Diagnostic automatique de l'état dépressif

S. Cholet H. Paugam-Moisy
Laboratoire de Mathématiques Informatique et Applications (LAMIA - EA 4540)
Université des Antilles, Campus de Fouillole - Guadeloupe
Stephane.Cholet@univ-antilles.fr

Résumé

Les troubles psychosociaux sont un problème de santé publique majeur, pouvant avoir des conséquences graves sur le court ou le long terme, tant sur le plan professionnel que personnel ou familial. Le diagnostic de ces troubles doit être établi par un professionnel. Toutefois, l'IA (l'Intelligence Artificielle) peut apporter une contribution en fournissant au praticien une aide au diagnostic, et au patient un suivi permanent rapide et peu coûteux. Nous proposons une approche vers une méthode de diagnostic automatique de l'état dépressif à partir d'observations du visage en temps réel, au moyen d'une simple webcam. A partir de vidéos du challenge AVEC'2014, nous avons entraîné un classifieur neuronal à extraire des prototypes de visages selon différentes valeurs du score de dépression de Beck (BDI-II).

Abstract

Psychosocial disorders are a major public health problem that can have serious consequences in the short or long term, on a professional, personal or family level. The diagnosis of these disorders must be made by a professional. However, AI (Artificial Intelligence) can make a contribution by providing the practitioner with diagnostic assistance, and the patient with rapid and inexpensive ongoing follow-up. We propose an approach towards an automatic diagnosis of the depressive state based on real-time facial observations, using a simple webcam. From videos of the AVEC'2014 challenge, we trained a neural classifier to extract prototypes of faces according to different values of Beck's depression score (BDI-II).

Mots Clefs

Informatique affective ; visages ; classifieur incrémental ; réseaux de neurones ; apprentissage de prototypes.

1 Introduction

1.1 Contexte

Les troubles psychosociaux, et singulièrement les troubles dépressifs, sont une maladie touchant plus de 300 millions de personnes dans le monde. Ces troubles mentaux se caractérisent par une tristesse, une perte d'intérêt ou de plaisir, des sentiments de culpabilité ou de dévalorisation de soi, un sommeil ou un appétit perturbé, une certaine fatigue et des problèmes de concentration [1]. La maladie se décline en plusieurs termes, souvent liés au contexte (par exemple, la dépression post-partum, après la grossesse ; ou la dépression saisonnière, liée au manque de lumière l'hiver) et à la durée des symptômes, qui doivent persister au moins deux semaines pour caractériser une dépression [2]. Elle peut durer de quelques semaines à plusieurs mois voire années. Les conséquences sur l'individu atteint peuvent être multiples et de gravité variable. Parmi celles-ci, on peut citer l'isolement, l'absentéisme au travail, voire même les mutilations ou le suicide. L'importance de venir en aide aux personnes touchées est plébiscitée, et ce à différentes échelles. Dans les entreprises, de plus en plus de mesures sont prises afin d'assurer le bien-être des employés et de réduire ainsi les facteurs de risque liés à la dépression. A moins de consulter un spécialiste, les malades ne sont pas toujours en mesure de réaliser qu'ils sont atteints d'un trouble qui, dans une grande majorité des cas, peut se guérir grâce à un suivi psychologique et/ou à la prescription de médicaments adaptés [3].

1.2 Travaux antérieurs

Ces dernières années ont vu le nombre de travaux relatifs à l'analyse automatique du comportement émotionnel humain progresser de manière significative [4]. Plusieurs tentatives pour modéliser les émotions humaines ont été proposées, dont certaines sont très largement utilisées : une modélisation soit continue (le *circumplex* de Russell [5]), soit discrète (les émotions de base de Ekman [6] : tristesse, joie, colère, peur, dégoût et surprise). L'usage de la vidéo s'est petit à petit imposé comme source de données de choix pour l'analyse émotionnelle, bien qu'historiquement, des procédés plus invasifs aient été préférés, comme l'électrocardiogramme ou la conductance cutanée. Deux cadres d'études se distinguent. Le premier concerne la reconnaissance des émotions et le second, sur lequel nous nous focalisons ici, la prédiction des états dépressifs.
Encourageant les travaux dans ce domaine, des challenges internationaux tels que AVEC [7] ou FERA [8] invitent les chercheurs à confronter leurs méthodes sur une base de données commune. Wen *et al.* [9] ont utilisé des descripteurs visuels dynamiques (LPQ-TOP) associés à une régression par vecteurs supports (SVR) pour diagnostiquer l'état dépressif, avec une erreur RMSE de 8.17 sur le cor-

pus du challenge AVEC'2014. Zhu *et al.* [10] obtiennent une erreur de 9.55, en associant des images de flux optique aux images statiques des visages dans des réseaux de neurones profonds. D'autres approches tiennent compte de la modalité auditive, utilisée notamment pour l'apport d'informations de contexte importantes. Ainsi, Williamson *et al.* [11] ont combiné des descripteurs faciaux (sélection d'unités d'action) et auditifs (durée des phonèmes et analyses des fréquences, notamment) dans des mixtures de modèles gaussiens et obtiennent une erreur de 8.50 sur le corpus AVEC'2014. Gong *et al.* [12] tirent profit des trois modalités visuelle, auditive et contextuelle (retranscription d'interviews) au moyen de régresseurs courants (forêts aléatoires, descente de gradient stochastique et SVM, machine à vecteurs supports) pour une erreur de 4.99 sur le corpus DAIC-WOZ. Si certains travaux accordent une majeure partie de leur effort à la sélection des descripteurs, d'autres optent pour des méthodes connues pour leur capacité à extraire l'information directement depuis les images ou les bandes audios à disposition. Le corpus AVEC'2014 est étiqueté en termes de scores BDI-II et DAIC-WOZ en termes de scores PHQ-8 (*Patient Health Questionnaire, ver. 8*), qui sont deux méthodes d'évaluation de l'état dépressif (voir partie 2).

1.3 Contribution

Le développement d'un système qui, suite à la collaboration d'experts du domaine de la psychiatrie, pourra fournir un diagnostic automatique de l'état dépressif est un axe de bataille offert par l'Intelligence Artificielle pour prévenir l'apparition de tels troubles. En ce sens, l'effort proposé ici est double. Le premier est un outil orienté vers l'usage individuel, permettant à l'utilisateur d'évaluer la sévérité de la dépression dont il souffre. De cette manière, il pourra décider de la suite à donner à l'évaluation en allant consulter un spécialiste. Le second effort s'oriente vers l'usage du système par les experts, notamment pour sa capacité à se spécialiser sur un individu et à augmenter sa précision au fil des entretiens. Ainsi, il disposera d'une aide au diagnostic adaptée à chaque patient.
Le système est basé sur un classifieur neuronal, adapté au traitement de vidéos enregistrées ou capturées en direct. Le traitement produit en sortie un score dépressif, en termes du test *Beck Depression Inventory* II (BDI-II). Dans la partie 2, l'on présentera les données utilisées avant de décrire le classifieur utilisé dans la partie 3. La partie 4 pose les conditions expérimentales retenues dans le cadre de cette étude, et la partie 5 présente les résultats obtenus. Enfin, une conclusion et une ouverture sur des perspectives composera la partie 6.

2 Données

2.1 Corpus AVEC'2014

L'*AudioVisual Emotion Challenge* (AVEC) [7] est un concours international invitant les chercheurs à confronter leurs méthodes et à comparer les performances obtenues sur un même jeu de données.

FIGURE 1 – Image extraite d'une vidéo de AVEC'2014.

Le jeu de données de l'édition 2014 [7] se présente sous la forme de 100 vidéos (tâche *Freeform*) où un individu est en interaction avec un avatar et répond à une question d'ordre général (e.g. comment vous sentez-vous ? pouvez-vous raconter un souvenir d'enfance ?) et de 100 autres (tâche *Northwind*) où l'individu lit un passage écrit, en langue allemande. Ces 200 vidéos sont réparties en deux sous-ensembles : la partition dite de développement (ensemble de motifs pour tester la *généralisation*) et la partition d'apprentissage (ensemble des *exemples* pour construire le modèle). Les vidéos sont constituées de *frames*, en nombre variable (les vidéos n'ayant pas toutes la même durée), enregistrées à raison de 30 par seconde, et contiennent des informations visuelles et auditives. Chaque vidéo est annotée d'un score, celui obtenu au test BDI-II (voir partie 2.2). Une troisième partition, dite partition de test, ne comprend que des vidéos (100 éléments), sans annotations. Les performances prises en compte par les organisateurs du challenge pour départager les participants sont calculées sur cette dernière partition.
Pour l'étude présentée dans cet article, nous retenons les données visuelles des 200 vidéos des tâches Freeform et Northwind. Cela représente un jeu de données de 291 155 images semblables à celles de la Figure 1.

2.2 *Beck Depression Inventory* II

Le test d'évaluation de l'état dépressif *Beck Depression Inventory* (BDI) [13] a été créé par Aaron T. Beck., père de la thérapie cognitive, en 1961. Il a subi plusieurs modifications, visant à l'améliorer. En 1996, sa version II (BDI-II) est un test auto-administré, comptant 21 questions.
Le score obtenu peut prendre une valeur de 0 à 63 ; il donne une indication sur la sévérité de la dépression dont souffre le patient, tel que précisé dans la Table 1. A l'époque de l'apparition du test, il va à contrecourant des pratiques, en

TABLE 1 – Interprétation du score au test BDI-II

Score obtenu	Sévérité de la dépression
0-13	Minimale
14-19	Moyenne
20-28	Modérée
29-63	Sévère

se focalisant sur la perception qu'a le patient de son propre état, plutôt que sur les enjeux psychologiques motivant son comportement et ses réactions à un environnement donné [14] (que l'on appelle, dans la littérature, la psychodynamique). Le test se fonde sur des années de collectes de données, de collaborations entre psychiatres, d'entretiens docteur-malade et de révisions [15]. Le test BDI-II bénéficie d'une corrélation positive avec l'échelle *Hamilton Rating Scale* (HRS) [13], qui est un test administré par un professionnel en psychiatrie. Il est important de noter que dans le processus d'évaluation, le test BDI-II ne tient pas compte de l'expression faciale ou verbale du sujet. L'étude présentée ici démontre qu'il existe bien une forte corrélation entre l'expression faciale et l'état dépressif puisque le classifieur construit à partir des visages extraits des vidéos permet de prédire de manière fiable la sévérité dépressive.

2.3 Extraction des descripteurs et changement de repère

Afin de classifier les vidéos selon leur score BDI-II, on extrait, pour chaque image, un ensemble de 68 points faciaux d'intérêt (voir Figure 2). Cela nécessite, en amont, la détection et le redimensionnement des visages. L'extracteur de points d'intérêts utilise le modèle de Kazemi et Sullivan [16]. Le détecteur de visages (entraîné sur l'ensemble i-BUG 300-W, voir Sagonas *et al.* [17]) implémente un classifieur linéaire sur une pyramide d'images dans des fenêtres temporelles, ainsi que sur des histogrammes de gradients orientés. L'outil Dlib [18] a été utilisé pour mettre en œuvre l'extraction.

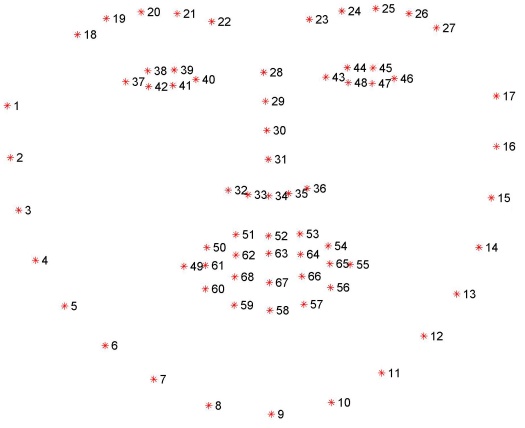

FIGURE 2 – Points faciaux d'intérêt - Modèle MultiPIE.

L'alignement des visages est également une étape importante [19], qui permet notamment de limiter le biais introduit par des facteurs tels que la distance à la webcam ou la morphologie faciale du sujet, mais aussi d'homogénéiser les données avant la classification. Afin d'aligner les yeux, de centrer les visages dans l'image et d'homogénéiser leur taille, les points subissent une translation de vecteur $\overrightarrow{T}$, ainsi qu'une rotation d'angle θ (facteur d'échelle : S). Cette transformation isométrique est un cas particulier de similitude calculée pour chaque visage. La matrice de transformation M est donnée par l'équation 1. Ces étapes sont réalisées au moyen de la librairie OpenCV [20].

$$M = \begin{bmatrix} s_x cos(\theta) & sin(\theta) & t_x \\ -sin(\theta) & s_y cos(\theta) & t_y \end{bmatrix} \quad (1)$$

3 Modèle à base de prototypes

Le choix du classifieur incrémental pour prédire l'état dépressif a été motivé à la fois par les inspirations biologiques sous-jacentes, comme démontré par Grossberg dès la fin des années 80 (voir [21] pour une synthèse ce sujet) et par le récent regain d'intérêt pour les modèles à base de prototypes : Biehl, Hammer et Villmann [22] affirment en 2016 que de tels systèmes sont très intéressants pour l'analyse de données complexes et de grande dimension.

3.1 Le classifieur incrémental

Le modèle ART (*Adaptive Resonance Theory*) de Grossberg [23] est un système de classification neuronal capable de s'adapter aux entrées dites significatives, tout en restant stable face aux entrées non-significatives. Ainsi, si l'on présente au système un exemple proche d'une représentation qu'il connaît, il la modifiera en conséquence. En revanche, si on lui présente un exemple inconnu, une nouvelle représentation sera créée pour le prendre en compte.
Le classifieur incrémental utilisé ici est inspiré du modèle ART et suit le même principe. Il a été proposé par Azcarraga [24] puis modifié par Puzenat [25], qui l'utilisait pour la reconnaissance de formes manuscrites. Il s'agit d'un réseau de neurones dont la couche d'entrée est, classiquement, adaptée à la dimension de l'espace des données. La seconde couche est constituée de "neurones-distance", les prototypes, qui sont totalement connectés aux neurones d'entrée. Ainsi, à chaque présentation d'un exemple, celui-ci est comparé à tous les prototypes en mémoire. Dans la troisième couche, chaque neurone est connecté à un seul et unique prototype (voir Figure 3) ; aucun apprentissage n'est effectué par la couche de sortie.

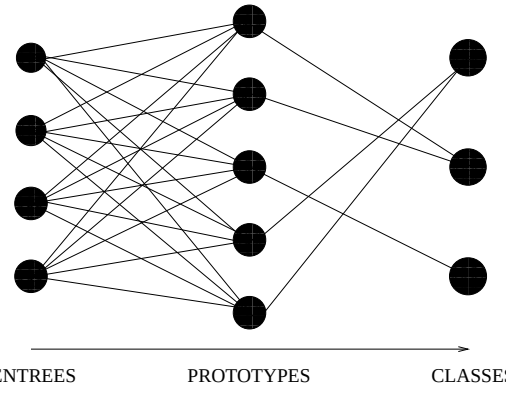

FIGURE 3 – Architecture du classifieur incrémental.

Selon le protocole précisé ci-dessous (partie 3.2), de nouveaux prototypes sont créés pendant le processus d'appren-

tissage. Néanmoins le fait que leur nombre, qui dépend de la taille de la base d'apprentissage ainsi que du nombre de classes, reste faible par rapport au nombre d'exemples garantira que le classifieur aura extrait des données une information synthétique. Les prototypes sont représentés dans le même espace que celui des données, ce qui rend aisée leur interprétation, ainsi que leur appréhension par des experts dans le cadre d'un travail transversal [22].

3.2 Apprentissage

La phase d'apprentissage consiste en une seule passe de la base d'exemples et elle suit un algorithme par compétition. Initialement, le premier exemple est recopié comme unique prototype et on l'associe en sortie à la classe de l'exemple. Par la suite, pour chaque nouvel exemple présenté X, on cherche le prototype $P_{meilleur}$ qui en est le plus proche, au sens de la mesure choisie (voir discussion ci-dessous). *A priori*, si la classe de $P_{meilleur}$ est celle de l'exemple, le prototype est gagnant et sa connexion avec l'exemple est modifiée afin de l'en rapprocher. Sinon, un nouveau prototype est créé à l'image de l'exemple et est associé en sortie à la classe de l'exemple.
Une exception à l'adaptation de $P_{meilleur}$ relève d'une condition plus subtile : on cherche, parmi les prototypes les plus proches de l'exemple, le premier prototype dont la classe est différente de celle de $P_{meilleur}$, et on le nomme P_{second}. S'il y a risque de confusion, i.e. dans une zone de l'espace d'entrée ou des motifs proches doivent être associés à des classes distinctes, alors un nouveau prototype est créé.
Le sens de proximité se réfère ici à une mesure de distance ou de similarité entre un exemple et un prototype. La mesure est choisie en accord avec le problème à traiter, ce qui rend les classifieurs incrémentaux flexibles et adaptables. On peut utiliser une distance de Mahalanobis, qui accorde un poids moins important aux composantes les plus dispersées, ou une distance de Minkowski (équation 2) qui permet un calcul plus rapide.

$$d_{mink_p} = \left(\sum_{i=1}^{n} \mid x_i - y_i \mid^p \right)^{1/p} \tag{2}$$

Pour $p = 2$, on retrouve la distance euclidienne qui est particulièrement adaptée aux situations où les descripteurs sont des coordonnées, comme c'est le cas pour les points faciaux d'intérêt. Après avoir vérifié que des valeurs $p > 2$ ne produisaient pas de résultats significativement meilleurs, nous avons opté pour la distance euclidienne et nous la noterons d.

3.3 Hyperparamètres de contrôle

L'algorithme d'apprentissage du classifieur incrémental propose trois hyperparamètres de contrôle afin de répondre au dilemme stabilité-plasticité (*stability-plasticity dilemma*), c'est-à-dire de tenir compte des nouveaux éléments à apprendre sans oublier ceux déjà mémorisés.

Seuil d'influence. Le seuil d'influence s_{inf} du modèle définit la valeur à laquelle doit être inférieure la distance entre un exemple et son meilleur prototype, tel que décrit par l'équation 3. Si cette condition n'est pas vérifiée, on crée un nouveau prototype.

$$d(P_{meilleur}, X) \leq s_{inf} \tag{3}$$

Seuil de confusion. Le seuil de confusion s_{conf} aide à lever les ambiguités dans les zones frontières entre classes distinctes et s'utilise comme décrit par l'équation 4. Si cette condition n'est pas vérifiée, on crée un nouveau prototype.

$$d(P_{meilleur}, X) - d(P_{second}, X) > s_{conf} \tag{4}$$

Coefficient de rapprochement. Lorsqu'aucun nouveau prototype n'est créé, l'adaptation de $P_{meilleur}$ à l'exemple X est contrôlée par le coefficient de rapprochement α. L'équation 5 décrit la modification des poids du neurone prototype.
Pour i tel que P_i = $P_{meilleur}$,

$$\forall j, w_{ji} \leftarrow w_{ji} + \alpha(x_j - w_{ji}) \tag{5}$$

Plus ce coefficient α est grand, plus la représentation modélisée par $P_{meilleur}$ se rapproche de l'exemple et plus on accroît la création de prototypes. A *contrario*, pour un coefficient petit, $P_{meilleur}$ sera peu modifié et le nombre de prototypes restera limité. On note que, pour $\alpha = 0.5$, on calcule le barycentre entre les deux entités.

3.4 Généralisation

La généralisation respecte les mêmes contraintes que l'apprentissage, mais il n'y a plus de création ni de modification de prototypes. Pour chaque nouveau motif n'ayant pas participé à l'apprentissage du modèle :

1. **Présenter** un motif X
2. **Rechercher** $P_{meilleur}$ tel que $d(P_{meilleur}, X)$ soit minimale
3. **Rechercher** P_{second} tel que la classe de P_{second} soit différente de celle de $P_{meilleur}$
4. **Si** $P_{meilleur}$ est trop éloigné de X ou s'il y a risque de confusion :

$$\begin{cases} d(P_{meilleur}, X) > s_{inf} \\ \text{ou} \\ d(P_{meilleur}, X) - d(P_{second}, X) \leq s_{conf} \end{cases}$$

Alors rejeter X (non-réponse)
Sinon Si la classe de $P_{meilleur}$ est celle de X,
Alors X est reconnu (bonne réponse)
Sinon X n'est pas reconnu (mauvaise réponse)

3.5 Non-réponses

Le classifieur incrémental est en mesure de produire, en sortie, trois types réponses : une "non-réponse", une

"bonne-réponse" ou une "mauvaise réponse". En particulier, une non-réponse est rendue lorsque le meilleur prototype $P_{meilleur}$ de l'exemple d'entrée est trop éloigné de ce dernier, ou lorsque la distance entre $P_{meilleur}$ et P_{second} est trop faible au regard de l'exemple. Cette réponse, en plus de se rapprocher du diagnostic que ferait un humain, peut être considérée comme un indicateur de fiabilité du score dépressif calculé pour un sujet donné (voir 5.3). Dans le cas où le système produirait un grand nombre de non-réponses, la classification pourrait être jugée peu fiable.Le cas échéant, le système peut être spécialisé sur l'individu, via un réapprentissage du modèle, sur décision d'un expert. Cette possibilité d'obtenir une "non-réponse" est une spécificité précieuse de ce type de classifieur, le rendant plus proche d'un diagnostic humain.

4 Conditions expérimentales

Le classifieur incrémental est entraîné pour associer à chaque image (cf. 2.1) le score BDI-II de la vidéo dont elle a été extraite. Afin de réduire les risques liés au sur-apprentissage, et pour disposer d'un plus grand nombre de classes représentées, nous avons mélangé les partitions de développement et d'apprentissage dont nous disposions. De plus, les exemples ont été stratifiés afin que chaque classe soit toujours représentée en quantité raisonnable dans les ensembles d'apprentissage et de généralisation.
A priori, il conviendrait d'apprendre un modèle en régression pour lire en sortie la valeur du score. Cependant les différentes valeurs sont en nombre limité (seulement 41 présentes dans les vidéos étudiées, parmi les 64 valeurs possibles en théorie) et chacune sera considérée comme une classe. Il est important de noter que le système ne sera pas en mesure de discriminer, en généralisation, une classe inexistante dans les données d'apprentissage. De plus, compte tenu de la Table 1, on pourra *a posteriori* regrouper les scores numériques dans des intervalles pour qualifier de manière descriptive la sévérité de la dépression.

4.1 Stratégies de test

Nous retenons trois stratégies pour les expériences :

- **Classique** : construction d'un modèle sur une base d'apprentissage puis estimation de la performance en généralisation sur une base disjointe ;
- **Validation croisée** : une partition $S = \cup_{m=1}^{M} S_m$ de la base de données S étant réalisée, apprentissage de M modèles, chacun sur $S = \cup_{k \neq m} S_k$, avec estimation de sa performance en généralisation sur S_m ;
- **Flux continu** : un modèle ayant été appris, utilisation en temps-réel pour prédire l'état dépressif d'un individu placé devant une webcam.

La stratégie "classique" a été mise en œuvre en premier, afin d'étudier le comportement du modèle et de valider les choix de prétraitements et d'extraction des descripteurs (présentés en 2.3). Au fil de ces expérimentations, la taille de la base de généralisation a été fixée à 3/10e des données, distinctes des 7/10e ayant servi à entraîner le modèle, comme le récapitule la Table 2.

TABLE 2 – Composition des ensembles de données pour la stratégie classique

	Freeform	Northwind		**Total**
Apprentissage	113 876	89 933		**203 809**
Généralisation	48 945	38 401		**87 346**
Total	162 821	128 334		**291 155**

Les meilleurs hyperparamètres pour le modèle ont été déterminés au moyen d'une recherche en grille (*grid search*) pour tenir compte des interactions entre hyperparamètres. Les résultats présentés ci-dessous ont été obtenus avec un seuil d'influence de 70, un seuil de confusion de 0.1 et un coefficient de rapprochement de 0.1.

4.2 Mesures de performances

Afin d'évaluer les performances de notre approche, nous retenons quatre indicateurs, dont deux estiment un taux de succès en classification et deux autres mesurent une erreur :

- Le taux de succès, en termes de **score** BDI-II
- Le taux de succès au sens des **intervalles** de sévérité de dépression (cf. Table 1)
- L'erreur quadratique moyenne (RMSE)
- L'erreur absolue moyenne (MAE)

Ces deux derniers indicateurs sont bien adaptés aux cas de la classification multi-classe, particulièrement lorsque les classes sont hétérogènes en nombre de données. Pour les taux de succès, il est important de noter que ces indicateurs seront calculés en tenant compte des images bien classées, et non des vidéos dans leur ensemble.

5 Résultats

5.1 Entraînement et validation croisée

Les meilleures performances obtenues dans le cadre de la stratégie "classique" sont présentées dans la Table 3 pour les taux de succès et la Table 4 pour les indicateurs d'erreur. Notons que ces résultats, en particulier les RMSE, ne sont pas directement comparables avec ceux du challenge AVEC'2014 cités en 1.2 dans la mesure où nous n'avons pas accès à la partition de test réservée aux organisateurs du challenge, et où nous n'avons pas choisi le même partitionnement des données pour nos essais.

TABLE 3 – Taux de succès pour la stratégie classique

	Freeform		Northwind	
	Bon score	Bon intervalle	Bon score	Bon intervalle
App.	92.43%	95.25%	92.29%	95.40%
Gén.	89.78%	93.70%	91.01%	94.52%

Cette stratégie oblige à construire le modèle en n'utilisant qu'une partie des données (70% ici). En revanche, la stratégie "validation croisée" permet, après estimation moyenne

TABLE 4 – Les erreurs pour la stratégie classique

	Freeform		Northwind	
	RMSE	MAE	RMSE	MAE
App.	4.07	0.83	3.79	0.8
Gén.	4.67	1.11	4.05	0.92

de la performance en généralisation sur M modèles, de construire un $M + 1^{eme}$ modèle qui apprend sur toutes les données à disposition. La Table 5 donne les performances pour une validation croisée avec $M = 10$, où l'algorithme apprend sur 262 040 exemples et généralise sur les 29 115 restants. En effet, les bases *Freeform* et *Northwind* ont été mélangées puisque la stratégie "classique" a démontré la similarité de leur comportement.

TABLE 5 – Performances en validation croisée

	Bon score	Bon intervalle	RMSE	MAE
Moyenne	90.73%	94.51%	4.30	0.97

On note un gain de performance d'environ 4 % en passant du nombre de bien classés par score BDI-II au nombre de bien classés par intervalle de sévérité dépressive. Cette amélioration confirme l'existence d'une continuité entre états dépressifs de sévérité proche, et témoigne de la capacité du système à la saisir. Le nombre de prototypes est de l'ordre de 15% à 18% du nombre d'exemples.

5.2 Comparaison avec d'autres classifieurs

TABLE 6 – Comparaison des performances des classifieurs de la littérature

	Bon score	Temps de classification (en sec., pour un ex.)
SVM	73.23 %	0.009
MLP	66.98 %	**0.001**
Random Forest	**94.87 %**	0.118
C. Incrémental	90.25%	0.023

Les performances du classifieur incrémental sont comparées aux performances de classifieurs de la littérature dans la Table 6. Les taux de succès ont été obtenus en généralisation sur 30% des données via la stratégie classique exposée en 4.1, après un apprentissage sur 70% de la base complète. Les temps de réponse des classifieurs à un nouveau motif présenté ont aussi été mesurés. Le CI n'est pas le plus performant en termes de taux de succès, mais présente le meilleur compromis entre qualité et rapidité de la réponse. Ce point est essentiel dans le cadre d'un outil d'aide au diagnostic puisque le système doit pouvoir donner une estimation en flux continu.

5.3 Flux continu

À l'issue de la validation croisée, le classifieur a été entraîné sur l'ensemble des 291 155 images disponibles. Les performances à considérer sont celles obtenues en moyenne (voir Table 5). Il peut désormais être utilisé en prédiction pour fournir une estimation automatique de l'état dépressif d'un individu faisant face à une simple webcam [26].

La fréquence des images est de 30 par seconde lors de la capture. Cependant, l'expression dépressive s'évaluant sur la durée, il n'est pas nécessaire de traiter toutes les images produites. On fixe un nombre d'images n à traiter par secondes (par exemple, $n = 10$) ainsi qu'une durée d'enregistrement. Les images sont prétraitées et les descripteurs extraits comme décrit dans la partie 2.3. Chacune est alors comparée aux prototypes par l'algorithme de généralisation (cf. partie 3.4).

La sortie du système est une valeur d'état dépressif du sujet filmé estimée par le score BDI-II majoritaire sur une période p donnée en secondes (par exemple : $p = 20$). Notons au passage que cette procédure permet d'effacer au fur et à mesure les données personnelles qui n'auront été enregistrées que temporairement. Comme suggéré dans la partie 3.5, les non-réponses pourront être à terme exploitées comme indicateur de fiabilité du classifieur.

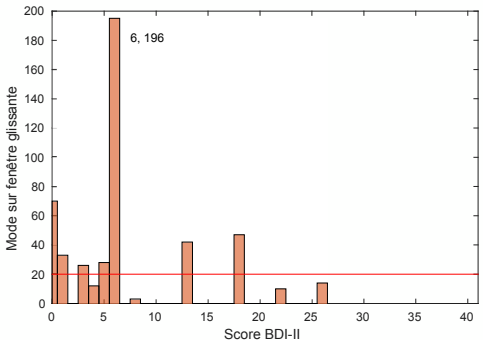

FIGURE 4 – Distribution des scores proposés par la classification en flux continu

Il est important de noter que, dans ce dernier contexte, il ne nous est pas encore possible d'évaluer de réelles performances. Par exemple, la pertinence des scores BDI-II fournis par le système ne pourra être validée que lorsqu'un protocole expérimental sera mis en place, en collaboration avec un expert humain (voir partie 6). Néanmoins, la faisabilité du traitement *on-line* a été établie par l'un des auteurs de cet article : en filmant son propre visage, il a obtenu un score stable de 6 sur une période d'une vingtaine de secondes, avec un nombre de non-réponses de 20 sur 500 matérialisé par la ligne horizontale sur la Figure 4.

6 Conclusion et discussion

Nous avons proposé un classifieur incrémental à base de prototypes afin de déterminer l'état dépressif d'un individu à partir d'une vidéo. Le prétraitement des images permet de réduire fortement le biais introduit par les différences d'échelle et les spécificités morphologiques des sujets. La classification rapide autorise, sous couvert de validation par un expert, le développement d'un module de classification en temps-réel de l'état dépressif, en capturant le flux vidéo directement via une webcam.

Le système pourra facilement être utilisé par un praticien comme outil d'aide au diagnostic et de suivi de patient, ce dernier pouvant lui-même effectuer des évaluations de son état à l'aide d'un matériel peu coûteux. Si cela s'avère nécessaire, l'outil pourra être ré-étalonné (phase d'apprentissage complémentaire) pour mieux s'adapter à un patient précis. Sur le plan technique, notons cependant que l'accroissement du nombre de prototypes aura pour effet de ralentir le traitement. Pour cela, nous proposons en perspective l'étude d'une procédure d'élagage, visant à réduire le nombre de prototypes. Ce type de procédure va de paire avec tout système incrémental, et est en phase active de développement, raison pour laquelle elle n'est pas présentée dans cet article.

Notons enfin que la plupart des travaux sur la dépression utilisent à la fois les modalités visuelle et auditive, à l'instar de *Yu et al.* [27]. La prochaine étape de ce travail consistera à entraîner, de manière indépendante et sur le même modèle, un classifieur permettant de prédire l'état dépressif à partir des données audio uniquement. La mise en commun des deux modèles pourra ensuite se faire au moyen d'un modèle de mémoire associative multimodale qui réalise la fusion des données à l'aide d'une *Bidirective Associative Memory* (BAM). Le modèle complet a déjà été développé [28], sur la base d'une modélisation cognitive, et il a démontré l'amélioration des performances par la prise en compte de plusieurs modalités [29].

Remerciements

Le travail décrit dans cet article a été réalisé en Python. En ce sens, ses auteurs souhaitent remercier les contributeurs de Numpy [30], Scipy [31] et Scikit-learn [32].

Références

[1] (2018) Depression. Organization World Health. Accessed 2018-04-02. [Online]. Available : http ://www.who.int/mediacentre/factsheets/fs369/fr/

[2] (2018) Depression. Health National Institute of Human. Accessed 2018-04-02. [Online]. Available : https ://www.nimh.nih.gov/health/topics/depression/

[3] R. J. DeRubeis, G. J. Siegle, and S. D. Hollon, "Cognitive therapy vs. medications for depression : Treatment outcomes and neural mechanisms," *Nat. Rev. Neurosci.*, no. 10, pp. 788–796, oct.

[4] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods : Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[5] J. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[6] P. Ekman, "Differential Communication Of Affect By Head And Body Cues.pdf," *J. Pers. Soc. Psychol.*, vol. 2, no. 5, pp. 726–735, 1965.

[7] F. Ringeval, M. Pantic, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, and M. Schmitt, "Avec 2017 - Real-life Depression, and Affect Recognition Workshop and Challenge," *Proc. 7th Annu. Work. Audio/Visual Emot. Chall. - AVEC '17*, pp. 3–9.

[8] M. F. Valstar, E. Sanchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang, L. Yin, and M. Pantic, "FERA 2017 - Addressing Head Pose in the Third Facial Expression Recognition and Analysis Challenge," *Autom. Face Gesture Recognit. (FG 2017)*, pp. 839–847, 2017.

[9] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 7, pp. 1432–1441, 2015.

[10] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated Depression Diagnosis based on Deep Networks to Encode Facial Appearance and Dynamics," *IEEE Trans. Affect. Comput.*, no. X, pp. 1–1.

[11] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and Facial Biomarkers of Depression based on Motor Incoordination and Timing," *Proc. 4th Int. Work. Audio/Visual Emot. Chall. - AVEC '14*, pp. 65–72.

[12] Y. Gong and C. Poellabauer, "Topic Modeling Based Multi-modal Depression Detection," *Proc. 7th Annu. Work. Audio/Visual Emot. Chall. - AVEC '17*, pp. 69–76.

[13] A. T. Beck, R. A. Steer, and G. K. Brown, "Beck depression inventory-II," *San Antonio*, vol. 78, no. 2, pp. 490–498, 1996.

[14] A. T. Beck, *Depression : Causes and Treatment*. University of Pennsylvania Press, 1972.

[15] L. R. Aiken, *Psychological Testing and Assessment, 4th edition*. Allyn & Bacon, 1982.

[16] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1867–1874.

[17] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 Faces In-The-Wild Challenge : database and results," *Image Vis. Comput.*, pp. 3–18.

[18] D. E. King, "Dlib-ml : A Machine Learning Toolkit," *J. Mach. Learn. Res.*, pp. 1755–1758.

[19] A. Rosebrock. (2017) Face Alignment using OpenCV and Python. Accessed 2018-05-24. [Online]. Available : https ://www.pyimagesearch.com/2017/05/22/face-alignment-with-opencv-and-python/

[20] G. Bradski, "The OpenCV Library," *Dr Dobbs J. Softw. Tools*, pp. 120–125.

[21] G. A. Carpenter and S. Grossberg, "Adaptive Resonance Theory," *Handb. brain theory neural networks*, pp. 87–90, 2003.

[22] M. Biehl, B. Hammer, and T. Villmann, "Prototype-based models in machine learning," *Wiley Interdiscip. Rev. Cogn. Sci.*, vol. 7, no. 2, pp. 92–111, 2016.

[23] S. Grossberg, "Adaptive Resonance Theory : How a brain learns to consciously attend, learn, and recognize a changing world," *Neural Networks*, pp. 1–47.

[24] A. P. Azcarraga, "Modèles neuronaux pour la classification incrémentale de formes visuelles," Ph.D. dissertation, Genoble INPG.

[25] D. Puzenat, "Parallélisme et modularité des modèles connexionnistes," p. 176.

[26] S. Cholet and H. Paugam-Moisy, "Démonstration du diagnostic automatique de l ' état dépressif," in *Conférence Natl. en Intell. Artif.* Nancy : Plateforme Intelligence Artificielle, 2018, p. To Appear.

[27] S. Yu, S. Scherer, D. Devault, J. Gratch, G. Stratou, L. P. Morency, and J. Cassel, "Multimodal prediction of psychosocial disorders : Learning verbal and non-verbal commonalities in adjacenc pairs," in *Proc. 17th Work. Semant. Pragmat. Dialogue*, 2013, pp. 160–169.

[28] E. Reynaud, A. Crépet, H. Paugam-Moisy, and D. Puzenat, "A computational model for binding sensory modalities," in *Abstr. Conscious. Cogn.* Academic Press, 2000, ch. 9, pp. 97–88.

[29] H. Paugam-Moisy and E. Reynaud, "Multi-network system for sensory integration," *Int. Jt. Conf. Neural Networks, Vols 1-4, Proc.*, vol. 1-4, no. February 2001, pp. 2343–2348, 2001.

[30] T. E. Oliphant, *A Guide to Numpy*. Trelgol Publishing, 2006.

[31] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy : Open source scientific tools for Python," 2001–. [Online]. Available : http ://www.scipy.org/

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn : Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.