode *low p-value*. Les modèles de régression ont été entraînés avec la méthode de cross validation leave-one-out (qui consiste à utiliser uniquement un échantillon dans la partie test) et les paramètres ont été estimés avec une validation croisée à 10 folds. Le meilleur résultat de prédictions est obtenu avec la régression Ridge sur toutes les features et sur les features réduites avec l'ACP ($R^2 = 0.36$ dans les deux cas). La régression Ridge a été utilisée aussi pour prédire la décision d'embauche avec des sous-ensembles différents de features (seulement les features vidéo, –ou audio ...) et en considérant des courts extraits de l'entretien.

Le même dataset a été exploité dans le papier [Finnerty et al., 2016] où 3 évaluateurs indépendants ont noté le niveau de stress (score de 1 à 5) pour chaque question posée pendant l'entretien et ont aussi donné une note sur le niveau de stress global. Les algorithmes appliqués étaient la régression Ridge et la Forêt Aléatoire avec l'emploi de la méthode de cross-validation leave-one-out pour partager les données en train et test. Pour le choix des paramètres, les auteurs ont appliqué une validation croisée avec 10 folds. Le meilleur résultat a été obtenu avec toutes les features et le Forêt Aléatoire ($R^2 = 0.195$), et le deuxième meilleur score ($R^2 = 0.190$) a été obtenu avec la régression Ridge en considérant seulement les features video auxquelles on a appliqué une réduction de dimension avec la méthode low p-value ($p < 0.05$).

Dans la deuxième partie de [Nguyen, 2015], l'auteur travaille sur beaucoup plus

de données : 900 vidéos, avec une note donnée par les employés d'Amazon Mechanical Turk sur plusieurs compétences (*hireability*, *professional skills*, *communication skills* et *social skills*). Les méthodes utilisées sont presque les mêmes que dans la première partie. Comme algorithmes, l'auteur applique la régression Ridge, la Forêt Aléatoire et aussi la régression Lasso. Dans ce cas, une validation croisée a été mise en place avec 10-folds afin d'estimer les paramètres ; il a été décidé que 90% des données feraient partie du jeu d'entraînement et 10% du test. Les meilleurs résultats ont été obtenus avec la Forêt Aléatoire sur toutes les features avec un score R^2 de 0.27. Les performances sont beaucoup moins bonnes lorsque l'on considère un groupe de features seulement.

Un modèle plus complexe à été développé dans [Hemamou et al., 2019] ; grâce à un corpus de plus de 7000 vidéos d'entretiens d'embauche mises à disposition par une entreprise, les auteurs ont construit un réseau de neurones profond pour prédire l'employabilité d'un candidat. La prédiction, dans ce cas, est binaire : employable ou non employable. L'entretien est modélisé comme une séquence de $n + 1$ éléments : l'intitulé du poste suivi de n couples de questions-réponses. L'intitulé du poste et les questions sont des séquences de mots, tandis que les réponses sont des données plus riches, puisqu'elles se composent également des descripteurs de niveau inférieur (audio, vidéo ou texte). Les descripteurs ont été encodés avec un GRU (*Gated Recurrent Unit*), encodeur qui est capable de traiter les séquences. Des mécanismes d'attention sont introduits pour donner un poids aux questions les plus importantes en relation au type de travail. Les données ont été partagées en train (80%), validation (10%) et test (10%). Le score $F1$ est utilisé comme métrique d'évaluation. La performance est évaluée pour chaque modalité audio, visuelle et textuelle. Le réseau ainsi construit produit de meilleurs résultats ($F1 = 0.64$, pour le textuel et l'audio) que les modèles non-séquentiels (SVM, Ridge et Forêt Aléatoire). Il est aussi meilleur que des modèles qui ne tiennent pas compte de la hiérarchie des descripteurs, des mécanismes d'attention, ou du contexte (intitulé du poste et questions). Les trois versions monomodales (audio, vidéo et texte) du réseau ont été fusionnées de deux manières différentes (à la dernière couche ou en faisant une moyenne sur l'output de chaque modalité) et le meilleur score ($F1 = 0.645$) a été obtenu dans le deuxième cas.

Le papier [Escalante et al., 2020] porte sur un challenge où on a demandé aux participants de prédire le 5 traits de la personnalité d'un candidat (*Openness*, *Conscientousness*, *Extroversion*, *Agreableness*, *Neuroticism*) et, en relation à ça, ses possibilités d'être embauché. Le challenge consistait en deux parties : la partie quantitative (la performance du modèle) et l'explicabilité (la production d'un texte qui explique pourquoi le modèle a décidé pour l'employabilité ou pas du candidat). Le dataset était composé de 10000 clips de 15 secondes de CV-videos en anglais. Les vidéos avaient été notées par les employés d'Amazon Mechanical Turk. Pendant le challenge, deux modèles en particulier ont été remarqués et, dans le papier, ils sont décrits en détails.

Le premier, appelé **BU-NKU**, est détaillé aussi dans [Kaya et al., 2017], et on va en esquisser les points principaux. Après avoir extrait les features audio et vidéo, on

utilise les kernel extreme learning machines (ELM) (un réseau de neurones, avec ELM au lieu de la back-propagation) pour obtenir des prédictions qui sont alors passées à une Forêt Aléatoire. Le résultat est un score pour chaque trait de personnalité ainsi que pour le niveau d'employabilité. À la fin du procès, un texte est produit, dans lequel la décision sur l'employabilité est expliquée en rapport aux valeurs de chaque trait de la personnalité. En plus, le texte final souligne le groupe de features qui ont déterminé les valeurs sorties par le modèle (par exemple, facial features). Pour l'évaluation quantitative du modèle, on a utilisé le Mean Absolute Error, qui a donné des résultats autour de 0.92.

L'autre modèle, le **TUD**, est aussi présenté en détail dans le papier ; il s'agit d'un modèle beaucoup plus simple du précédent, avec performances plus basses, mais une bonne explicabilité. Les features considérées dans **TUD** sont de type vidéo et texte et elles sont regroupées en quatre ensembles : *OpenFace* (vidéo), *face movement* (vidéo), *readability* (texte) et *text representation* (texte). Pour chaque ensemble, on applique la PCA retenant 90% de la variance et aux variables retenues, on applique une simple régression linéaire. Enfin, les scores de chaque groupe ont été fusionnés en un score pour chaque trait de la personnalité ainsi que pour l'employabilité. Comme annoncé, le Mean Absolute Error n'est pas le meilleur du challenge, puisqu'il a une valeur autour de 0.88. Par contre, le texte produit pour expliquer le modèle est plus précis et exhaustif que dans **BU-NKU** : pour chaque personne, le texte reporte si le score obtenu est inhabituel par rapport aux personnes avec des caractéristiques similaires et donne aussi une description des features qui ont contribué au maximum à la prédiction.

5 Moyens nécessaires

5.1 Répartition des tâches

AFPA L'AFPA a défini le contenu de l'entretien permettant d'évaluer le niveau de stress du candidat :

— Le questionnaire est proposé via une interface digitale avant et après la mise en situation ;

— La mise en situation elle-même.

Une fois cette tâche réalisée, l'AFPA s'est chargée de faire passer les entretiens à trente-et-un candidats (stagiaires en formation, formateurs) afin de constituer les enregistrements nécessaires à l'entraînement de notre algorithme multimodal de détection du stress.

Étudiants Les tâches sont réparties selon 2 axes :

— L'extraction technique des features via les 3 modalités : vidéo, audio et texte.

  - Features Audio : Laurent s'est chargé de l'extraction des features audio.
  - Features Video : Vincent et Valentina se sont occupés de l'extraction des features video.
  - Texte : Alexandre s'est chargé du traitement des features textuelles.
- L'analyse de performance et sélection du modèle alimenté par ces features. Une approche multimodale et/ou ensembliste a été étudiée afin de trouver la meilleure combinaison de modèles entre les différentes modalités (étude des modèles standalone et agrégés).

Les parties les plus denses étaient celles de l'analyse de modèle : elles ont nécessité donc un appui de la part de l'ensemble des participants du projet.

5.2 Ressources (calcul, logiciel, etc.)

Concernant les calculs, nous avons utilisé nos ordinateurs personnels.

Afin d'exploiter des features textuelles, nous avons eu besoin d'une retranscription textuelle des vidéos. Nous avons utilisé des programmes d'*Automatic Speech Recognition* (ASR), en particulier la version gratuite de l'ASR de Google.

6 Données

6.1 Description des données

L'AFPA a fait passer le test à 31 personnes, résultant en 31 entretiens d'une durée totale cumulée de 7h21 d'enregistrement, avec une durée moyenne de 14 minutes. Il s'agit d'une quantité de vidéos significative. En effet, à titre de comparaison, le challenge MuSe 2021 (Multimodal Sentiment Analysis ChallengeSE) propose 35 heures de vidéos (MuSE-CaR) dont 5h47 spécifiquement sur le stress (Ulm-TSST).

Les entretiens concernent principalement des stagiaires de deux centres de l'AFPA ainsi que quelques formateurs. On remarque une proportion plus importante de personnes de genre féminin avec 20 femmes et 11 hommes.

Il est intéressant de noter que, en raison des mesures sanitaires en vigueur, les entretiens ont été tournées en majorité aux domiciles des personnes, avec leurs matériels, entraînant des vidéos de résolutions différentes, des qualités sonores et lumineuses hétérogènes ainsi que des arrière-plans variés.

6.2 Qualité des données

Une première revue des données a permis de constater une qualité suffisante pour 29 entretiens. Cependant, deux entretiens présentent les anomalies suivantes :

— le cadrage du visage pour une entretien n'était pas suffisant, d'où un visage détecté à seulement 47% par OpenFace ;

— une entretien incomplète pour laquelle il manque les trois dernières questions.

6.3 Déroulement des entretiens

Les candidats sont soumis à 23 séquences défilantes (automatiquement ou par interaction).

Les 7 premières séquences et les 3 dernières sont deux questionnaires de l'AFPA où les candidats répondent par écrit. Notre analyse vidéo se concentre sur les 13 séquences restantes, qu'on regroupera en 11 pour effectuer les annotations.

Durée théorique de chaque séquence (sec)											
D1-7	D8	D9	D10	D11	D12-14	D15	D16	D17	D18	D19	D20-23
33	50	100	66	97	62	63	28	31	42	39	74

Table 1 – Longueur supposée des séquences

Certaines réponses aux questions sont écrites et d'autres sont orales. Les participants sont enregistrés durant toute la durée de l'exercice et parlent souvent pendant les questions nécessitant pourtant seuelement une réponse écrite.

Voici la liste des questions :

— D8 - En moins d'une minute, pouvez-vous vous présenter et nous expliquer vos motivations pour cette formation ? (réponse orale)

— D9 - Pouvez-vous développer un aspect de votre parcours en lien avec cette formation ? (réponse orale)

— D10 - Quelles sont vos motivations pour cette formation ? (réponse orale)

— D11 - Que savez-vous du métier auquel vous voulez vous former ? (réponse orale)

— D12-14 - Mémoriser une série de chiffres et de lettres affichée à l'écran (réponse écrite)

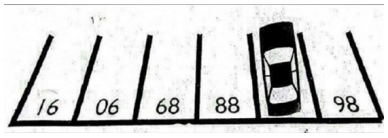
— D15 - Mémoriser une liste de 10 mots (réponse écrite)

— D16 - Que voyez-vous dans l'image ? (réponse écrite)

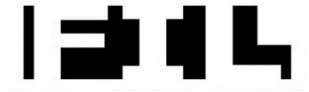
— 

— D17 - Lire une liste de mots écrits en couleur ex : [bleu, vert , rouge, vert, jaune, noir ..] (réponse orale)

— D18 - Indiquez la place de parking cachée par la voiture (réponse écrite)

— 

— D19 - Que lisez-vous ? (réponse écrite)

— 

— D20 - Un escargot cherche à atteindre le sommet d'un poteau de 12 mètres. Il monte de 3 mètres chaque jour mais redescend de 2 mètres chaque nuit. Combien de temps lui faudra-t-il pour atteindre le sommet du poteau ? (réponse écrite)

Les annotations de stress ont été réalisées sur chacune de ces séquences. Cependant, les participants peuvent en passer manuellement certaines. Pour avoir un temps fixe entre chaque séquence, on décide de regrouper les annotations en moyennant de la manière suivante :

— Les annotations D12 à D16 sont regroupées (moyenne)

— Les annotations D18 à D20 sont regroupées (moyenne)

On obtient un total de 8 séquences annotées.

6.4 Annotations

Les entretiens ont été découpées en 8 séquences temporelles et pour chacune d'elles, le stress été évalué selon quatre niveaux allant de 0 à 3, 0 désignant une personne non stressée et 3 une personne très stressée.

En outre, les informations relatives au genre du participant (homme ou femme) et du type de participant (stagiaire ou formateur) ont été ajoutées.

Séquence	Durée moyenne de chaque séquence (sec)	Écart-type de chaque séquence (sec)
D1-7	168	72
D8	50	0
D9	99	0
D10	67	0
D11	96	0
D12-16	162	26
D17	29	0
D18-23	184	59

Table 2 – Longueur des séquences annotées

6.5 Evaluation des annotations

- Chaque entretien est annotée par 2 personnes de manière indépendante
- Les écarts d'annotations des niveaux de stress sont analysés selon deux critères :
  - Distance L1 (somme des valeurs absolues des écarts pour toutes les annotations d'un entretien). Les 9 entretiens présentant une distance L1 supérieure ou égale à 7 ont été ré-annotées par une troisième personne. Ensuite, les deux annotations globales les plus proches (distance L1 la plus petite) ont été conservées. Enfin, si les distances L1 étaient toujours importantes (supérieure à 7), l'entretien faisait l'objet d'une ré-annotation par une quatrième personne, ce qui a été le cas pour un entretien
  - Différence de niveau de stress pour une séquence supérieur ou égal à 2. Quatre annotations spécifiques ont été revues par une troisième personne pour définir l'annotation finale. Suite à cette revue, la moyenne a été retenue dans 3 cas et le minimum dans 1 cas
- Une fois les annotations ajustées, une annotation finale est retenue comme le maximum des niveaux de stress annotés.

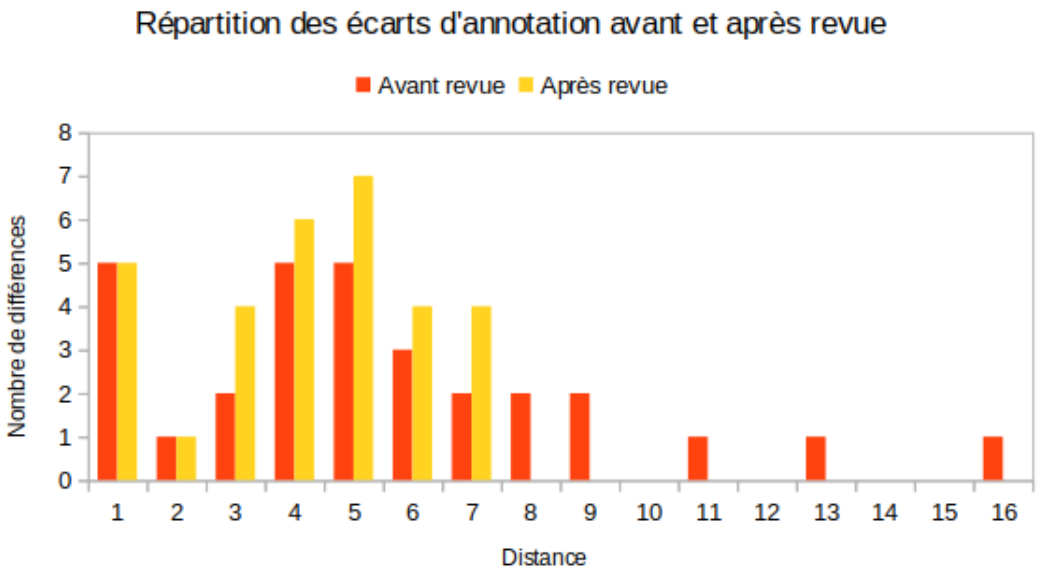


Figure 2 – Différences d'annotation (distance L1)

Les écarts inter-annotateurs sont étudiés à l'aide du cohen kappa définit comme suit :

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

où $\Pr(a)$ est l'accord relatif entre codeurs et $\Pr(e)$ la probabilité d'un accord aléatoire. Si les codeurs sont totalement en accord, $\kappa = 1$. S'ils sont totalement en désaccord (ou en accord dû uniquement au hasard), $\kappa \leq 0$.

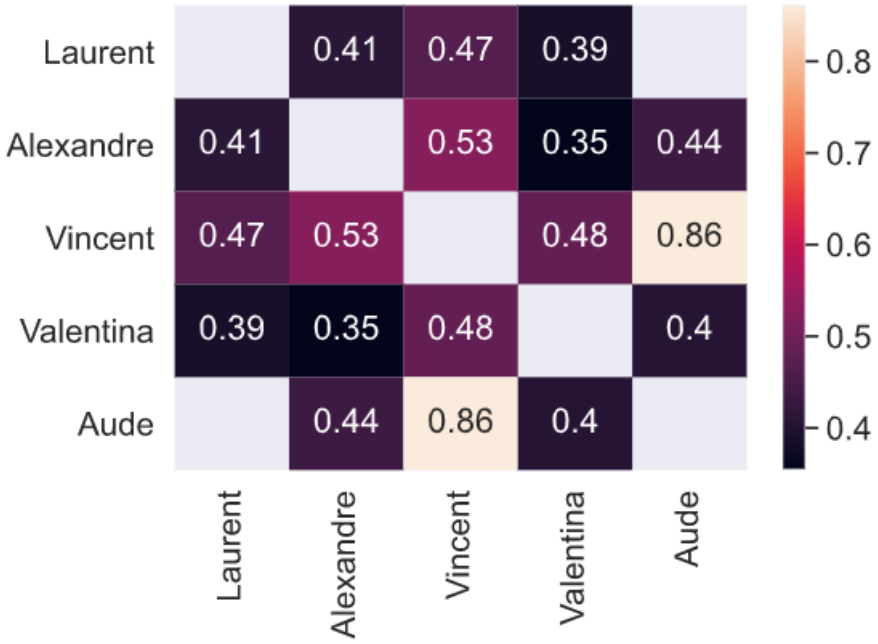


Figure 3 – Différences inter-annotateurs (Cohen Kappa)

Le coefficient de kappa peut atteindre un maximum de 1. [Landis and Koch, 1977] proposent la table suivante pour interpréter les résultats.

κ	**Interprétation**
< 0	Désaccord
0.0-0.2	Accord très faible
0.2-0.4	Accord faible
0.4-0.6	Accord modéré
0.6-0.8	Accord fort
0.8-1.0	Accord presque parfait

Table 3 – Interprétation du coefficient Kappa

Cependant il est à noter que ces ordres de grandeurs ne font pas consensus dans la communauté scientifique. Notamment parce que le nombre de catégories influe sur l'estimation obtenue – moins il y a de catégories, plus le κ est élevé.

L'évaluation de nos annotations par le coefficient Kappa est plus optimiste que les résultats obtenus à l'aide du Cronbach Alpha.

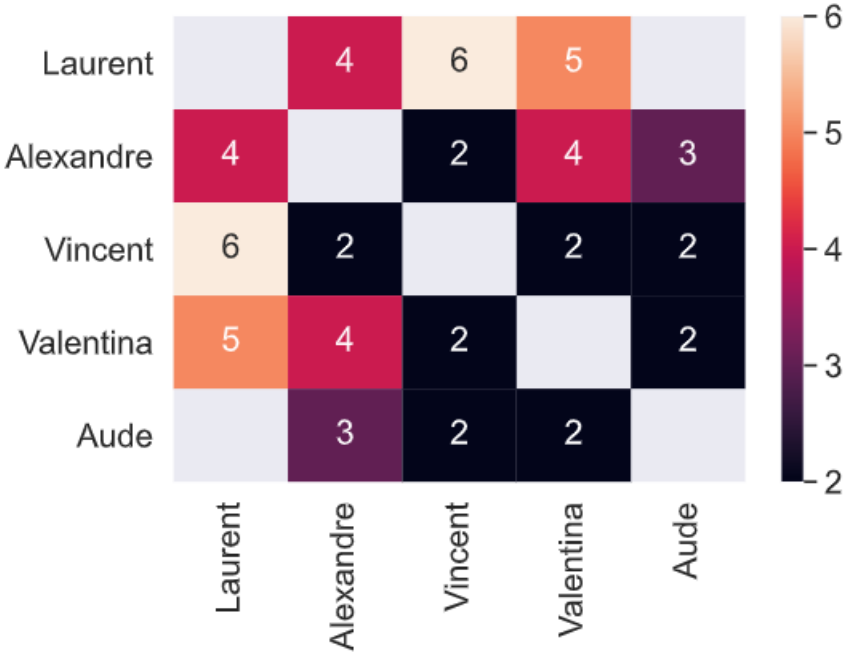


Figure 4 – Nombre de entretiens annotées en commun

Nous avons tenté de répartir équitablement les annotations. Chaque annotateur a annoté en moyenne 12 entretiens (en excluant l'annotateur "Aude"). La répartition des couples d'annotateurs, quant à elle, peut cependant être améliorée.

L'évaluation du stress est une tâche relativement difficile. Comment différencier l'amusement de la nervosité ? Le candidat parait-il stressé ou pressé d'en finir ? Les manifestations du stress sont variées parmi les individus. Nous aurions besoin de plus d'annotateurs pour affiner les scores et obtenir un meilleur consensus.

6.6 Annotations finales

Les 30 entretiens retenues permettent d'avoir 30 niveaux de stress globaux et 240 niveaux de stress au niveau de chaque séquence. Ces niveaux de stress sont présentés dans les graphiques ci-dessous.

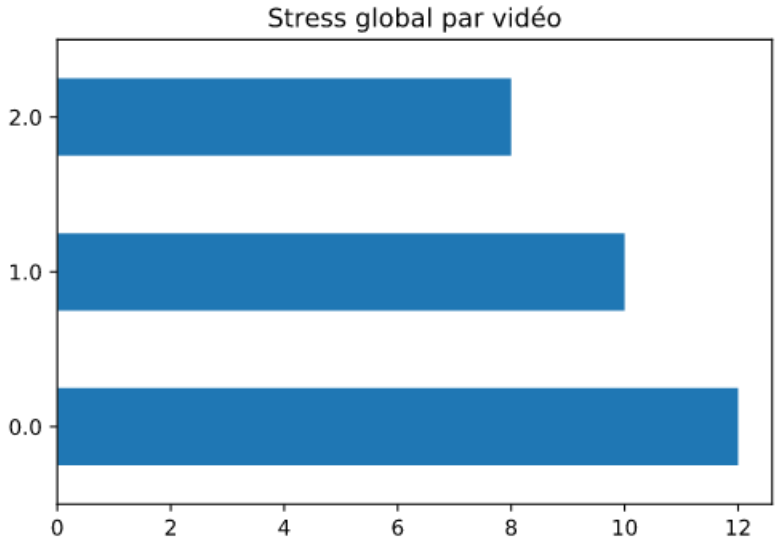


Figure 5 – Répartition du niveau de stress global

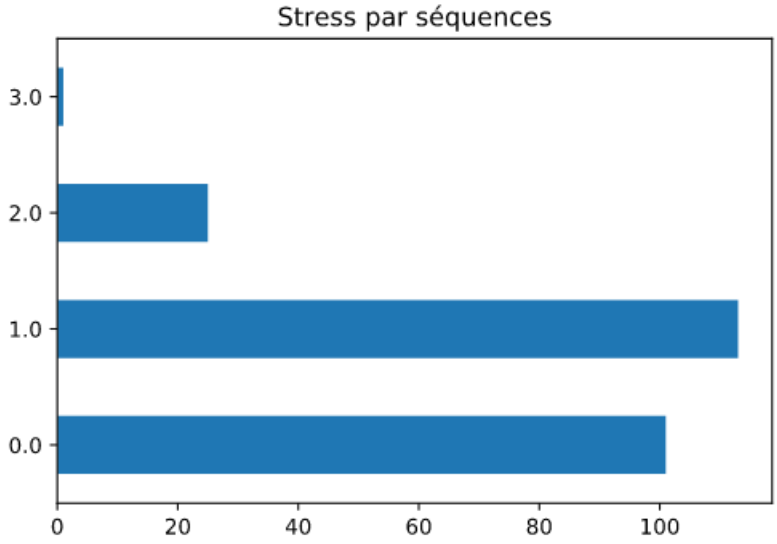


Figure 6 – Répartition du niveau de stress par séquence

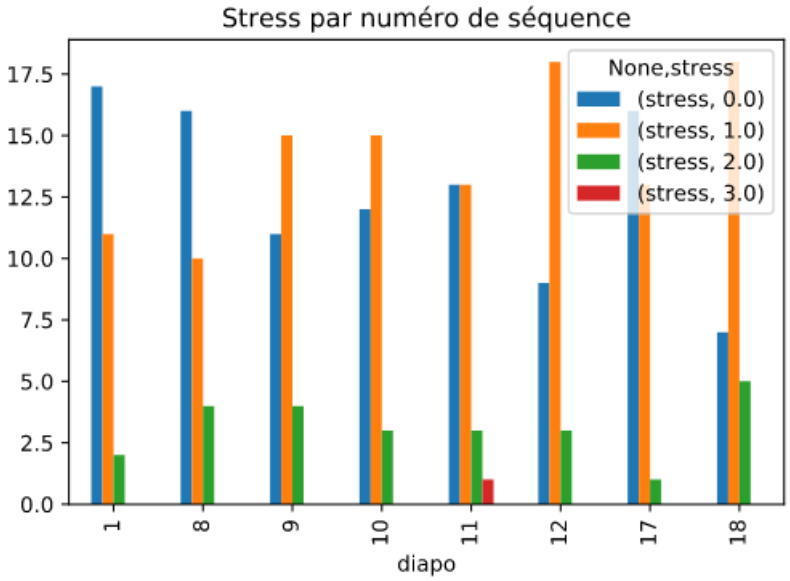


Figure 7 – Répartition du niveau de stress par numéro de séquence

On remarque une très faible quantité d'annotation de niveau 3 de stress. De plus, le niveau de stress annoté augmente au fur et à mesure de la entretien, les dernières séquences semblant générer un stress perçu plus important.

7 Extraction des caractéristiques

7.1 Caractéristiques vidéos

Pour extraire les features vidéos nous avons utilisé le toolkit OpenFace [Baltrusaitis et al., 2018]. Il s'agit d'un outil composé de plusieurs modèles pré-entraînés, qui extrait plusieurs features sous forme d'un csv. Dans notre cas OpenFace crée un fichier csv pour chaque entretien. Chaque ligne du fichier correspond à un timestamp de 0.03/0.06 secondes.

Des 31 entretiens fournies par l'AFPA initialement, OpenFace détecte un visage dans plus de 92% de timestamp pour 30 vidéos. La vidéo (WIN_20210402_14 _27_50_Pro) a été retiré pour manque de détection du visage (seulement 47% de la vidéo) dû à un mauvais éclairage.

En localisant les landmarks du visage et des yeux, OpenFace extrait des features spatiales, mais détecte aussi les points d'intérêt du système FACS à l'aide de SVMs pré-entrainés (cf. section 4.2.1). En particulier il détecte les unités d'action (AU, *i.e.* action

units), qui décrivent les contractions ou décontractions des muscles du visage.

AU	Full name	Illustration
AU1	Inner brow raiser	
AU2	Outer brow raiser	
AU4	Brow lowerer	
AU5	Upper lid raiser	
AU6	Cheek raiser	
AU7	Lid tightener	
AU9	Nose wrinkler	
AU10	Upper lip raiser	
AU12	Lip corner puller	
AU14	Dimpler	
AU15	Lip corner depressor	
AU17	Chin raiser	
AU20	Lip stretched	
AU23	Lip tightener	
AU25	Lips part	
AU26	Jaw drop	
AU28	Lip suck	
AU45	Blink	

Figure 8 – AUs détectés pas OpenFace

OpenFace permet de récupérer un total de 714 features. De ces features on retiendra la position du vecteur du regard pour chaque œil (3×2 features), l'angle de rotation du regard (2 features), la position et la rotation de la tête (6 features) la présence des AUs (18 features) et l'intensité des AUs (17 features, pour l'AU28 seulement la présence est détectée), pour un total de 49 features. La présence des AUs a comme valeurs 0 ou 1, et l'intesité est un flottant entre 0 et 5. Il faut remarquer que les données sur les AUs sont contradictoires : comme l'intensité et la présence ont été entraînées sur des datasets différents, il y a des cas où pour une certaine AU on à une valeur de 0 comme présence, mais une valeur différente de 0 pour l'intensité.

Open Face	Facial landmarks	Head position	Head rotation	Eyes landmarks	Gaze direction	Action Units intensity	Action Units presence	Features retained
Number of features	340	3	3	280	8	17	18	**49**

Une vidéo extraite avec OpenFace = un fichier de taille n ~ 24000 / p = 714 (une ligne toutes les 30 ms)

Figure 9 – Récapitulatif features vidéos

7.2 Caractéristiques audios

Dans un premier temps, les pistes audios sont extraites au format wav. Dans un deuxième temps, les fichiers obtenus sont utilisés avec le logiciel open-source openSmile [Florian Eyben, 2010] pour en extraire des caractéristiques audios. Les caractéristiques sont extraites toutes les 10ms sur des fenêtres temporelles de 40ms. Deux ensembles de caractéristiques audios ont été expérimentés : celles avec les configurations emobase (INTERSPEECH Paralinguistic Challenge feature set) [Eyben et al., 2009] et eGeMaps (extended Geneva Minimalistic Acoustic ParameterSet for Voice Research and Affective Computing) [Eyben et al., 2015] ainsi qu'une association des caractéristiques emobase et eGeMAPS.

7.2.1 emobase

Emobase contient les 26 descripteurs de bas niveau (low-level descriptors (LLD)) suivants ainsi que leurs dérivées :

— Prosodie : fréquence fondamentale F0 (fréquence de l'harmonique de premier rang d'un son complexe), probabilité de voix, pitch et *loudness* (intensité normalisée élevée à la puissance 0.3) et intensité

— MFCC : les coefficients cepstraux 1 à 12 des *Mel-Frequencies* (*Mel-Frequency cepstral coefficients*)

— LSF : les coefficients *line spectral pair frequencies* 1 à 8

— *Zero-Crossing Rate (ZCR)* : taux de changement de signe du signal

7.2.2 eGeMAPS

eGeMAPS contient les 23 descripteurs de bas niveau suivants :

— Caractéristiques relatives à la fréquence : *pitch*, *jitter* (déviation de la fréquence fondamentale F_0), fréquences formants 1, 2 and 3, bandes passantes des formants 1, 2 and 3

— Caractéristiques relatives à l'amplitude : *shimmer* (différence des amplitudes des pics de périodes F_0 consécutives), *loudness* (estimation de l'intensité du signal perçu), *Harmonics-to-Noise Ratio (HNR)* (relation de l'énergie dans les composantes harmoniques à l'énergie des composantes de bruit)

— Caractéristiques spectrales : ratio alpha (ratio des énergies cumulées entre $50 - 1000Hz$ et $1 - 5kHz$), index Hammarberg (ratio des pics d'énergie les plus forts entre $0 - 2kHz$ et $2 - 5kHz$), *spectral slope* $0 - 500Hz$ et $500 - 1500Hz$ (coefficients des régressions linéaires du spectre de la puissance logarithmique des deux bandes

de fréquences), énergies relatives formant 1, 2 and 3, différences harmoniques $H1-H2$ et $H1-H3$ (ratio des énergies des harmoniques considérées)

— MFCC : les coefficients cepstraux 1 à 4 des *Mel-Frequencies* (*Mel-Frequency cepstral coefficients*)

— Flux spectral : différence de spectres entre 2 frames consécutives

7.2.3 Mix eGeMAPS et emobase

Les caractéristiques eGeMAPS et emobase peuvent paraître complémentaires. D'une part, emobase présente des caractéristiques spectrales plus détaillées, par exemple avec plus de coefficients cepstraux de MFCC ainsi que des coefficients spectraux (LSF). D'autre part, eGeMAPS est plus complet sur les descripteurs relatifs à la voix, notamment avec les *jitter*, *shimmer*, *formants* et *Harmonics-to-Noise Ratio*. Nous allons tenter de combiner les caractéristiques de emobase et eGeMAPS pour avoir des caractéristiques audios plus complètes. Nous avons retenus :

— eGeMAPS
  — Caractéristiques relatives à la fréquence : *pitch*, *jitter* (déviation de la fréquence fondamentale F_0), fréquences formant 1, 2 and 3, bandes passantes du formant 1, 2 and 3 [7]
  — Caractéristiques relatives à l'amplitude : *shimmer* (différence des amplitudes des pics de périodes F_0 consécutives), *loudness* (estimation de l'intensité du signal perçu), *Harmonics-to-Noise Ratio (HNR)* (relation de l'énergie dans les composantes harmoniques à l'énergie des composantes de bruit)
  — Caractéristiques spectrales : ratio alpha (ratio des énergies cumulées entre $50-1000Hz$ et $1-5kHz$), index Hammarberg (ratio des pics d'énergie les plus forts entre $0-2kHz$ et $2-5kHz$), *spectral slope* $0-500Hz$ et $500-1500Hz$ (coefficients des régressions linéaires du spectre de la puissance logarithmique des deux bandes de fréquences), énergies relatives formant 1, 2 and 3, différences harmoniques $H1-H2$ et $H1-H3$ (ratio des énergies des harmoniques considérées)
  — Flux spectral : différence de spectres entre 2 frames consécutives
— emobase
  — Prosodie : fréquence fondamentale F0, probabilité de voix, pitch et *loudness* (intensité normalisée élevée à la puissance 0.3) et intensité
  — MFCC : les coefficients cepstraux 1 à 12 des *Mel-Frequencies* (*Mel-Frequency cepstral coefficients*)
  — LSF : les coefficients *line spectral pair frequencies* 1 à 8
  — *Zero-Crossing Rate (ZCR)* : taux de changement de signe du signal

7. Un formant (acoustique) d'un son de parole représente l'un des maxima d'énergie du spectre sonore de ce son de parole. Il dépend du type son de parole, en particulier de la manière avec laquelle le phonème a été prononcé et du contexte phonétique

7.3 Caractéristiques textuelles

Les caractéristiques textuelles correspondent aux transcriptions (sous forme de fichier txt) depuis la bande-son extraite pour les caractéristiques audios au format wav à la section précédente. Cette transcription est obtenue de manière automatique grâce à un ASR (Automatic Speech Recognition).

Sur l'ensemble des séquences d'entretien seulement 5 (séquences 8-11 et 17) contiennent des réponses orales. Sur ces 5 séquences, seulement 4 (séquences 8-11) sont des réponses à des questions ouvertes et présentent donc un intérêt sémantique : elles correspondent aux questions de présentation générale, de l'expérience et de la motivation du candidat ainsi que de sa connaissance du métier visé. Nous avons donc restreint l'extraction des *features* à ces 4 séquences.

D'un point de vue pratique, il a fallu annoter sur chaque entretien les temps de début de parole afin de localiser les parties de la bande son à extraire (la version gratuite de l'ASR de Google ne permettant pas d'extraire des séquences trop longues) et de découper la bande son en 4 parties.

N.B :
- les questions étant à durée déterminée, il n'a pas été nécessaire d'annoter les temps de fin de parole, l'ASR ne retranscrivant rien si la personne ne parle pas durant le temps imparti ;
- la qualité très hétérogène de la transcription automatique a nécessité une revue manuelle partielle.

7.3.1 Pré-traitement des données

Une première étape préliminaire à l'extraction des données consiste à retirer les mots vides et à opérer une tokenisation puis une lemmatisation du texte afin de ne traiter qu'une seule fois les mots qui ont une racine grammaticale commune (par exemple les conjugaisons d'un même verbe). La tokenisation ainsi que la lemmetisation ont été effectuées à l'aide de la librairie Spacy en Python.

Dans les 4 représentations suivantes, les séquences d'une interview (4 dans notre cas, soit un total de 124 pour 31 entretiens) sont considérées comme des observations (auxquelles on peut associer leur label grâce aux annotations pas séquence).

En revanche, les caractéristiques de ces séquences peuvent êtres extraites de manière différente.

7.3.2 Représentation TF-IDF

Dans la représentation TF-IDF ("Term Frequency-Inverse Document Frequency"), les *features* correspondent au vocabulaire utilisé par le stagiaire pour répondre à la question.

Pour chaque mot de vocabulaire, la fréquence d'apparition du mot dans la séquence est pondérée par l'inverse de sa fréquence d'apparition dans les autres séquences afin d'accorder plus de poids aux mots moins fréquents (sinon les mots vides caractériseraient toutes les séquences).

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$tfidf(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

Le choix du vocabulaire a une grande incidence sur cette représentation et de manière classique nous avons choisi l'ensemble des mots utilisés dans l'ensemble des séquences afin de diminuer la dimension de la matrice. Dans notre cas, le vocabulaire était constitué de 1460 mots.

Le désavantage de cette représentation est l'aspect creux de la matrice (peu d'observations dans un espace à grande dimension), l'utilisation d'une grande partie du vocabulaire étant peu réalisable lors de réponses de quelques minutes.

7.3.3 Représentation LIWC

Afin de lutter contre la grande dimension du vocabulaire, nous pouvons aussi regrouper les mots prononcés par catégories selon un dictionnaire spécifique. Conformément à la littérature, nous avons utilisé le dictionnaire LIWC (*Linguistic Inquiry Word Count*) afin d'associer chaque mot à une ou plusieurs émotions (par exemple le mot "stress" est associé aux catégories *affect*, *émotion négative* et *anxiété*). Cette représentation permet de passer d'un espace à 1460 dimensions à un espace à 64 dimensions.

7.3.4 Plongement lexical - fastText

Une manière de remédier à la sparsité de la matrice représentant nos caractéristiques textuelles est de représenter l'ensemble des mots de notre vocabulaire par une représentation vectorielle dans un espace (généralement à 300 dimensions) où des mots avec un sens similaire seraient proches.

Ces représentations sont généralement apprises par des réseaux de neurones grâce à deux modèles : le modèle *CBOW (Common Bag of Words)* et le modèle *Skip-Gram*. L'objectif du premier est de prédire un mot-cible en fonction des mots qui l'entourent alors que le second suit l'objectif inverse.

Étant donné le faible volume de données textuelles (124 réponses d'une durée maximale d'une minute trente secondes), il n'a évidemment pas été question d'entraîner notre propre modèle mais d'utiliser un modèle existant pré-entraîné.

Notre choix s'est porté sur un modèle (fastText) pré-entraîné et mis à disposition par le laboratoire de recherche en IA de Facebook. (`https://fasttext.cc/docs/en/crawl-vectors.html`

Comme indiqué sur leur site, cette représentation fasText a été pré-entraînée sur *Common Crawl* et Wikipedia en utilisant le modèle *CBOW* en dimension 300 avec des caractères n-grammes de taille 5. Cette représentation est disponible dans 157 langues et notamment en français.

L'avantage de cette représentation est qu'elle permet aussi de réduire la dimension de notre représentation à 300 mais elle est complètement exogène à notre exercice dans la mesure où n'avons même pas effectué de *fine-tuning*.

D'un point de vue pratique, une des contraintes particulières de cette représentation est qu'un mot lui-même est représenté par un vecteur à 300 dimensions (*embedding*) ; une séquence étant représentée par plusieurs mots, il faut donc trouver une manière d'agréger ces représentations vectorielles en une seule pour éviter d'avoir à traiter une représentation en 3 dimensions (documents*mots*embedding). Nous avons donc fait le choix d'agréger les mots d'une séquence par leur moyenne.

7.3.5 Représentation TF-IDF avec plongement lexical

Enfin, la dernière représentation est une combinaison de TF-IDF et de plongement lexical : elle consiste à pondérer l'agrégation des vecteurs par le poids qui leur a été assigné dans la représentation TF-IDF. D'un point de vue algébrique, il suffit d'effectuer le produit matriciel de la représentation TF-IDF par la matrice d'*embedding* du vocabulaire (vocabulaire*embedding).

Ces quatre représentations seront donc ensuite combinées à un modèle afin de déterminer celle qui est la plus pertinente.

Cependant, il convient de souligner le désavantage commun à ces 3 méthodes : elles ne prennent pas en compte l'ordre dans lequel les mots sont prononcés.

8 Modèles mono-modaux

8.1 Aspects méthodologiques

Les annotations des entretiens ont consisté à évaluer les niveaux de stress par séquence ainsi que le niveau de stress global. L'objectif de la tâche de prédiction présente donc plusieurs niveaux de hiérarchie :

— prédire le niveau de stress d'un stagiaire sur une séquence donnée ;

— prédire le niveau de stress global d'un stagiaire sur la durée de l'entretien. Dans ce cas, plusieurs niveaux de fusion des séquences sont possibles : on peut simplement agréger les séquences afin de prédire le stress global ou bien on peut utiliser les modèles précédents pour prédire le stress par séquence afin de prédire le stress global (*modèles en cascade*).

8.1.1 Validation croisée

Afin d'éviter *l'overfitting*, compte-tenu du volume faible de données, nous avons eu recours à une validation croisée pour entraîner nos modèles. Cependant, les observations sont constituées de plusieurs séquences d'une même personne interviewée et nous risquons donc d'avoir de la corrélation non désirée entre les différentes séquences appartenant à un même entretien. Pour cette raison, nous effectuons une validation croisée excluant un entretien (composé de 8 séquences) à itération de la validation croisée (*Leave One Group Out*). Ainsi, l'ensemble des séquences d'une entretien permettront chacune à leur tour de tester le modèle entraîné sur l'ensemble des séquences de tous les autres entretiens.

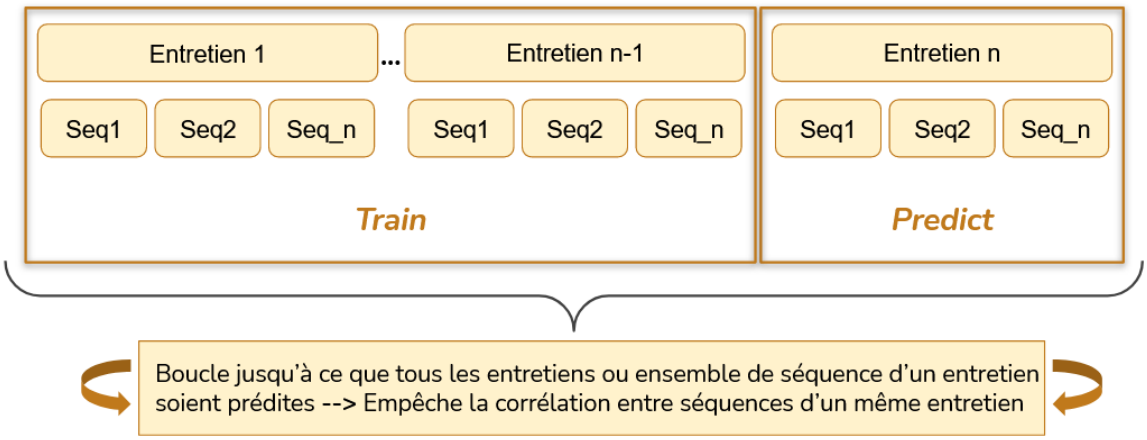


Figure 10 – Méthode de validation : Leave-One-Group-Out

8.1.2 Classification vs. Régression

L'utilisation de modèles de classification ou de régression a été une des premières problématiques que nous avons rencontrées. Les niveaux de stress sont des entiers allant de 0 à 3, pouvant être présentés comme des classes distinctes, donc une tâche de classification. Cependant, chaque niveau est bien ordonné logiquement. Les niveaux mesurent tous le stress. En recherchant dans la littérature existante, des solutions existes pour ces cas ambigus. La solution qui a particulièrement attirée notre attention est celle de [Frank and Hall, 2001]. "A simple approach to Ordinal Classification" propose de combiner plusieurs modèles de classification binaire pour retenir l'information de "l'ordre" des classes, comme le ferait une régression. Nous avons implémenté et testé cette méthode qui améliore en effet légèrement les résultats. Cependant, entraîner de multiples classifieurs est coûteux en temps de calcul.

Par ailleurs, les métriques de F1 score ou d'Accuracy (classification) sont des métriques plus intéressantes à présenter que la Mean Squared Error (regression) pour les besoins de l'AFPA. De plus, les matrices de confusion sont de bons outils d'analyse et de présentation des résultats à l'AFPA. C'est principalement cette dernière raison qui nous a poussé à choisir une modélisation en classification plutôt qu'en régression.

8.1.3 Ajout de caractéristiques

En plus des caractéristiques propres aux différentes modalités, nous avons rajouté des caractéristiques globales (notamment le **genre de la personne interviewée**) en espérant qu'elles puissent améliorer les performances de nos modèles.

Pour les modèles mono-modaux, le genre s'est avéré utile pour la tâche de prédiction seulement pour la modalité vidéo.

8.1.4 Critère de sélection

Dans la mesure où nous nous trouvons dans le cas de classes déséquilibrées (très peu de personnes présentant un stress de niveau 2 ou 3), nous avons préféré utiliser le ***F1-score*** pondéré afin de prendre en compte à la fois le *recall* et la *precision*. Le module *F1-score* de *sklearn* permet de calculer (grâce à l'option *weighted* de l'argument *average*) un F1-score pour chaque classe afin de pondérer le F1-score par le cardinal des observations de chaque classe.

8.1.5 Traitement des données

Les caractéristiques vidéos et audios, extraites avec les toolkits OpenFace et OpenSmile, permettent d'avoir des données indexées par *frames* (toutes les $30ms$ pour la vidéo et toutes les $10ms$ pour l'audio). Afin d'alimenter les modèles, elles ont fait l'objet de réductions temporelles :

— à l'échelle des séquences de la entretien, soit 8 séquences par interview. Les fonctions d'agrégations suivantes ont pu être utilisées : moyenne, écart-type, minimum, maximum, médiane, centiles 25% et 75%, kurtosis et asymétrie (skewness) ;

— dans des fenêtres temporelles de 5 secondes avec les mêmes fonctions d'agrégation.

En outre, certaines prédictions de stress ont pu être utilisées comme variables explicatives d'autres modèles (modèles hiérarchiques). Dans ces cas, ces prédictions ont pu faire l'objet de traitement :

— chaque prédiction de stress par séquence a été utilisée comme variable explicative, sans traitement ;

— à l'échelle d'une séquence ou de l'entretien, les stress prédits ont pu faire l'objet d'agrégations comme la moyenne, écart-type, min, max, médiane, centiles 25% et 75%, kurtosis et asymétrie (skewness)

— à l'échelle d'une séquence ou de l'entretien, les pourcentage de stress prédits ont été calculés (ex : 20% de stress prédit 0, 70% de stress prédit 1 et 10% de stress prédit 2)

8.2 Architecture des modèles non séquentiels mono-modaux audios et vidéos

Les features des modalités vidéo et audio extraites pour chaque entretien sont sous la forme d'un fichier ".csv" avec une ligne par frame, avec un timestamp. Nous allons

présenter ici les approches développées de manière parallèle pour les deux modalités.

Vu que le but est de prédire le niveau de stress de chaque personne, la question de comment utiliser les données temporelles , en particulier pour des séquences de durées variables. Une solution est de choisir de ne pas aplatir la temporalité, et d'utiliser un modèle séquentiel : cette approche sera développée en 8.6. Dans cette section on se concentre plutôt sur les différentes stratégies pour aplatir la temporalité et entraîner des modèles non séquentiels.

La deuxième question qui se pose est la manière d'utiliser les annotations du stress de chaque séquence pour prédire le stress d'une personne. On a mis en place différentes méthodes pour utiliser ces annotations, notamment des approches hiérarchiques.

8.2.1 Première architecture

La première architecture qu'on a utilisée est décrite dans la figure 11, où il y a deux pipelines qui sont représentées par deux flèches :

— La flèche en bas, qui va directement des features à la prédiction du stress global ;
— la flèche en haut qui va des features à la prédiction du stress global, en passant par la prédiction du stress par séquence.

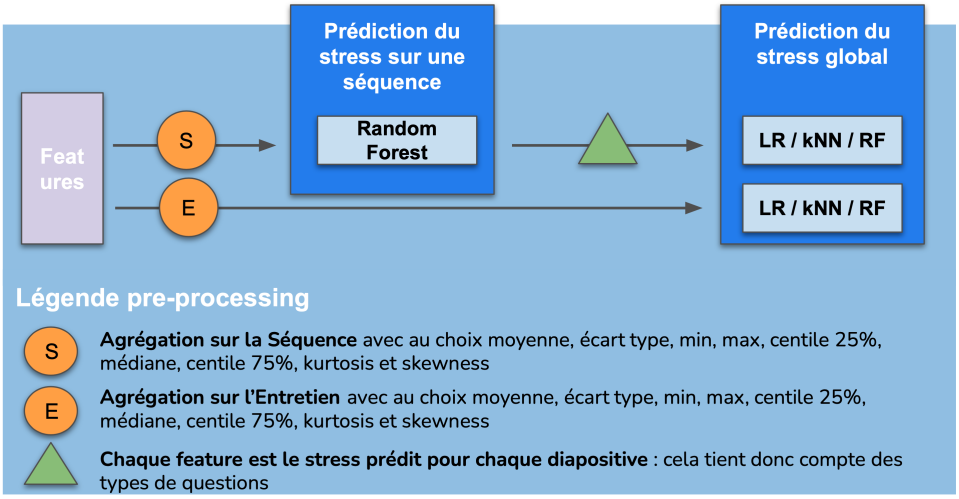


Figure 11 – Première architecture

L'architecture la plus simple pour prédire le stress global d'une personne est représentée par la flèche en bas de la figure 11. Dans cette pipeline, on a agrégé les features

de chaque entretien à travers plusieurs fonctions (moyenne, écart type, min, max, centile 25%, médiane, centile 75%, kurtosis et skewness), en obtenant un dataframe avec le même nombre de lignes que d'entretiens (*i.e.* le nombre des personnes). Nous avons alors prédit le stress de chaque personne, en utilisant les annotations au niveau de l'entretien. Nous avons testé plusieurs modèles pour cette prédiction (forêt aléatoire, régression logistique, k-NN après éventuellement avoir réduit le nombre des features avec une analyse en composantes principales (ACP)).

La flèche en haut de le figure 11 représente une pipeline plus complexe : elle est constituée de 2 modèles en cascade. Dans ce cas, on a agrégé les features de chaque séquence avec plusieurs fonctions d'agrégation, en obtenant un dataframe avec autant de lignes que le produit du nombre d'entretiens par celui des séquences. Nous avons pu alors utiliser les annotations au niveau de chaque séquence pour prédire le niveau de stress de la séquence avec une forêt aléatoire. Nous avons considéré alors le dataframe qui a pour lignes les entretiens et pour colonnes les différentes séquences, et où les valeurs sont les niveaux de stress prédits à l'étape précédente. Nous avons alors testé plusieurs modèles (forêt aléatoire, régression logistique, k-NN après éventuellement avoir réduit le nombre des features avec une ACP) pour cette prédiction.

8.2.2 Deuxième architecture

Nous avons ensuite décidé de construire une architecture plus fine, qui prend en compte le fait que le stress peut être détecté dans un intervalle temporel beaucoup plus court que la durée d'une séquence. Nous avons alors rajouté une étape à l'architecture précédente, en agrégeant d'abord les features par fenêtres temporelles de 5 secondes. En utilisant comme label le niveau de stress annoté sur la séquence dont la fenêtre fait partie, nous avons prédit le niveau de stress sur chaque fenêtre avec une forêt aléatoire. Ceci est représenté par la partie gauche de la figure 12.

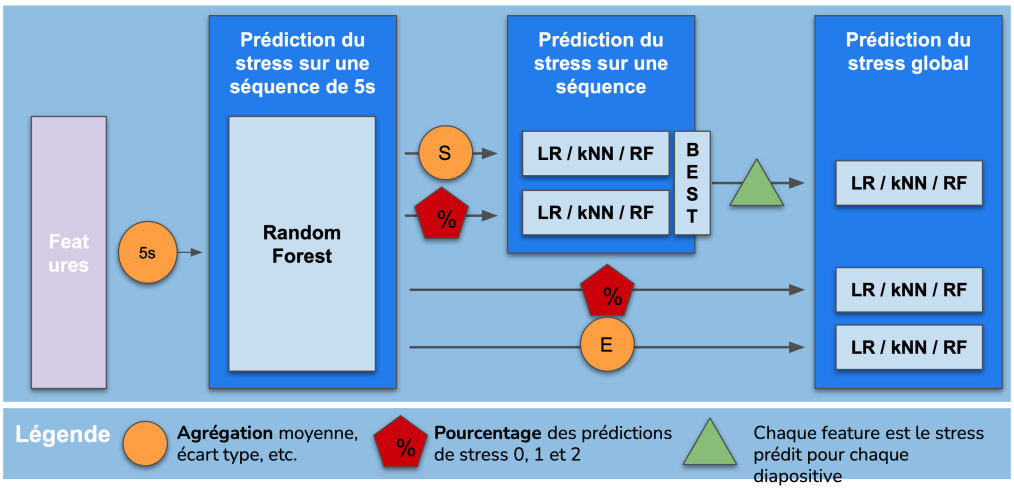


Figure 12 – Deuxième architecture

À partir de cette prédiction, on peut procéder comme dans la section 8.2.1 : il y a deux alternatives, décrites dans la suite.

La première alternative est représentée par les deux flèches en bas dans la partie droite de la figure 12, et consiste à utiliser les prédictions sur chaque fenêtre de 5 secondes pour prédire le stress sur chaque entretien. Nous avons exploité deux manières d'agréger les données :

— à travers plusieurs fonctions (moyenne, écart type, min, max, centile 25%, médiane, centile 75%, kutorsis et skewness)

— en considérant les pourcentages de stress prédits 0, 1, 2 et 3 pour chaque entretien.

Nous avons ensuite, comme avant, testé plusieurs modèles (forêt aléatoire, régression logistique, k-NN après éventuellement avoir réduit le nombre des features avec une ACP) pour cette prédiction.

La deuxième alternative consiste à agréger les prédictions sur les fenêtres de 5 secondes sur chaque séquence, faire une prédiction en analysant plusieurs modèles et ensuite utiliser les résultats obtenus pour prédire le stress sur chaque entretien. Cela est représenté dans la partie en haut de la figure 12.

8.3 Modèles mono-modaux vidéo non séquentiels

Les modèles testés sur les données vidéo ont été construits en suivant les architectures présentées dans la section 8.2. Nous présentons ici les meilleurs résultats obtenus en termes de **F1-score**. En outre, même si le but final est la prédiction du niveau du stress global de chaque personne, nous présentons aussi les résultats de la prédiction du

niveau de stress sur chaque séquence. Nous avons expérimenté les modèles avec deux combinaisons différentes des features :

— features liées aux AUs et features spatiales ;

— features liées aux AUs.

8.3.1 Features engineering

Dans le cadre des modèles non-séquentiels nous avons rajouté des features pour mesurer le changement au cours du temps. Si $Z = (z)_i$ est une colonne du dataframe, on a créé la dérivée première $Z' = (z')_i$ et la dérivée seconde $Z'' = (z'')_i$, définies comme suit

$$z'_i = |z_i - z_{i-1}|$$
$$z''_i = |z'_i - z'_{i-1}|$$

On a alors calculé les dérivées premières et secondes des features spatiales et des features qui mesurent l'intensité des AUs.

8.3.2 Résultats en considérant seulement les features connexes aux AUs

On présente ici le meilleur résultat obtenu en considérant seulement les features liées aux AUs.

Prédiction du stress par séquence Le modèle le plus performant présente un **F1** de 53.7% et une accuracy de 55.6%. Il est obtenu en

1. agrégeant les features sur chaque fenêtre de 5 secondes (la moyenne) ;
2. entraînant une forêt aléatoire pour prédire le stress sur chaque fenêtre ;
3. agrégeant les prédictions obtenues à niveau de chaque séquence (moyenne, écart-type, min, max, médiane, centiles 25% et 75%, kurtosis et asymétrie (skewness)) [8] ;
4. entraînant sur les prévisions précédemment agrégées un modèle constitué d'une analyse en composantes principales (APC avec $n = 3$) et un k-NN (avec $k = 3$)

Prédiction du stress global Le modèle le plus performant présente un **F1** de 56.8% et une accuracy de 62%. Il est obtenu en utilisant le modèle décrit dans le paragraphe précèdent. Au fait une fois obtenues les prédictions du niveau du stress par chaque séquence nous obtenons les scores en entraînent sur ces données un k-NN avec $k = 7$.

8. Des scores très similaires sont obtenus avec l'agrégation avec pourcentages de 0, de 1, de 2 et de 3 prédits pour chaque entretien, mais en ce cas le modèle n'arrive pas à prédire les 2 ou les 3.

8.3.3 Résultats en considérant les features connexes aux AUs et les features spatiales

Nous avons constaté que l'architecture du modèle décrite dans le paragraphe précédent se trouve être la plus performante même lorsqu'on utilise toutes les features. Nous décrivons ici les détails pour la commodité du lecteur.

Prédiction du stress par séquence Le modèle le plus performant donne un score de 53.7% en termes de **F1** et un score de 56.9% en termes d'accuracy. Il est obtenu en

1. agrégeant les features sur chaque fenêtre de 5 secondes (la moyenne) ;
2. entraînant une forêt aléatoire pour prédire le stress sur chaque fenêtre ;
3. agrégeant les prédictions obtenues à niveau de chaque séquence (moyenne, écart-type, min, max, médiane, centiles 25% et 75%, kurtosis et asymétrie (skewness)) ;
4. entraînant sur les prévisions précédemment agrégées un modèle constitué d'une analyse en composantes principales (APC avec $n = 3$) et un k-NN (avec $k = 4$)

Prédiction du stress global Le modèle le plus performant présente un **F1** de 67.3% et une accuracy de 68.9%. Il est obtenu en utilisant le modèle décrit dans le paragraphe précèdent. Au fait, une fois obtenues les prédictions du niveau du stress par chaque séquence nous obtenons les scores en entraînant sur ces données un k-NN avec $k = 10$, après une analyse en composantes principales (APC avec $n = 3$).

8.3.4 Feature importance

Il est intéressant d'observer les résultats d'un modèle très simple pour la prédiction du niveau du stress par séquence : il s'agit de l'agrégation des features par séquence et de la prédiction du niveau du stress par séquence faite avec une forêt aléatoire. Même si les performances ne sont pas exceptionnelles (**F1** de 43.9% et accuracy de 46.9%) on peut, grâce au fait que nous utilisons une forêt aléatoire, analyser l'importance de chaque feature. On remarque que les features les plus importantes pour la détection du stress sont liées aux muscles autour des yeux (AU07 : *Orbicularis oculi, pars palpebralis* et AU05 : *Orbicularis oculi, pars palpebralis*).

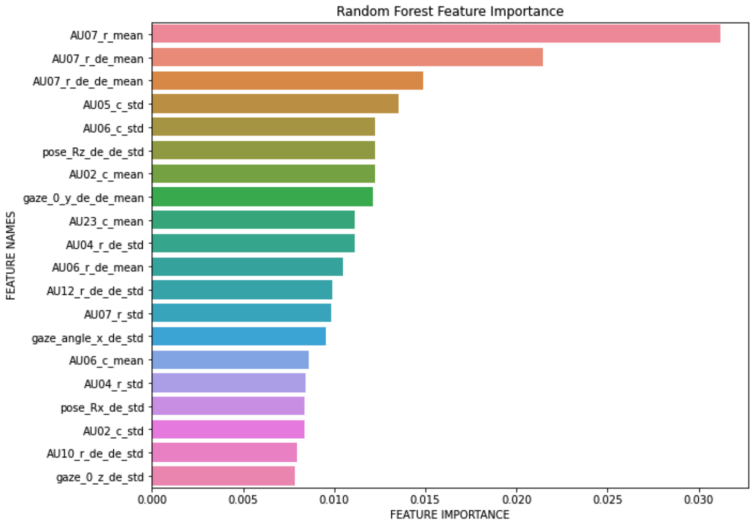


Figure 13 – Features importance

8.4 Modèles mono-modaux audios

Les modèles mono-modaux audios ont été testés selon les architectures présentées en 8.2. Ils ont été testés sur 3 ensembles de caractéristiques audios (emobase, eGeMAPS et mélange de emobase et eGeMAPS) d'une part et sur des durées de séquences d'entretien différentes d'autre part. En effet, dans ce dernier cas, nous avons voulu expérimenté des modèles s'appuyant sur (1) toutes les 8 séquences de l'entretien et (2) seulement les 5 séquences pour lesquelles une réponse orale était attendu, c'est-à-dire les séquences 8, 9, 10, 11 et 17.

8.4.1 Cas 1 : utilisation de l'ensemble des 8 séquences de l'entretien

Prédiction du stress par séquence Parmi les modèles testés, le score F1 le plus élévé est de 49.9% et il présente une accuracy de 51.3%. Il s'agit d'un modèle hiérachique.

1. Dans un premier temps, les caractéristiques audio, mélange de emobase et eGeMAPS, sont agrégées dans des fenêtres temporelles de 5 secondes (moyenne, écart-type, minimum, maximum, médiane, centiles 25% et 75%, kurtosis et asymétrie (skewness))
2. Les prédictions sont ensuite agrégrées au niveau de chacune des 8 séquences via leur pourcentage de niveau de stress prédit
3. Un modèle de forêts aléatoires est ensuite entraîné sur ces agrégats, résultant un score F1 de 41.7% et une accuracy de 44.7%

4. Les prédictions de ce modèle alimentent ensuite un nouveau modèle de forêts aléatoires et obtient les indicateurs précisés ci-dessus.

Prédiction du stress global Le modèle le plus performant a obtenu un score F1 et une accuracy respectivement de de 55.3% et 56.7%. Il s'agit également d'un modèle hiérarchique, mais celui-ci n'utilise par les fenêtres temporelles de 5 secondes.

1. Dans un premier temps, les caractéristiques audio eGeMAPS sont agrégées au niveau de chacune des 8 séquences (moyenne, écart-type, minimum, maximum, médiane, centiles 25% et 75%, kurtosis et asymétrie (skewness))
2. Un modèle de forêts aléatoires est ensuite entraîné sur ces agrégats, résultant un score F1 de 31.1% et une accuracy de 40.0%
3. Les prédictions de ce modèle alimentent ensuite un nouveau modèle de régression logistique et obtient les indicateurs précisés ci-dessus.

8.4.2 Cas 2 : utilisation des 5 séquences pour lesquelles une réponse orale est attendue

Prédiction du stress par séquence Le modèle le plus performant présente un F1 et une accuracy respectives de 51.5% et 54.0%. Il est obtenu en :

1. agrégeant les caractéristiques audio emobase au niveau de chacune des 8 séquences (moyenne, écart-type, minimum, maximum, médiane, centiles 25% et 75%, kurtosis et asymétrie (skewness))
2. entraînant un modèle de forêts aléatoires sur ces agrégats, résultant les scores précisés ci-dessus.

Prédiction du stress global Le modèle le plus performant présente un score F1 et une accuracy respectives de 60.5% et 60.0%. Il s'agit d'un modèle hiérarchique basé sur les caractéristiques emobase.

1. Dans un premier temps, les caractéristiques audio emobase sont agrégées dans des fenêtres temporelles de 5 secondes (moyenne, écart-type, minimum, maximum, médiane, centiles 25% et 75%, kurtosis et asymétrie (skewness))
2. Un modèle de forêts aléatoires est ensuite entraîné sur ces agrégats, résultant un score F1 de 44.6% et une accuracy de 46.3%
3. Les prédictions de ce modèle sont agrégées au niveau de chacune des 8 séquences (moyenne, écart-type, minimum, maximum, médiane, centiles 25% et 75%, kurtosis et asymétrie (skewness))
4. ELles alimentent ensuite un nouveau modèle constitué d'une analyse en composantes principales (ACP) et d'un k plus proches voisins, avec des F1 et accuracy respectifs de 38.7% et 47.4%.

5. Les prédictions de ce modèle alimentent ensuite un nouveau modèle constitué d'une analyse en composantes principales (ACP) et d'un k plus proches voisins, et donne les F1 et accuracy donnés plus-hauts, soit 60.% et 60.0% respectivement

8.5 Modèles mono-modaux textuels

8.5.1 Prédiction du stress par séquence

Pour la prédiction du stress par séquence pour la modalité textuelle, le modèle le plus performant en termes de F1-score est la combinaison de caractéristiques LIWC et d'une Forêt Aléatoire avec un score de 52.57% (pour une *accuracy* de 49.19%).

Le deuxième meilleure performance est atteinte avec une représentation *word embedding* et une forêt aléatoire avec un F1-score de 50.32%.

Il est intéressant de noter que la performance de la combinaison des représentations TF-IDF et *word embedding* est meilleure ques ces deux-ci seulement pour le modèle SVC ($41.39\% < 41.75\% < 46.77\%$). Avec une Forêt Aléatoire, cette combinaison donne des résultats (42.74%) meilleurs que la représentation TF-IDF (38.07%) mais moins bons que la représentation *word embedding* seule (50.32%).

Modèle (stress par séquence)	F1-score	Accuracy
TF-IDF + RF	38.07%	34.68%
TF-IDF + SVC	41.39%	37.10%
LIWC + RF	52.57%	49.19%
LIWC + SVC	44.81%	39.52%
fastText + RF	50.32%	45.97%
fastText + SVC	41.75%	36.29%
TF-IDF + fastText + RF	42.74%	44.84%
TF-IDF + fastText + SVC	46.77%	38.43%

8.5.2 Prédiction du stress global par entretien

En ce qui concerne la prédiction du stress global par entretien, le modèle le plus performant en F1-score est la combinaison de caractéristiques *word embedding* et d'une Forêt Aléatoire avec un score de 50.06% (pour une *accuracy* de 51.90%).

Globalement, les scores sont moins bons que pour le stress par local (par séquence). Ceci est sûrement dû au fait que la modalité textuelle ne concerne que 4 séquences alors que le stress global porte sur l'ensemble de l'entretien (8 séquences).

Modèle (stress par entretien)	F1-score	Accuracy
TF-IDF + RF	27.30%	34.76%
TF-IDF + SVC	24.24%	35.76%
LIWC + RF	35.41%	38.10%
LIWC + SVC	49.17%	50.95%
fastText + RF	40.22 %	44.76%
fastText + SVC	50.06 %	51.90%
TF-IDF + fastText + RF	37.57 %	42.38%
TF-IDF + fastText + SVC	45.16 %	36.19%

8.6 Modèle LSTM Vidéo

L'utilisation d'un réseau de neurones Long Short-Term Memory (LSTM) avait principalement pour objectif d'évaluer la pertinence d'un modèle utilisant la séquentialité des entretiens. La modalité vidéo étant la plus prometteuse, nous avons choisi d'entraîner notre LSTM seulement sur les caractéristiques extraites à partir d'OpenFace. Les features ainsi extraites correspondent à des niveaux d'intensité des Action Units (muscles faciaux) ainsi qu'aux coordonnées de rotation et position de la tête.

8.6.1 Architecture

On utilise des fenêtres temporelles d'1 seconde, avec un chevauchement de 0.5 seconde, pour effectuer une moyenne glissante sur l'ensemble des entretiens. On fixe la durée d'une séquence LSTM à une taille 15. Chaque séquence prédite a donc une durée de 8 secondes. On assigne le label de stress local (stress annoté sur les 8 séquences de l'entretien) à chaque séquence LSTM correspondante.

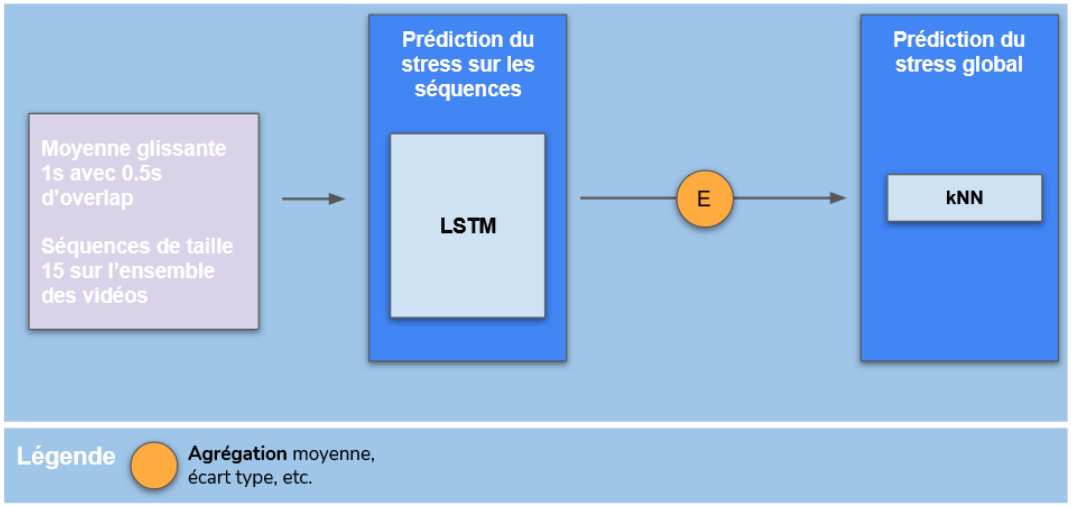


Figure 14 – Architecture du LSTM sur la modalité Vidéo

Pour valider le modèle, on prédit toutes les séquences de 15s de chaque entretien en utilisant le principe du Leave-One Group Out (8.1.1) . On agrège ensuite les prédictions (moyenne, min, max, écart-type) pour chaque entretien. Enfin, suivant la même méthode de validation, un k-NN est entrainé pour prédire le stress global.

8.6.2 Résultats

Modèle	Accuracy	F1-score
LSTM + k-NN	50.1%	51.8%

Table 4 – Prédiction du stress global

Un des problèmes de notre modèle est sa difficulté à prédire les classes minoritaires (niveau de stress 2 et 3). Malgré l'ajout de poids spécifiques à ces classes, les performances sont faibles. Par ailleurs, la méthode de validation choisie est trop coûteuse pour le réseau de neurones dont le temps d'apprentissage est élevé pour chaque cross validation. Il faudrait utiliser un jeu d'entrainement et de validation simple pour modifier l'architecture du réseau et les hyperparamètres associés. De plus, un second niveau de LSTM associé à la prédiction du stress global plutôt que d'utiliser un k-NN serait une piste d'amélioration potentielle.

Nous avons choisi de ne pas continuer à investiguer les réseaux de neurones puisque leur utilisation est très coûteuse en temps de calcul et les performances sur nos modèles non séquentiels étaient prometteuses.

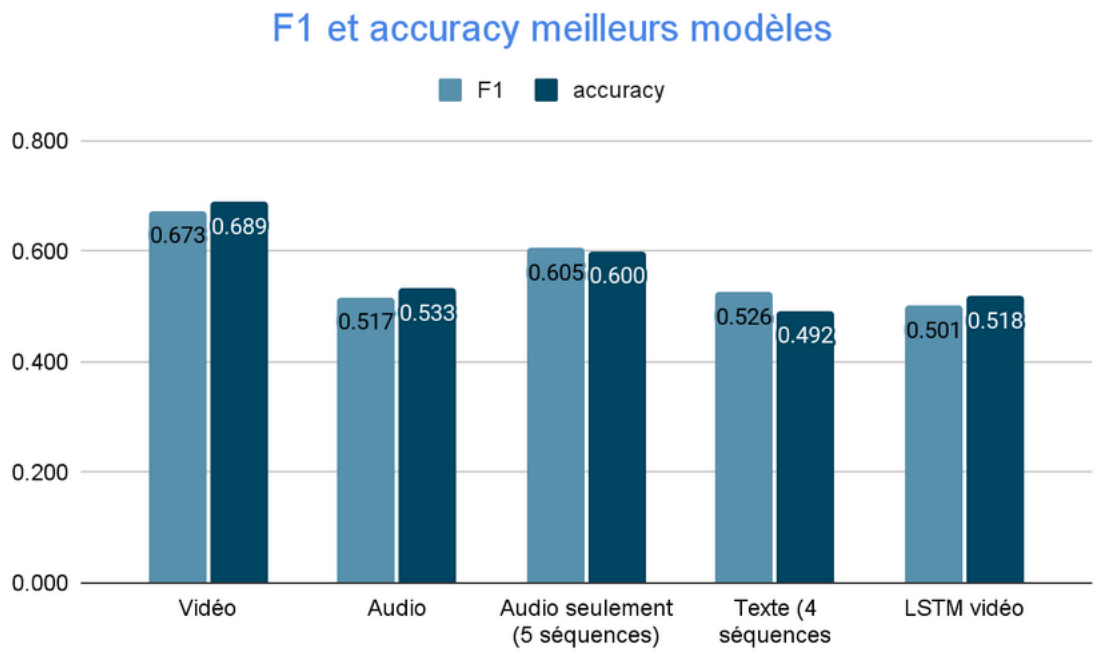


Figure 15 – Meilleurs résultats des modèles mono-modaux par modalité

9 Modèles multi-modaux

9.1 Architecture

De la même manière que dans la partie 8.1 pour les différentes séquences d'un entretien sur une modalité donnée, plusieurs hiérarchies d'agrégation des modalités sont possibles : pour prédire le stress de manière "multimodale", on peut simplement agréger les caractéristiques de chaque modalité en amont de l'entraînement d'un modèle commun aux trois modalités (figure 16) ou bien en aval d'une première étape de classification propre à chaque modalité (figure 17).

9.1.1 *Early fusion*

Comme précisé en introduction, l'architecture *Early fusion* correspond à la fusion en amont de l'ensemble des caractéristiques extraites sur toutes les modalités dans une grande *features matrix* qui sera alors entraînée par un classifieur unique (Régression Logistique, KNN ou Forêt Aléatoire). Contrairement à l'architecture suivante, elle n'utilise pas les prédictions du stress de chaque modalité.

Comme l'indique les différents blocs à droite de la figure ci-dessous, cette fusion est réalisée pour la tâche de prédiction du stress par séquence ainsi que celle du stress global.

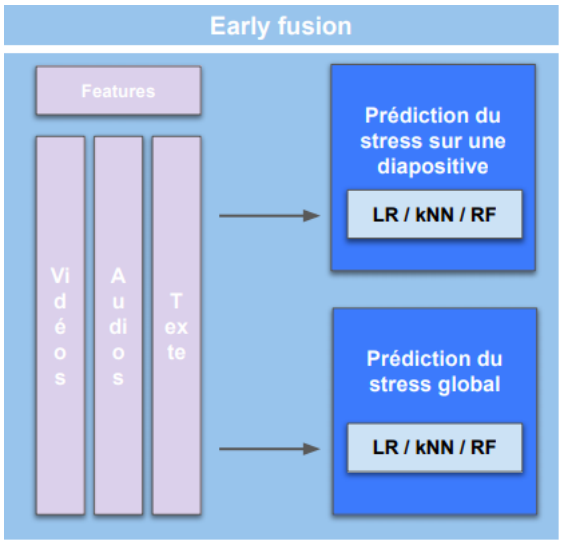


Figure 16 – Early Fusion

9.1.2 *Late fusion*

En ce qui concerne l'architecture *Late fusion*, elle correspond à une fusion des modalités en aval d'une première étape de prédiction monomodale. Elle se déroule en deux étapes :

1. entraînement de classifieurs monomodaux et prédiction du stress par modalité ;
2. agrégation et injection des prédictions monomodales dans un nouveau classifieur (Régression Logistique, KNN ou Forêt Aléatoire).

Il est important de noter que les classifieurs utilisés pour chaque modalité dans la première étape peuvent être différents.

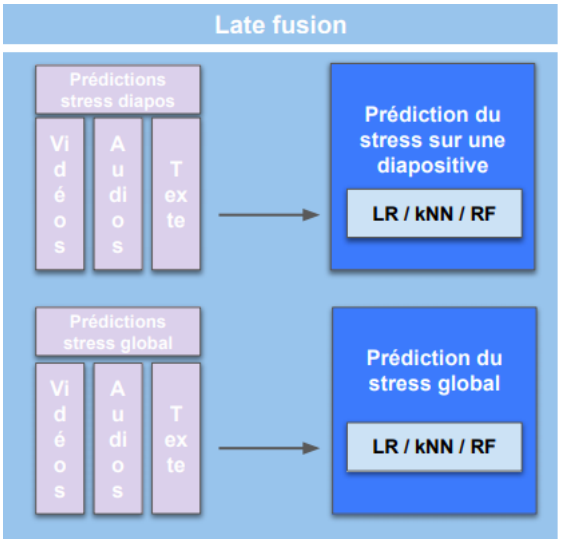


Figure 17 – Late Fusion

9.2 Résultats

9.2.1 Stress par séquences

Early fusion La meilleure performance pour la prédiction multimodale du stress par séquences avec *early fusion* est obtenue en combinant les trois modalités dans une Forêt Aléatoire (55.39%). Cette performance est meilleure que le maximum des modalités prises séparément (54.71% pour la video, 54.21% pour l'audio et 47.77% pour le texte).

Modèle (stress par séquence)	F1-score	Accuracy
Video + Audio + Texte	55.39%	56.04%
Video + Audio	53.64%	53.76%

Late fusion La meilleure performance pour la prédiction multimodale du stress par séquences avec *late fusion* est obtenue en combinant les trois modalités en effectuant au préalable un préprocessing (pourcentage de scores prédits vidéo, audio et texte) puis une régression logistique (54.8%).

Modèle (stress par séquence)	F1-score	Accuracy
Video + Audio + Texte	54.8%	56.0%
Video + Audio only	51.4%	54.7%
Video + Audio	53.6%	54.7%

9.2.2 Stress global

Early fusion La meilleure performance pour la prédiction multimodale du stress global avec *early fusion* est obtenue en combinant seulement les modalités vidéos en appliquant au préalable une ACP à trois composantes puis en effectuant une régression logistique.

Modèle (stress par entretien)	F1-score	Accuracy
Video + Audio + Texte	42.49%	44.82%
Video + Audio	49.72%	51.72%

Late fusion La meilleure performance pour la prédiction multimodale du stress global avec *late fusion* est obtenue en combinant seulement les modalités vidéo et audio et en effectuant au préalable une ACP à deux composantes puis en effectuant un KNN (78.9%).

Modèle (stress par entretien)	F1-score	Accuracy
Video + Audio + Texte	75.7%	75.9%
Video + Audio (5 séquences)	65.8%	65.5%
Video + Audio	78.9%	79.3%
Video	67.3%	68.9%

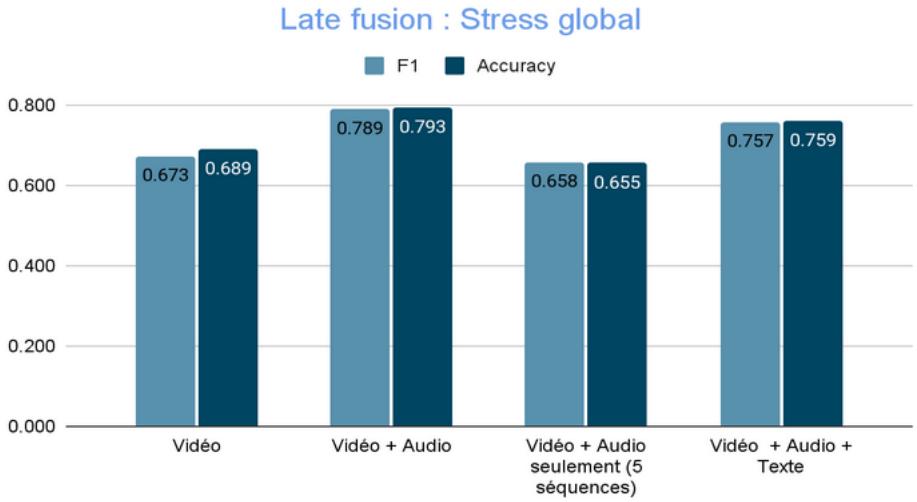


Figure 18 – Meilleurs résultats des modèles multi-modaux

10 Conclusion

10.1 Résultats

En conclusion, la modalité la plus importante pour la détermination du niveau de stress global s'avère être la vidéo (F1 de 67,3%), devant le texte (F1 de 52,6%) et l'audio (F1 de 51,7%). En effet, intuitivement, le stress peut être détecté plus facilement par des expressions du visage que par la voix ou le vocabulaire utilisé. Ces deux dernières modalités peuvent par ailleurs être modifiées plus facilement pour masquer le stress.

L'ajout des modalités audio et texte à la modalité vidéo s'avère bénéfique et permet d'améliorer le modèle mono-modal vidéo. Le meilleur modèle (F1 de 78.9%) *late fusion* est une combinaison des modalités vidéos et audios. La combinaison avec la modalité texte ne permet pas d'améliorer le modèle, probablement car seules 4 séquences sont analysées et que le texte contient moins d'information en qualité et en quantité.

En conclusion, les modèles non séquentiels produisent de bons résultats.

10.2 Problématiques rencontrées

La réalisation des entretiens à domicile été un gros challenge pour les stagiaires, en particulier en raison de la crise sanitaire. En outre, la contrainte de l'exercice en ligne, filmé, ainsi que les réactions différentes des individus à ce type d'exercice, n'a pas permis d'avoir des données équilibrées en termes de stress. En effet, le nombre d'annotations de stress élevé et moyen étaient faibles, d'où la difficulté de construire des modèles adéquats.

En outre, l'évaluation du stress est une tâche relativement difficile. Comment différencier l'amusement de la nervosité ? Le candidat paraît-il stressé ou pressé d'en finir ? Les manifestations du stress sont variées parmi les individus.

Enfin, la gestion de la temporalité et de la hiérarchisation du stress local / stress global a donné lieu à une multiplicité de combinaisons de traitement des données et des modèles.

10.3 Améliorations possibles

Le protocole d'annotation du stress peut être amélioré, notamment avec des directives précises des manifestations du stress. Cela requiert l'aide d'un spécialiste.

— Modalité vidéo : Améliorer le modèle LSTM à l'aide d'un second niveau qui utiliserait les prédictions de séquence du premier LSTM. Utiliser la mécanique d'attention

dans le réseau. Les modèles séquentiels peuvent être plus explorés ;

— Modalité audio : Utiliser des CNN sur les spectrogrammes audio pour extraire des caractéristiques locales, et ensuite utiliser des LSTM qui prennent en compte la séquentialité ;

— Modalité texte : Utiliser un word embedding prenant en compte le contexte (ex : BERT).

Références

[Acharyya et al., 2020] Acharyya, R., Das, S., Chattoraj, A., and Tanveer, M. I. (2020). Fairyted : A fair rating predictor for TED talk data. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 338–345. AAAI Press.

[Baltrusaitis et al., 2018] Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0 : Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66.

[Escalante et al., 2020] Escalante, H. J., Kaya, H., Salah, A. A., Escalera, S., Güçlütürk, Y., Güçlü, U., Baró, X., Guyon, I., Jacques, J. C. S., Madadi, M., Ayache, S., Viegas, E., Gurpinar, F., Wicaksana, A. S., Liem, C., Van Gerven, M. A. J., and Van Lier, R. (2020). Modeling, recognizing, and explaining apparent personality from videos. *IEEE Transactions on Affective Computing*, pages 1–18.

[Eyben et al., 2015] Eyben, F., Scherer, K., Schuller, B., Sundberg, J., Andre, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., and Truong, K. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7 :1–1.

[Eyben et al., 2009] Eyben, F., Wöllmer, M., and Schuller, B. (2009). Openear — introducing the munich open-source emotion and affect recognition toolkit. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6.

[Finnerty et al., 2016] Finnerty, A. N., Muralidhar, S., Nguyen, L. S., Pianesi, F., and Gatica-Perez, D. (2016). Stressful first impressions in job interviews. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 325–332, New York, NY, USA. Association for Computing Machinery.

[Florian Eyben, 2010] Florian Eyben, Martin Wöllmer, B. S. (2010). opensmile - the munich versatile and fast open-source audio feature extractor. *ACM Multimedia (MM)*, pages 1459–1462.

[Frank and Hall, 2001] Frank, E. and Hall, M. (2001). A simple approach to ordinal classification. In De Raedt, L. and Flach, P., editors, *Machine Learning : ECML 2001*, pages 145–156, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Hasan et al., 2019] Hasan, M., Ullah, S., Khan, M. J., and Khurshid, K. (2019). Comparative Analysis of SVM, ANN and CNN for Classifying Vegetation Species Using Hyperspectral Thermal Infrared Data. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4213 :1861–1868.

[Hemamou et al., 2019] Hemamou, L., Felhi, G., Vandenbussche, V., Martin, J.-C., and Clavel, C. (2019). Hirenet : A hierarchical attention model for the automatic analysis of asynchronous video job interviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01) :573–581.

[Kar and Corcoran, 2017] Kar, A. and Corcoran, P. (2017). A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access*, 5 :16495–16519.

[Kaya et al., 2017] Kaya, H., Gürpinar, F., and Salah, A. A. (2017). Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1651–1659.

[Landis and Koch, 1977] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1) :159–174.

[Mora and Odobez, 2016] Mora, K. A. F. and Odobez, J.-M. (2016). Gaze estimation in the 3d space using rgb-d sensors - towards head-pose and user invariance. *International Journal of Computer Vision*, 118(2) :194–216.

[Nguyen, 2015] Nguyen, L. (2015). *Computational Analysis Of Behavior In Employment Interviews And Video Resumes*. PhD thesis, EPFL.

[Viola and Jones, 2001] Viola, P. A. and Jones, M. J. (2001). Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518. IEEE Computer Society.

[Wu and Qu, 2020] Wu, A. and Qu, H. (2020). Multimodal analysis of video collections : Visual exploration of presentation techniques in ted talks. *IEEE transactions on visualization and computer graphics*, 26(7) :2429—2442.