ETUDE DES ALGORITHMES DE CLASSIFICATION AUTOMATIQUE POUR ZOOCAM 2018

Grâce au logiciel « Etude classifieurs»

Fiche documentaire

Titre du rapport : Etude des algorithmes de classification automatique pour Zoocam	
Référence interne : REM/RDT/LCSM 2018-144 **Diffusion :** ☐ libre (internet) ☐ restreinte (intranet) – date de levée d'embargo : 2023/08/29 ☒ interdite (confidentielle) – date de levée de confidentialité : 2023/08/29	**Date de publication :** 2018/08/29 **Version :** 1.0.0 **Référence de l'illustration de couverture** Crédit photo/titre/date **Langue(s) : Française**
Résumé/ Abstract : L'étude démontre l'intérêt de l'utilisation des algorithmes de réseau de neurones dans la classification de vignettes générées par Zoocam. Cepandant, certaines caractérisques des échantillons d'apprentissages et l'ajustement de paramètres sont à évaluer pour rendres ces algorithmes performants.	
Mots-clés/ Key words : Classification automatique, intelligence artificielle, réseaux de neurones, Random Forest, Zoocam 2018	
Comment citer ce document :	
Disponibilité des données de la recherche :	
DOI :	

Commanditaire du rapport :	
Nom / référence du contrat : ☐ Rapport intermédiaire (réf. bibliographique : XXX) ☐ Rapport définitif (réf. interne **du rapport intermédiaire** : R.DEP/UNIT/LABO AN-NUM/ID ARCHIMER)	
Projets dans lesquels ce rapport s'inscrit (programme européen, campagne, etc.) :	
Auteur(s) / adresse mail	**Affiliation / Direction / Service, laboratoire**
Bertrand Forest / bforest@ifremer.fr	REM/RDT/LCSM
Encadrement(s) :	
Destinataire :	
Validé par :	

Sommaire

Table des matières

1. Présentation de l'étude

1. Avant propos

Ce document ne contient pas d'introduction aux principes de fonctionnement des algorithmes de classification *Random Forest* et *Réseau neuronaux*, car cela à déjà été rédigé, et exposé.

2. Introduction

Le logiciel ZooCam 2018 utilise l'algorithme *Random Forest* pour la classification des vignettes. Cet algorithme est actuellement assez satisfaisant, mais depuis quelques années, d'autres algorithmes de classification tels que les réseaux de neurones pourraient être plus efficaces. Il est donc intéressant d'évaluer leurs performances, avantages et inconvénients On sait que ces derniers sont gourmands en puissance de calculs lors de l'étape d'apprentissage, cependant, la classification est très rapide. L'évolution technologique des cartes GPU (composés plusieurs milliers de processeurs), rend les temps de calculs de l'apprentissage des réseaux de neurones raisonnables tandis que leurs coûts tendent à la baisse.

Les résultats de cette étude pourront être employés dans d'autres domaines que la biologie, aussi je parlerai de « catégories » dans le logiciel et de « taxons » pour l'application à la biologie.

3. Les différents algorithmes de classifications évalués

Dans cette étude, je vais évaluer 4 méthodes de classification:

- Le **Random Forest (RF)** qui a pour entrées les 50 paramètres morphologiques des individus issus du *Process* de Zoocam (*fichier learning.pid)*
- Le **Deep Neural Network (DNN)** , les mêmes 50 paramètres.
- Le **Convolutional Neural Network (CNN)**. L'algorithme ne traite ici que les pixels des vignettes des individus.
- Le **Random Forest Enrichi** au CNN avec deux variantes :
    - Variante A : Lorsqu'un individu est prédit avec RF, on regarde quel aurait été le 2eme choix suivant les plus hautes probabilités. Si l'écart est faible (doute) et si la prédiction avec CNN correspond à l'une des deux premières prédictions de RF, alors la prédiction finale revient à CNN.
    - Variante B : Le Ramdom Forest et ses 50 paramètres morphologies sont complétés des sorties CNN (probabilités de chaque catégories). Bien entendu le temps de calcul est légèrement plus long.

4. Le logiciel «Etude Classifieurs »

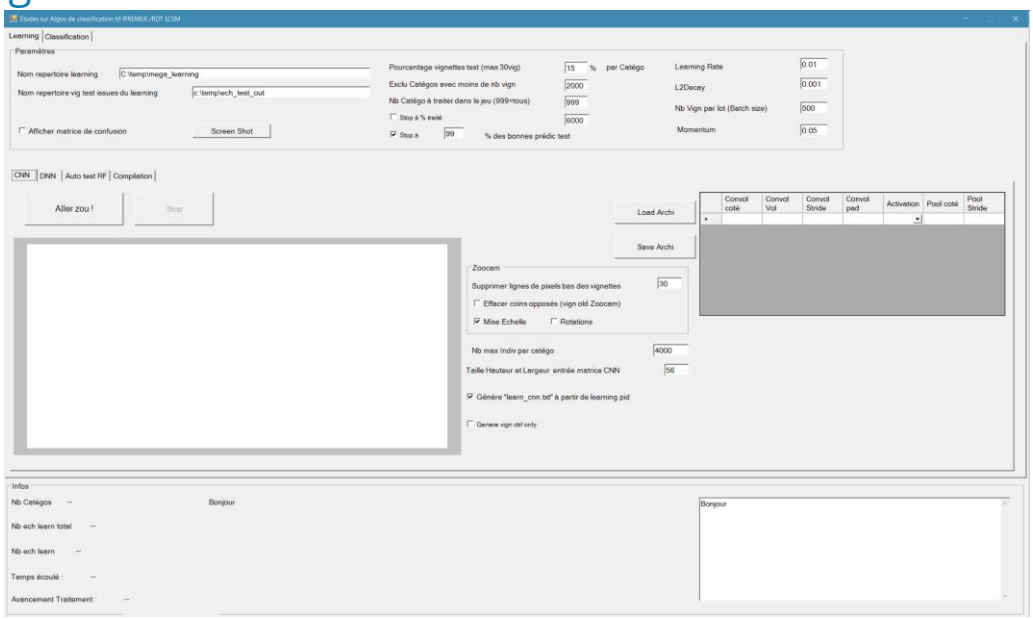


Fig 1 : Page principale du logiciel

Le logiciel permet :

- D'effectuer l'apprentissage des classifieurs CNN, RF et DNN , d'évaluer leur performance par rapport à des individus test, de sauvegarder les états des réseaux neuronaux.
- CNN recherche la présence d'un fichier learning PID et s'il existe, il y a création d'un nouveau learning constitué des bonnes vignettes prédites par CNN. Ce nouveau learning pourra être utilisé par RF et DNN.
- De sélectionner un nombre d'individus mini et maxi par catégorie
- D'afficher la progression, l'état de prédiction du jeu test, la matrice de confusion
- D'effectuer des prédictions suivant toutes les méthodes sur un échantillon totalement indépendant.
- De générer un « MegaLearning» (assemblage d'échantillons validés) et de mettre à jour le fichier PID ainsi obtenu après déplacements de vignettes par l'opérateur. Il est possible de limiter le nombre d'individus dans chaque catégorie.
- De comparer un répertoire d'un échantillon prédit avec cet échantillon validé, et d'afficher les statistiques par groupes de catégories.
- De pouvoir traiter les anciennes vignettes Zoocam : effacer les 2 « coins », d'effectuer 3 rotations par vignettes.
- A chaque fin de calcul, un fichier *« c:\temp\ log_frm_control.txt »* est généré et contient les statistiques que nous retrouverons dans ce document. Chaque ligne correspond ainsi :
  Nom de la catégorie [nb individus bien prédits parmi les individus test / nb d'individus test / pourcentage]

5. Origines des algorithmes

Jonathan Perchoc, premier développeur du Zoocam, a donné une indication sur l'origine de son code Random Forest:
http://semanticquery.com/archive/semanticsearchart/downloads/RFtest.zip
mais ce lien n'est pas sécurisé selon mon navigateur WEB. En tapant sur Google « semanticsearchart » on arrive à http://ezcodesample.com/SemanticSearchArt/researchRF.html où un code C# est téléchargeable. J'y ai apporté quelques modifications :

- les données sont centrées/réduites (en fait soustraites au minimum et divisées par l'amplitude pour avoir les valeurs entre zéro et un)
- la détection et suppression des arbres très peu contributeurs à la décision.

Pour le CNN et le DNN, je me suis inspiré du code C# « ConvNetSharp» disponible sur le site web du Github https://github.com/cbovar/ConvNetSharp sous licence MIT. Ce code permet d'utiliser un GPU NVIDIA pour effectuer les calculs et travailler en simple et double précision. Le gros défaut de ce code est qu'il n'est absolument pas documenté. Une bonne initiation aux paramètres des réseaux de neurones est nécessaire avant d'y comprendre quelque chose. Le deuxième problème est que l'on ne peut pas anticiper la consommation de mémoire GPU nécessaire à l'apprentissage suivant les paramètres employés : un plantage du logiciel survient alors. Un jeu de test est fourni : MNIST (100000 vignettes 28*28 pixels correspondant aux 10 chiffres en écriture manuscrite)

Parallèlement à l'écriture de ce logiciel, j'ai passé quelques heures en avec le couple Python-TensorFlow de Google. J'ai testé un code assez simple avec un learning composé de 10000 vignettes en couleur de chiens et autant d'autres de chats. Le résultat est très décevant : 65% de bonnes prédictions, je n'ai pas poursuivi dans cette voie. J'ai aussi exploré le site de Microsoft Azure : à ce que j'ai plus ou moins compris, on peut gratuitement faire des tests d'apprentissages mais dans des domaines déjà bien ciblés comme la traduction automatique ou la détection de visages. (https://gallery.azure.ai/machineLearningAPIs), je n'ai pas trouvé d'applications dans notre cadre de recherches.

6. Jeux de données

Je dispose du jeu de données de PELGAS 2018 leg 1 produits sur le navire La Thalassa de l'IFREMER., Il y a de 318 échantillons CUFES, mais il y a peu œufs de sardines par rapport aux œufs d'anchois. Normalement, tous les échantillons sont validés, particulièrement pour les œufs, c'est en effet le sujet de la campagne PELGAS.

7. Learnings

Dans toutes les études de learning, 15% des individus de chaque catégorie sont réservés pour le test de prédiction, mais il ne peut pas y en avoir plus de 100 pour CNN pour éviter de soustraire trop de précieux individus dédiés à l'apprentissage.

a. Pelgas 2018 Martin Huret

Martin Huret a créé à bord du navire La Thalassa pendant la campagne Pelgas 2018 Leg 1, un learning composé de 6842 individus répartis en 23 taxons.

Taxons	Effectifs
00_Artefact	60
00_Bubble	322
00_Bubbles_large	201
00_Detritus	759
00_Fiber	323
00_Halosphera	121
00_Paint_rust	45
03_Copepoda_large	914
03_Copepoda_small	815
04_Malacostraca_large	290
05_Cladocera	427
13_Engraulis_egg_1	318
13_Engraulis_egg_2_3	297
13_Engraulis_egg_4_6	337
13_Engraulis_egg_7_8	314
13_Engraulis_egg_9_11	89
13_Fish_egg	299
13_Sardine_egg_1	10
13_Sardine_egg_2_3	345
13_Sardine_egg_4_6	304
13_Sardine_egg_7_8	17
13_Sardine_egg_9_11	22
13_Sardine_egg_dam	213

Tableau 1 : inventaire du learning « Pelgas 2018 Martin Huret »

Comme on peut le voir, dans le tableau 1, il y a peu d'œufs de sardines, ce qui est contrariant pour le CNN, et même pour les autres classifieurs quand il y a moins de 150 individus.

b. Pelgas 2018 MegaLearning

Une fonctionnalité du logiciel d'étude permet de générer un « MegaLearning » composé de tous les échantillons validés de Pelgas 2018 leg 1 sauf l'échantillon 308 (choisi parce

qu'il contient un nombre équilibré d'œufs d'anchois et de sardines, contenu listé en annexe). C'est échantillon est validé par plusieurs personnes et sert de test de prédictions indépendant dans cette étude. Dans certaines catégories de PELGAS 2018 leg 1, le nombre d'individus peut être inutilement très important (copépodes, débris), je l'ai limité à 3000 (valeur modifiable). Certaines catégories sont peu validées : artéfact, bulles… car les validateurs se sont surtout concentrés sur les œufs. Aussi j'ai passé un peu de temps à déplacer les vignettes d'un répertoire à l'autre particulièrement dans les débris et bulles. Il faut alors mettre à jour la colonne « validation » du fichier *learning.pid*, ce qui est prévu dans le logiciel pour le faire d'un clic de souris. Ce learning comporte ainsi 46296 individus avec 32 catégories (Tableau 2).

Taxons	Effectifs
00_Artefact	2 037
00_Artefact_2	106
00_Bubble	3 004
00_Bubbles_large	1 777
00_Detritus	2 820
00_Fiber	2 995
00_Halosphera	882
00_Paint_rust	343
00_Phytoplankton	321
00_Spirale	18
03_Copepoda_large	2 991
03_Copepoda_Pontellidae	161
03_Copepoda_small	2 999
04_Malacostraca_large	2 998
05_Cladocera	2 764
06_Cnidaria	24
13_Engraulis_egg_1	3 000
13_Engraulis_egg_2_3	3 000
13_Engraulis_egg_4_6	3 000
13_Engraulis_egg_7_8	3 000
13_Engraulis_egg_9_11	367
13_Fish_egg	2 393
13_Sardine_egg_1	59
13_Sardine_egg_2_3	1 915
13_Sardine_egg_4_6	1 125
13_Sardine_egg_7_8	205
13_Sardine_egg_9_11	179
13_Sardine_egg_dam	658
14_Fish_larvae	2
15_Multiple	366
17_Gasteropoda	319
17_Zooplankton	468

Tableau 2 : Inventaire du Megalearning

c. *Learning extrait d'Ecotaxa, Pelgas 2016*

Validé par Jean-Baptiste Romagnan, les vignettes JPEG sont malheureusement excessivement compressées, ce qui dégrade la qualité des images ; en plus, des coins noirs y ont été dessinés. Les effectifs par catégories est probablement insuffisant pour du CNN. Le gros souci de ce learning, c'est qu'il n'est pas accompagné du fichier PID, je ne peux pas le comparer au DNN et RF.

Taxons	Effectifs
Actinopterygii_egg	923
artefact_bubble	985
Bacillariophyceae_Diatoma	935
Brachyura_megalopa	510
Calanoida_Acartiidae	942
Calanoida_Calanidae	883
Calanoida_Centropagidae	775
Calanoida_Euchaetidae	53
Calanoida_Pontellidae	264
Calanoida_Temoridae	901
Centropagidae_Centropages	889
Centropagidae_Isias	457
Cirripedia_nauplii	535
Copepoda_dead	864
Copepoda_Harpacticoida	45
Copepoda_multiple	819
Copepoda_Poecilostomatoida	871
Crustacea_larvae	812
Crustacea_nauplii	407
Cyclopoida_Oithonidae	972
Decapoda_zoea	876
detritus_fiber	653
Diplostraca_Cladocera	808
Engraulidae temp_egg 1 temp	746
Engraulidae temp_egg 2 3 temp	690
Engraulidae temp_egg 4 6 temp	656
Engraulidae temp_egg 7 8 temp	536
Engraulidae temp_egg 9 11 temp	389
Engraulidae temp_egg unkn temp	11
Eumalacostraca_Amphipoda	184
Eumalacostraca_Decapoda	94
Eumalacostraca_Euphausiacea	732
Gnathostomata_Actinopterygii	84
Harosa_Rhizaria	593
Limacinidae_Limacina	488
Maxillopoda_Copepoda	959

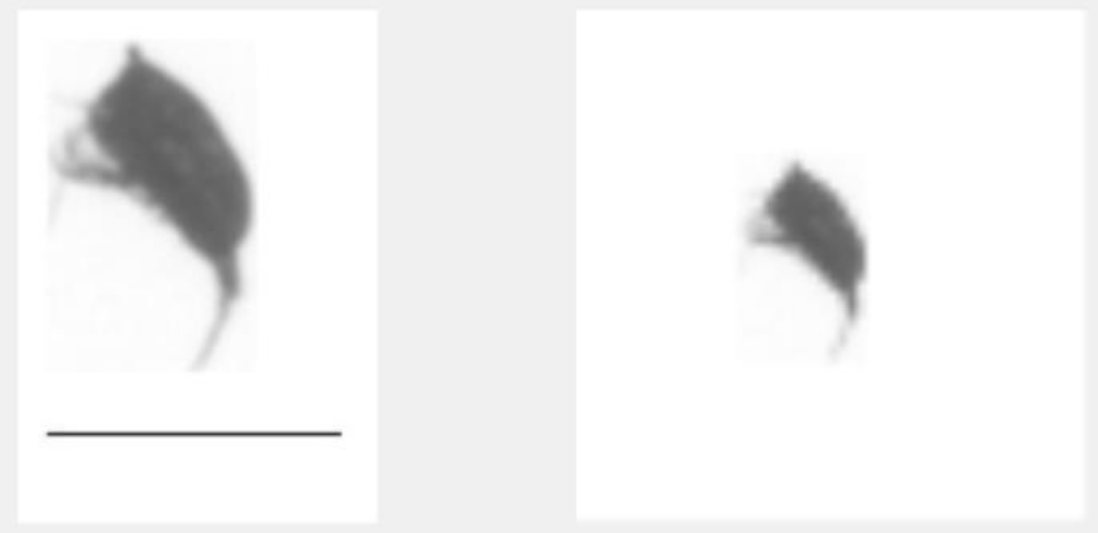
Metazoa_Chaetognatha	349
Metazoa_Echinodermata	81
Mollusca_egg	156
Noctilucaceae_Noctiluca	387
not-living_artefact	461
not-living_detritus	884
other_egg	463
other_gelatinous	441
other_multiple	797
Sardina temp_egg 1 temp	221
Sardina temp_egg 2 3 temp	807
Sardina temp_egg 4 6 temp	623
Sardina temp_egg 7 8 temp	319
Sardina temp_egg 9 11 temp	152
Sardina temp_egg unkn temp	943
Tunicata_Appendicularia	80

Tableau 3 : Inventaire du learning Ecotaxa Pelgas 2016

2.Utilisation du MegaLearning

1. CNN en apprentissage

Le principal paramètre du réseau de neurones convolutif est la taille de matrice d'entrée, correspondant idéalement à la dimension en pixels de la vignette (Largeur, hauteur). En pratique ici, toutes les vignettes sont de dimensions différentes, il faut donc les redimensionner en une dimension constante. Rarement, certaines vignette ont une taille au-delà de 700 pixels de coté (parfois, plusieurs organismes sont superposés sur la vignette) . Pour conserver la taille réelle des organismes dans une vignette de taille fixe, je suppose une taille maximum des vignettes : 300*300 (de manière statistique, cela doit correspondre à plus de 95% des individus), puis j'effectue une réduction correction d'échelle. Si cette correction n'était pas faite, alors la matrice d'entrée d'une grosse bulle par exemple serait sensiblement la même qu'une petite bulle ou d'un Halosphera, ce qui complique le travail du classifieur. Si on augmente la taille de la matrice d'entrée, on gagnera en prédiction à condition que le nombre d'individus par catégorie soit très important (sans compter que cela requerra plus de mémoire GPU). A l'inverse, trop réduire la matrice d'entrée supprimera trop d'informations sur les appendices des organismes. J'ai effectué des tests à 28*28, 56*56, 96*96, 120*120, 150*150.



*A gauche la vignette originale, à droite la vignette en 56*56*
*telle qu'elle serait aperçue dans une vignette 300*300*

Les paramètres définis ici sont :

- Matrice d'entrée : 56*56
- Learning rate : 0.01
- L2Decay : 0.001
- Batchsize : 500
- Momentum : 0.05
- Architecture :2 couches
  - Convol coté : 8
  - Convol Volume : 16
  - Convol stride : 1
  - Convol pad : 2
  - Activation :Relu
  - Pool coté : 2
  - Pool Stride :2

Durant un calcul d'apprentissage CNN ainsi paramétré, le CPU n'est pas très sollicité, ni le GPU (15%). La consommation de la mémoire PC est aussi faible. Il est possible que le goulot d'étranglement pourrait se situer sur la bande passante mémoire CPU. Les 8 Go de la mémoire GPU par contre ne sont pas forcement de trop : il est facile de la remplir et faire planter le programme si on demande un nombre d'individus par lots (batchsize) plus important et/ou plus de volume dans les couches, ou l'agrandissement de la matrice d'entrée. La lecture du taux d'utilisation de la mémoire GPU peut s'effectuer avec le gestionnaire de tâches de Windows 10.

A la fin du calcul, un fichier « *learning_cnn.txt* » est généré, il contient une entête contenant les valeurs minimums et maximums des catégories du fichier « *learning.pid* », puis le contenu du fichier « *learning.pid* » augmenté de plusieurs colonnes : elles correspondent à la dernière couche d'activation du CNN et le code de la catégorie prédite par CNN. Seuls les individus correctement prédits figurent dans ce fichier. La couche d'activation est un vecteur qui contient pour chaque catégorie, une probabilité de bonne prédiction. La somme de la couche vaut 1. L'indice de la colonne ayant la valeur la plus élevée (la plus proche de 1) donne l'indice de la catégorie prédite. Normalement l'ajout de ces colonnes devrait permettre de meilleures prédictions avec le RF et le DNN. C'est ce que cette étude va essayer de montrer, ou pas !
Le fichier de sortie (matrice d'apprentissage du réseau CNN) est sauvegardé dans le fichier « *_cnn_json.txt* », qui est utilisé pour les futures prédictions. Un répertoire « *_bin_cnn*» est aussi créé, il contient des fichiers d'optimisations utiles pour refaire un apprentissage avec de nouvelles valeurs des paramètres « learning rate », « L2decay » « Momentum » et une autre architecture sans refaire l'étape du redimensionnement.
Important : Il faut détruire ces fichiers si les paramètres de tris des catégories/individus changent sinon, il y aura un disfonctionnement du logiciel.

Voici ce que le logiciel calcule en 29400 secondes (8 heures) : Le nuage de points noirs indique la progression en pourcentage des bonnes prédictions des vignettes de la partie réservée à l'apprentissage, et le nuage rouge indique les bonnes prédictions des vignettes réservées au test.

Comme on peut l'observer, la prédiction des individus sélectionnés pour l'apprentissage tend vers 90-95% alors que les individus tests plafonnent à 60% : CNN arrive bien à apprendre avec les vignettes qui composent son learning, mais un phénomène bloquant empêche la bonne progression des prédictions des autres vignettes.

On remarque aussi 2 décrochages numériques aux alentours des itérations 30000 et 80000 (erreur mémoire ?)
J'ai passé du temps à tester l'influence des paramètres « Learning Rate », «L2Decay», «Momentum». Leur influence se situe surtout à la vitesse de convergence des courbes, mais le résultat au bout d'un certain temps me semble identique.

Voici sous forme de tableau la sortie du fichier *« c:\temp\ log_frm_control.txt »* :

Taxon	Nb individus biens prédits	Nb Individus Total	% Bonnes prédictions
00_Artefact	88	100	88.00
00_Artefact_2	0	16	0.00
00_Bubble	94	100	94.00
00_Bubbles_large	99	100	99.00
00_Detritus	69	100	69.00
00_Fiber	76	100	76.00
00_Halosphera	73	100	73.00
00_Paint_rust	25	51	49.02
00_Phytoplankton	5	48	10.42
00_Spirale	0	3	0.00
03_Copepoda_large	67	100	67.00
03_Copepoda_Pontellidae	1	24	4.17
03_Copepoda_small	63	100	63.00
04_Malacostraca_large	44	100	44.00
05_Cladocera	88	100	88.00
06_Cnidaria	0	4	0.00
13_Engraulis_egg_1	88	100	88.00
13_Engraulis_egg_2_3	67	100	67.00
13_Engraulis_egg_4_6	14	100	14.00
13_Engraulis_egg_7_8	98	100	98.00
13_Engraulis_egg_9_11	1	55	1.82
13_Fish_egg	92	100	92.00
13_Sardine_egg_1	1	9	11.11
13_Sardine_egg_2_3	81	100	81.00
13_Sardine_egg_4_6	57	100	57.00
13_Sardine_egg_7_8	5	31	16.13
13_Sardine_egg_9_11	4	27	14.81
13_Sardine_egg_dam	42	99	42.42
14_Fish_larvae	0	0	0.00
15_Multiple	10	55	18.18
17_Gasteropoda	4	48	8.33
17_Zooplankton	3	70	4.29

Groupe 00 : 73.68 %

Groupe 03 : 58.48 %
Groupe 04 : 44.00 %
Groupe 05 : 88.00 %
Groupe 06 : 0.00 %
Groupe 13 : 59.72 %
Groupe 14 : NaN %
Groupe 15 : 18.18 %
Groupe 17 : 5.93 %
Total : 60.67 %

L'extrait de la matrice de confusion suivante montre que les œufs d'anchois sont assez bien classés, par contre la différenciation des stades d'évolution des œufs n'est pas satisfaisante. Il y a au plus 100 individus test (/100) . La somme des valeurs d'une ligne vaut le nombre d'individus test. Il y a 10 individus anchois stade 1 qui ont été prédit par erreur en stade 2-3

13 Engraulis egg 1	13 Engraulis egg 2 3	13 Engraulis egg 4 6	13 Engraulis egg 7 8	13 Engraulis egg 9 11	13 Fish egg	13 Sardine egg 1	13 Sardine egg 2 3	13 Sardine egg 4 6	13 Sardine egg 7 8	13 Sardine egg 9 11	13 Sardine egg dam
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
2	2	0	2	0	3	0	0	0	0	0	0
0	1	0	1	0	0	0	0	0	0	0	0
0	1	0	0	0	1	0	1	1	1	0	7
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	2	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	2	0	0	0
88 / 100	10	0	0	0	0	0	0	0	0	0	0
9	67 / 100	4	20	0	0	0	0	0	0	0	0
0	20	14 / 100	64	0	0	0	0	0	0	0	0
0	1	1	98 / 100	0	0	0	0	0	0	0	0
0	2	1	51	1 / 55	0	0	0	0	0	0	0
0	0	0	1	0	92 / 100	0	1	3	0	0	2
0	0	0	0	0	0	1 / 9	7	0	0	0	1
0	0	0	0	0	4	1	81 / 100	11	2	0	1
0	0	0	1	0	1	0	29	57 / 100	6	2	4
1	1	0	1	0	1	0	7	11	5 / 31	2	1
0	0	0	1	0	1	0	1	15	2	4 / 27	1
1	0	2	9	0	3	1	20	10	2	1	42 / 99
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	2	1	0	0	5	0	0	1	1
0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	1	0	1	0	0	1	0	0	2

De plus, le manque de vignettes d'œufs de sardines fait écrouler le score du groupe 13 (ensemble des œufs)

2. Random Forest

Sur la Thalassa, le RF de Zoocam est défini à 650 arbres de décision. Mais ici le MegaLearning comprend beaucoup plus d'individus (plus de 46000, au lieu de quelques milliers), ce qui consomme beaucoup de mémoire CPU. J'ai donc réduit à 120 arbres, la consommation mémoire est ici déjà supérieure à 10 Go. Par contre, le calcul n'est pas très long : une vingtaine de secondes. N'oublions pas que relancer le même calcul donnera un résultat différent à 2 ou 3% près.

d. Utilisation du fichier learning.pid

6934 individus test, 46215 individus Learning sélectionnés 32 catégories

Taxon	Nb individus biens prédits	Nb Individus Total	% Bonnes prédictions
00_Artefact	271	306	88.56
00_Artefact_2	2	16	12.5
00_Bubble	428	450	95.11
00_Bubbles_large	265	267	99.25
00_Detritus	342	423	80.85
00_Fiber	416	449	92.65
00_Halosphera	93	132	70.45
00_Paint_rust	31	51	60.78
00_Phytoplankton	15	48	31.25
00_Spirale	0	3	0
03_Copepoda_large	275	448	61.38
03_Copepoda_Pontellidae	14	24	58.33
03_Copepoda_small	350	450	77.78
04_Malacostraca_large	413	450	91.78
05_Cladocera	361	415	86.99
06_Cnidaria	0	4	0
13_Engraulis_egg_1	410	450	91.11
13_Engraulis_egg_2_3	237	449	52.78
13_Engraulis_egg_4_6	203	444	45.72
13_Engraulis_egg_7_8	368	449	81.96
13_Engraulis_egg_9_11	1	55	1.82
13_Fish_egg	331	359	92.2
13_Sardine_egg_1	1	9	11.11
13_Sardine_egg_2_3	260	287	90.59
13_Sardine_egg_4_6	93	168	55.36
13_Sardine_egg_7_8	0	31	0
13_Sardine_egg_9_11	13	27	48.15
13_Sardine_egg_dam	50	99	50.51
14_Fish_larvae	0	0	0
15_Multiple	1	54	1.85
17_Gasteropoda	3	48	6.25
17_Zooplankton	19	69	27.54

Bonnes prédictions = 75.94%
Groupe 00 : 86.85 %
Groupe 03 : 69.31 %

Groupe 04 : 91.78 %
Groupe 05 : 86.99 %
Groupe 06 : 0.00 %
Groupe 13 : 69.58 %
Groupe 14 : NaN %
Groupe 15 : 1.85 %
Groupe 17 : 18.80 %
Total : 75.94 %

La matrice de confusion suivante montre ici aussi que si les taxons sont assez bien classés, la différentiation dans les stades est perfectible. Le nombre d'individu test n'est pas limité à 100 comme pour le CNN. 63 individus anchois 7-8 ont été classés par erreur en stade 4-6, parmi 449 . 368 sont biens classés.

13 Engraulis egg 1	13 Engraulis egg 2 3	13 Engraulis egg 4 6	13 Engraulis egg 7 8	13 Engraulis egg 9 11	13 Fish egg	13 Sardine egg 1	13 Sardine egg 2 3	13 Sardine egg 4 6	13 Sardine egg 7 8	13 Sardine egg 9 11	13 Sardine egg dam
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	1	0	0	0	0	0	2
0	0	0	0	0	0	0	0	0	0	0	0
1	6	0	2	0	19	0	1	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	8	0	1	0	0	0	6
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	1
410 / 450	27	3	1	0	2	0	0	0	0	0	0
42	237 / 449	110	40	0	7	0	0	0	0	0	0
1	64	203 / 444	157	0	12	0	1	0	0	0	0
0	8	63	368 / 449	1	3	0	0	2	0	2	1
0	0	5	45	1 / 55	2	0	0	1	0	0	0
1	6	0	4	0	331 / 359	0	4	2	0	0	1
0	0	0	0	0	1	1 / 9	5	1	0	0	1
0	0	0	0	0	4	0	260 / 287	19	0	0	4
0	0	0	2	0	6	0	54	93 / 168	2	3	6
0	0	0	2	0	0	1	10	11	0 / 31	2	5
0	0	0	0	0	0	0	2	11	0	13 / 27	1
1	0	0	1	0	9	0	20	6	0	0	50 / 99
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	2	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	4

e. *Utilisation du fichier learning_cnn.txt*

Sortie du fichier « c:\temp\ log_frm_control.txt »

Random Forest
6934 individus test, 46215 individus Learning sélectionnés 32 catégories
00_Artefact [285/306/93.14%]
00_Artefact_2 [0/16/0.00%]
00_Bubble [434/450/96.44%]
00_Bubbles_large [265/267/99.25%]
00_Detritus [372/423/87.94%]
00_Fiber [419/449/93.32%]
00_Halosphera [119/132/90.15%]
00_Paint_rust [45/51/88.24%]
00_Phytoplankton [33/48/68.75%]

00_Spirale [0/3/0.00%]
03_Copepoda_large [412/448/91.96%]
03_Copepoda_Pontellidae [24/24/100.00%]
03_Copepoda_small [406/450/90.22%]
04_Malacostraca_large [442/450/98.22%]
05_Cladocera [368/415/88.67%]
06_Cnidaria [4/4/100.00%]
13_Engraulis_egg_1 [427/450/94.89%]
13_Engraulis_egg_2_3 [371/449/82.63%]
13_Engraulis_egg_4_6 [311/444/70.05%]
13_Engraulis_egg_7_8 [390/449/86.86%]
13_Engraulis_egg_9_11 [42/55/76.36%]
13_Fish_egg [343/359/95.54%]
13_Sardine_egg_1 [6/9/66.67%]
13_Sardine_egg_2_3 [282/287/98.26%]
13_Sardine_egg_4_6 [160/168/95.24%]
13_Sardine_egg_7_8 [25/31/80.65%]
13_Sardine_egg_9_11 [19/27/70.37%]
13_Sardine_egg_dam [95/99/95.96%]
14_Fish_larvae [0/0/0.00%]
15_Multiple [49/54/90.74%]
17_Gasteropoda [24/48/50.00%]
17_Zooplankton [36/69/52.17%]
Bonnes prédictions = 89.53%
Groupe 00 : 91.93 %
Groupe 03 : 91.32 %
Groupe 04 : 98.22 %
Groupe 05 : 88.67 %
Groupe 06 : 100.00 %
Groupe 13 : 87.41 %
Groupe 14 : NaN %
Groupe 15 : 90.74 %
Groupe 17 : 51.28 %
Total : 89.53 %

13 Engraulis egg 1	13 Engraulis egg 2 3	13 Engraulis egg 4 6	13 Engraulis egg 7 8	13 Engraulis egg 9 11	13 Fish egg	13 Sardine egg 1	13 Sardine egg 2 3	13 Sardine egg 4 6	13 Sardine egg 7 8	13 Sardine egg 9 11	13 Sardine egg dam
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0
1	2	1	0	0	2	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	5	0	0	0	0	0	2
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
427 / 450	20	1	0	0	0	0	0	0	0	0	0
14	371 / 449	51	10	1	1	0	0	0	0	0	0
0	44	311 / 444	85	1	0	0	1	0	0	0	0
0	5	48	390 / 449	2	1	0	0	2	0	0	0
0	0	0	13	42 / 55	0	0	0	0	0	0	0
0	0	0	3	0	343 / 359	0	1	0	0	0	0
0	0	0	1	0	0	6 / 9	1	1	0	0	0
0	0	0	0	0	1	0	282 / 287	2	0	0	2
0	0	0	0	0	1	0	5	160 / 168	0	0	2
0	0	0	1	0	0	0	2	2	25 / 31	0	1
0	0	0	0	0	0	0	1	5	1	19 / 27	1
0	0	0	0	0	0	0	1	2	0	0	95 / 99
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	1	1	0	0	0

Comme on peut le voir, sur les œufs (groupe 13), la prédiction par le learning issu du CNN améliore significativement la prédiction 69.5-> 87.4 %

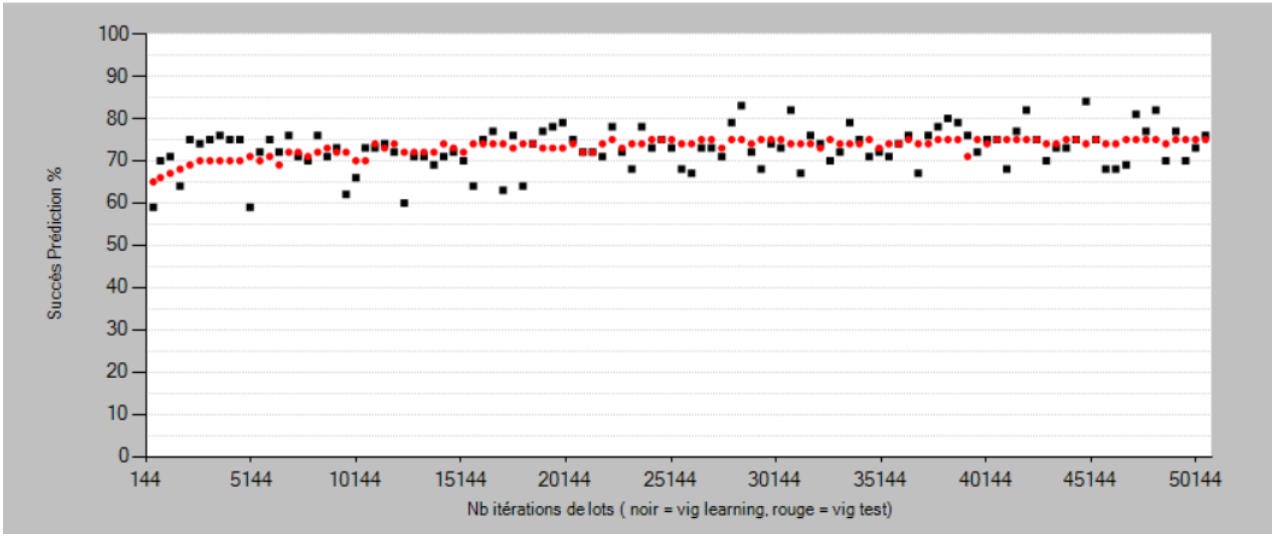
3. DNN

Les paramètres sont les suivants :

Learning rate : 0.012
L2Decay : 0.001
Batchsize : 500
Momentum : 0.1
3 couches de neurones : 26, 52,26
Temps de calcul : 15 minutes

Comme pour le CNN, les ressources CPU, GPU sont sous sollicitées. Le fichier d'apprentissage « *_dnn_json.txt* » est généré pour les futures prédictions.

f. Utilisation du fichier learning.pid



Le taux global de bonnes prédictions converge assez rapidement vers 70%

00_Artefact [272/306/88.89%]
00_Artefact_2 [1/16/6.25%]
00_Bubble [416/450/92.44%]
00_Bubbles_large [266/267/99.63%]
00_Detritus [353/423/83.45%]
00_Fiber [398/449/88.64%]
00_Halosphera [83/132/62.88%]
00_Paint_rust [37/51/72.55%]
00_Phytoplankton [13/48/27.08%]
00_Spirale [0/3/0.00%]
03_Copepoda_large [335/448/74.78%]
03_Copepoda_Pontellidae [12/24/50.00%]
03_Copepoda_small [362/450/80.44%]
04_Malacostraca_large [358/450/79.56%]
05_Cladocera [332/415/80.00%]
06_Cnidaria [0/4/0.00%]
13_Engraulis_egg_1 [413/450/91.78%]
13_Engraulis_egg_2_3 [258/449/57.46%]
13_Engraulis_egg_4_6 [245/444/55.18%]
13_Engraulis_egg_7_8 [340/449/75.72%]
13_Engraulis_egg_9_11 [2/55/3.64%]
13_Fish_egg [329/359/91.64%]
13_Sardine_egg_1 [0/9/0.00%]
13_Sardine_egg_2_3 [248/287/86.41%]
13_Sardine_egg_4_6 [111/168/66.07%]
13_Sardine_egg_7_8 [0/31/0.00%]
13_Sardine_egg_9_11 [4/27/14.81%]
13_Sardine_egg_dam [50/99/50.51%]
14_Fish_larvae [0/0/0.00%]

15_Multiple [4/54/7.41%]
17_Gasteropoda [3/48/6.25%]
17_Zooplankton [11/69/15.94%]
Groupe 00 : 85.73 %
Groupe 03 : 76.90 %
Groupe 04 : 79.56 %
Groupe 05 : 80.00 %
Groupe 06 : 0.00 %
Groupe 13 : 70.75 %
Groupe 14 : NaN %
Groupe 15 : 7.41 %
Groupe 17 : 11.97 %
Total : 75.80 %

13 Engraulis egg 1	13 Engraulis egg 2 3	13 Engraulis egg 4 6	13 Engraulis egg 7 8	13 Engraulis egg 9 11	13 Fish egg	13 Sardine egg 1	13 Sardine egg 2 3	13 Sardine egg 4 6	13 Sardine egg 7 8	13 Sardine egg 9 11	13 Sardine egg dam
1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	7	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	1	0	0	0	0	0	2
0	0	0	0	0	0	0	0	0	0	0	0
6	2	0	2	0	26	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	4	0	0	0	0	0	7
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
413 / 450	23	0	0	0	1	0	0	0	0	0	0
51	258 / 449	116	6	0	3	0	0	0	0	0	0
1	99	245 / 444	86	0	3	0	1	0	0	0	0
0	6	86	340 / 449	1	9	0	1	0	0	0	0
0	3	6	43	2 / 55	0	0	0	1	0	0	0
2	1	1	2	0	329 / 359	0	5	2	0	0	0
0	0	0	0	0	0	0 / 9	2	0	0	0	7
0	0	0	0	0	8	0	248 / 287	21	0	0	9
0	0	1	0	0	7	0	37	111 / 168	0	1	11
0	1	0	0	0	4	0	7	18	0 / 31	0	1
0	0	0	0	0	1	0	2	19	0	4 / 27	1
0	0	1	0	0	5	0	9	17	0	0	50 / 99
0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	1	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	2	0	1	0	0	0	0	0	3

g. Utilisation du fichier learning_cnn.txt

DNN
00_Artefact [290/306/94.77%]
00_Artefact_2 [2/16/12.50%]
00_Bubble [428/450/95.11%]
00_Bubbles_large [265/267/99.25%]
00_Detritus [357/423/84.40%]
00_Fiber [410/449/91.31%]
00_Halosphera [119/132/90.15%]
00_Paint_rust [47/51/92.16%]
00_Phytoplankton [30/48/62.50%]
00_Spirale [0/3/0.00%]

03_Copepoda_large [411/448/91.74%]
03_Copepoda_Pontellidae [24/24/100.00%]
03_Copepoda_small [403/450/89.56%]
04_Malacostraca_large [430/450/95.56%]
05_Cladocera [363/415/87.47%]
06_Cnidaria [0/4/0.00%]
13_Engraulis_egg_1 [421/450/93.56%]
13_Engraulis_egg_2_3 [371/449/82.63%]
13_Engraulis_egg_4_6 [323/444/72.75%]
13_Engraulis_egg_7_8 [382/449/85.08%]
13_Engraulis_egg_9_11 [39/55/70.91%]
13_Fish_egg [344/359/95.82%]
13_Sardine_egg_1 [7/9/77.78%]
13_Sardine_egg_2_3 [279/287/97.21%]
13_Sardine_egg_4_6 [161/168/95.83%]
13_Sardine_egg_7_8 [25/31/80.65%]
13_Sardine_egg_9_11 [22/27/81.48%]
13_Sardine_egg_dam [90/99/90.91%]
14_Fish_larvae [0/0/0.00%]
15_Multiple [44/54/81.48%]
17_Gasteropoda [24/48/50.00%]
17_Zooplankton [38/69/55.07%]
Groupe 00 : 90.82 %
Groupe 03 : 90.89 %
Groupe 04 : 95.56 %
Groupe 05 : 87.47 %
Groupe 06 : 0.00 %
Groupe 13 : 87.16 %
Groupe 14 : NaN %
Groupe 15 : 81.48 %
Groupe 17 : 52.99 %
Total : 88.68 %

13 Engraulis egg 1	13 Engraulis egg 2 3	13 Engraulis egg 4 6	13 Engraulis egg 7 8	13 Engraulis egg 9 11	13 Fish egg	13 Sardine egg 1	13 Sardine egg 2 3	13 Sardine egg 4 6	13 Sardine egg 7 8	13 Sardine egg 9 11	13 Sardine egg dam
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	2	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
4	1	2	1	0	2	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	6	0	0	0	0	0	3
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
421 / 450	21	0	1	0	0	0	0	0	0	0	0
22	371 / 449	45	3	0	3	0	0	0	0	0	0
0	47	323 / 444	70	0	2	0	0	0	0	0	0
0	4	56	382 / 449	4	1	0	0	1	0	0	0
0	1	2	11	39 / 55	2	0	0	0	0	0	0
0	0	1	4	0	344 / 359	0	0	0	0	0	0
0	0	0	0	0	0	7 / 9	2	0	0	0	0
0	1	0	0	0	2	0	279 / 287	5	0	0	0
0	0	0	0	0	0	0	4	161 / 168	0	1	2
0	0	0	1	0	2	0	1	1	25 / 31	0	0
0	0	0	2	0	0	0	0	3	0	22 / 27	0
0	0	0	0	0	2	0	2	2	0	0	90 / 99
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	0	0	0	0	0	0	3
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0

Les prédictions sont bonnes et comparables avec le RF utilisant le fichier *learning_cnn.txt*

3. Prédictions de l'échantillon 308 avec Megalearning

Les résultats précédant sont encourageants, mais on va prédire l'échantillon CUFES 308 et comparer les résultats avec ce qui a été validé à bord du navire. Les catégories *_Autres*, *Cut* et *Duplicats* sont exclues de tous les tests. Je n'ai pas touché à l'échantillon validé. Si nous l'examinons, nous remarquerons un mélange dans les détritus et bulles, il y a des petites bulles dans Cladocera…

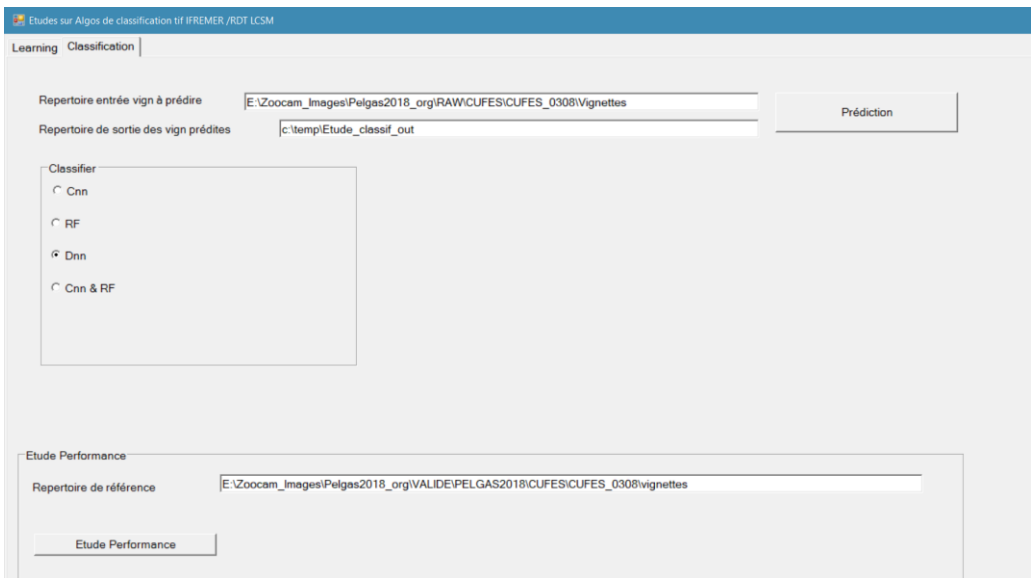


Fig 2 : Page classifications du logiciel

Nous ne pouvons pas utiliser ici le fichier *'learning_cnn.txt'* puisque le nombre de colonnes features (morphos50 + catégo23=83) n'est pas le même que ceux du fichier PELGAS2018_CUFES_308.pid (=50).

1. Avec CNN

00_Artefact[0/2/.0%]
00_Artefact_2[0/0/NaN%]
00_Bubble[32/44/72.7%]
00_Bubbles_large[5/8/62.5%]
00_Detritus[316/331/95.5%]
00_Fiber[50/115/43.5%]
00_Halosphera[9/11/81.8%]
00_Paint_rust[0/2/.0%]
00_Phytoplankton[0/4/.0%]
00_Spirale[0/0/NaN%]
03_Copepoda_large[235/284/82.7%]
03_Copepoda_Pontellidae[0/7/.0%]
03_Copepoda_small[347/415/83.6%]
04_Malacostraca_large[24/103/23.3%]
05_Cladocera[14/32/43.8%]
06_Cnidaria[0/0/NaN%]
13_Engraulis_egg_1[61/61/100.0%]
13_Engraulis_egg_2_3[24/41/58.5%]
13_Engraulis_egg_4_6[6/10/60.0%]
13_Engraulis_egg_7_8[2/5/40.0%]
13_Engraulis_egg_9_11[0/1/.0%]
13_Fish_egg[14/28/50.0%]
13_Sardine_egg_1[0/0/NaN%]
13_Sardine_egg_2_3[30/39/76.9%]
13_Sardine_egg_4_6[7/9/77.8%]
13_Sardine_egg_7_8[0/2/.0%]
13_Sardine_egg_9_11[0/0/NaN%]
13_Sardine_egg_dam[4/21/19.0%]
14_Fish_larvae[0/0/NaN%]
15_Multiple[0/11/.0%]
17_Gasteropoda[0/0/NaN%]
17_Zooplankton[0/16/.0%]
_Autres[0/0/NaN%]
Groupe 00 : 79.69 %
Groupe 03 : 82.44 %
Groupe 04 : 23.30 %
Groupe 05 : 43.75 %

Groupe 06 : NaN %
Groupe 13 : 68.20 %
Groupe 14 : NaN %
Groupe 15 : 0.00 %
Groupe 17 : 0.00 %
Total : 73.66 %

Le nombre important de vignettes œufs d'anchois stade 1 et sa relative simplicité permet au CNN d'avoir un score parfait.

2. Avec Random Forest

00_Artefact[0/0/NaN%]
00_Artefact_2[0/0/NaN%]
00_Bubble[31/43/72.1%]
00_Bubbles_large[6/6/100.0%]
00_Detritus[327/328/99.7%]
00_Fiber[64/126/50.8%]
00_Halosphera[10/11/90.9%]
00_Paint_rust[0/0/NaN%]
00_Phytoplankton[0/6/.0%]
00_Spirale[0/0/NaN%]
03_Copepoda_large[240/264/90.9%]
03_Copepoda_Pontellidae[0/2/.0%]
03_Copepoda_small[348/369/94.3%]
04_Malacostraca_large[39/175/22.3%]
05_Cladocera[18/51/35.3%]
06_Cnidaria[0/0/NaN%]
13_Engraulis_egg_1[64/65/98.5%]
13_Engraulis_egg_2_3[21/31/67.7%]
13_Engraulis_egg_4_6[6/10/60.0%]
13_Engraulis_egg_7_8[4/7/57.1%]
13_Engraulis_egg_9_11[0/0/NaN%]
13_Fish_egg[14/19/73.7%]
13_Sardine_egg_1[0/0/NaN%]
13_Sardine_egg_2_3[31/41/75.6%]
13_Sardine_egg_4_6[9/13/69.2%]
13_Sardine_egg_7_8[0/0/NaN%]
13_Sardine_egg_9_11[0/0/NaN%]
13_Sardine_egg_dam[7/18/38.9%]
14_Fish_larvae[0/0/NaN%]
15_Multiple[0/1/.0%]
17_Gasteropoda[0/1/.0%]

17_Zooplankton[0/15/.0%]
_Autres[0/0/NaN%]
Groupe 00 : 84.23 %
Groupe 03 : 92.60 %
Groupe 04 : 22.29 %
Groupe 05 : 35.29 %
Groupe 06 : NaN %
Groupe 13 : 76.47 %
Groupe 14 : NaN %
Groupe 15 : 0.00 %
Groupe 17 : 0.00 %
Total : 77.34 %

CNN est moins bon que RF pour l'ensemble des stades des œufs d'anchois

3. Avec DNN

00_Artefact[0/6/.0%]
00_Artefact_2[0/0/NaN%]
00_Bubble[28/33/84.8%]
00_Bubbles_large[6/7/85.7%]
00_Detritus[341/352/96.9%]
00_Fiber[62/124/50.0%]
00_Halosphera[3/3/100.0%]
00_Paint_rust[0/0/NaN%]
00_Phytoplankton[0/4/.0%]
00_Spirale[0/0/NaN%]
03_Copepoda_large[272/298/91.3%]
03_Copepoda_Pontellidae[0/1/.0%]
03_Copepoda_small[364/392/92.9%]
04_Malacostraca_large[31/116/26.7%]
05_Cladocera[24/47/51.1%]
06_Cnidaria[0/0/NaN%]
13_Engraulis_egg_1[67/74/90.5%]
13_Engraulis_egg_2_3[24/35/68.6%]
13_Engraulis_egg_4_6[6/8/75.0%]
13_Engraulis_egg_7_8[4/10/40.0%]
13_Engraulis_egg_9_11[0/0/NaN%]
13_Fish_egg[7/7/100.0%]
13_Sardine_egg_1[0/0/NaN%]
13_Sardine_egg_2_3[32/42/76.2%]
13_Sardine_egg_4_6[7/7/100.0%]

13_Sardine_egg_7_8[0/0/NaN%]
13_Sardine_egg_9_11[0/0/NaN%]
13_Sardine_egg_dam[7/20/35.0%]
14_Fish_larvae[0/0/NaN%]
15_Multiple[0/4/.0%]
17_Gasteropoda[0/0/NaN%]
17_Zooplankton[0/12/.0%]
_Autres[0/0/NaN%]
Groupe 00 : 83.18 %
Groupe 03 : 92.04 %
Groupe 04 : 26.72 %
Groupe 05 : 51.06 %
Groupe 06 : NaN %
Groupe 13 : 75.86 %
Groupe 14 : NaN %
Groupe 15 : 0.00 %
Groupe 17 : 0.00 %
Total : 80.21 %

DNN est un peu meilleur que RF et bien meilleur CNN

4. Avec RF aidé par CNN (Variante A)

00_Artefact[0/0/NaN%]
00_Artefact_2[0/0/NaN%]
00_Bubble[32/45/71.1%]
00_Bubbles_large[6/6/100.0%]
00_Detritus[326/331/98.5%]
00_Fiber[60/126/47.6%]
00_Halosphera[9/11/81.8%]
00_Paint_rust[0/0/NaN%]
00_Phytoplankton[0/2/.0%]
00_Spirale[0/0/NaN%]
03_Copepoda_large[236/253/93.3%]
03_Copepoda_Pontellidae[0/0/NaN%]
03_Copepoda_small[368/401/91.8%]
04_Malacostraca_large[39/169/23.1%]
05_Cladocera[17/38/44.7%]
06_Cnidaria[0/0/NaN%]
13_Engraulis_egg_1[62/62/100.0%]
13_Engraulis_egg_2_3[23/33/69.7%]
13_Engraulis_egg_4_6[7/10/70.0%]
13_Engraulis_egg_7_8[4/7/57.1%]

13_Engraulis_egg_9_11[0/0/NaN%]
13_Fish_egg[13/19/68.4%]
13_Sardine_egg_1[0/0/NaN%]
13_Sardine_egg_2_3[33/44/75.0%]
13_Sardine_egg_4_6[9/9/100.0%]
13_Sardine_egg_7_8[0/0/NaN%]
13_Sardine_egg_9_11[0/0/NaN%]
13_Sardine_egg_dam[7/20/35.0%]
14_Fish_larvae[0/0/NaN%]
15_Multiple[0/2/.0%]
17_Gasteropoda[0/0/NaN%]
17_Zooplankton[0/14/.0%]
_Autres[0/0/NaN%]
Groupe 00 : 83.11 %
Groupe 03 : 92.35 %
Groupe 04 : 23.08 %
Groupe 05 : 44.74 %
Groupe 06 : NaN %
Groupe 13 : 77.45 %
Groupe 14 : NaN %
Groupe 15 : 0.00 %
Groupe 17 : 0.00 %
Total : 78.09 %

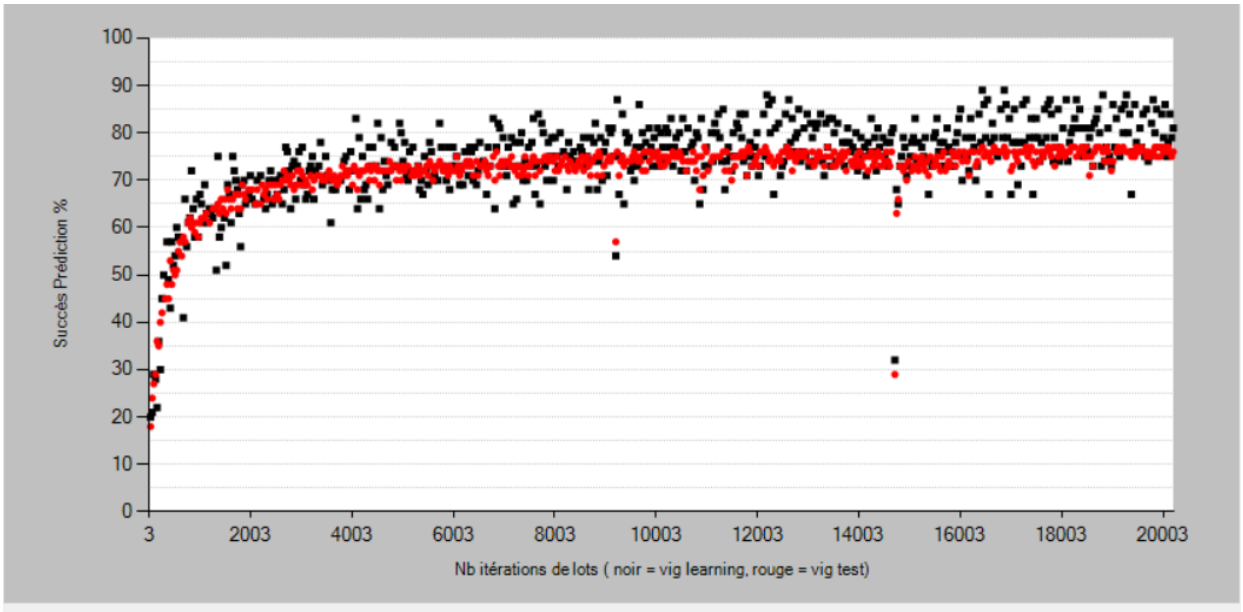
L'augmentation du taux de bonnes prédictions n'est pas significative par rapport au RF seul.

4. Utilisation du MegaLearning réduit

Je crée une sélection de catégories composées de plus de 2000 individus à partir de Megalearning.

1. En CNN

On obtient :



Les prédictions sont globalement meilleures qu'avec toutes les catégories du MegaLearning.

00_Artefact [89/100/89.00%]
00_Bubble [98/100/98.00%]
00_Detritus [59/100/59.00%]
00_Fiber [81/100/81.00%]
03_Copepoda_large [75/100/75.00%]
03_Copepoda_small [83/100/83.00%]
04_Malacostraca_large [41/100/41.00%]
05_Cladocera [79/100/79.00%]
13_Engraulis_egg_1 [90/100/90.00%]
13_Engraulis_egg_2_3 [72/100/72.00%]
13_Engraulis_egg_4_6 [51/100/51.00%]
13_Engraulis_egg_7_8 [87/100/87.00%]
13_Fish_egg [91/100/91.00%]
Groupe 00 : 81.75 %
Groupe 03 : 79.00 %
Groupe 04 : 41.00 %
Groupe 05 : 79.00 %
Groupe 13 : 78.20 %
Total : 76.62 %
Temps de calcul : 1h30
En laissant plusieurs heures, le taux de bonnes prédictions global dépasse légèrement 81%

2. En RF

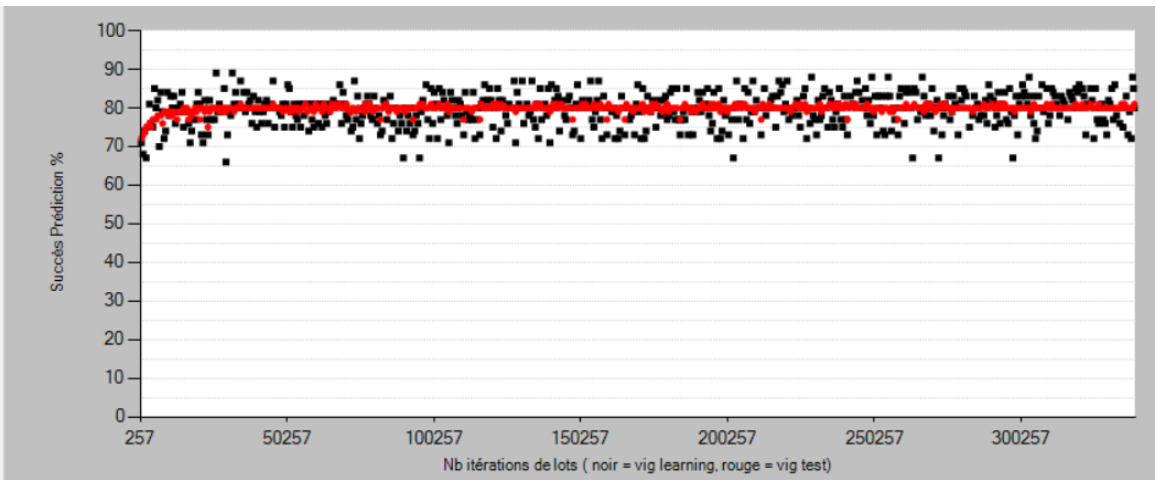
Random Forest
5542 individus test, 36940 individus Learning sélectionnés 13 catégories
00_Artefact [279/306/91.18%]

00_Bubble [431/450/95.78%]
00_Detritus [319/423/75.41%]
00_Fiber [414/449/92.20%]
03_Copepoda_large [276/448/61.61%]
03_Copepoda_small [341/450/75.78%]
04_Malacostraca_large [410/450/91.11%]
05_Cladocera [377/415/90.84%]
13_Engraulis_egg_1 [417/450/92.67%]
13_Engraulis_egg_2_3 [230/449/51.22%]
13_Engraulis_egg_4_6 [227/444/51.13%]
13_Engraulis_egg_7_8 [349/449/77.73%]
13_Fish_egg [328/359/91.36%]
Bonnes prédictions = 79.36%
Groupe 00 : 88.64 %
Groupe 03 : 68.71 %
Groupe 04 : 91.11 %
Groupe 05 : 90.84 %
Groupe 13 : 72.11 %
Total : 79.36 %

RF est moins performant aue le CNN sur le groupe 13.

3. En DNN

On obtient :



DNN
00_Artefact [280/306/91.50%]
00_Bubble [434/450/96.44%]
00_Detritus [362/423/85.58%]
00_Fiber [391/449/87.08%]
03_Copepoda_large [326/448/72.77%]
03_Copepoda_small [370/450/82.22%]
04_Malacostraca_large [384/450/85.33%]
05_Cladocera [318/415/76.63%]

13_Engraulis_egg_1 [419/450/93.11%]
13_Engraulis_egg_2_3 [269/449/59.91%]
13_Engraulis_egg_4_6 [255/444/57.43%]
13_Engraulis_egg_7_8 [335/449/74.61%]
13_Fish_egg [347/359/96.66%]
Groupe 00 : 90.11 %
Groupe 03 : 77.51 %
Groupe 04 : 85.33 %
Groupe 05 : 76.63 %
Groupe 13 : 75.55 %
Total : 81.02 %

DNN est aussi performant avec toutes les catégories du learning, mais CNN est le plus performant.

5. Prédictions de l'échantillon 308 avec Megalearning réduit

1. Avec le CNN

00_Artefact[0/1/.0%]
00_Bubble[31/44/70.5%]
00_Detritus[323/339/95.3%]
00_Fiber[53/139/38.1%]
03_Copepoda_large[275/331/83.1%]
03_Copepoda_small[328/371/88.4%]
04_Malacostraca_large[30/103/29.1%]
05_Cladocera[14/49/28.6%]
13_Engraulis_egg_1[68/72/94.4%]
13_Engraulis_egg_2_3[23/31/74.2%]
13_Engraulis_egg_4_6[6/7/85.7%]
13_Engraulis_egg_7_8[4/6/66.7%]
13_Fish_egg[17/33/51.5%]
Groupe 00 : 77.82 %
Groupe 03 : 85.90 %
Groupe 04 : 29.13 %
Groupe 05 : 28.57 %
Groupe 13 : 79.19 %
Total : 76.80 %

On s'intéresse surtout ici aux œufs puisque ces catégories sont présentes dans le learning réduit et dans l'échantillon validé. La prédiction du groupe 13 (œufs) montre la bonne performance de CNN en learning réduit (68.20% avec toutes les catégories).

2. Avec le RF

00_Artefact[0/0/NaN%]
00_Bubble[32/44/72.7%]
00_Detritus[354/356/99.4%]
00_Fiber[64/126/50.8%]
03_Copepoda_large[234/259/90.3%]
03_Copepoda_small[347/369/94.0%]
04_Malacostraca_large[39/180/21.7%]
05_Cladocera[18/50/36.0%]
13_Engraulis_egg_1[65/66/98.5%]
13_Engraulis_egg_2_3[23/30/76.7%]
13_Engraulis_egg_4_6[9/11/81.8%]
13_Engraulis_egg_7_8[4/6/66.7%]
13_Fish_egg[16/29/55.2%]
Groupe 00 : 85.55 %
Groupe 03 : 92.52 %
Groupe 04 : 21.67 %
Groupe 05 : 36.00 %
Groupe 13 : 82.39 %
Total : 78.96 %

... Mais RF semble ici faire mieux...

3. Avec DNN

00_Artefact[1/5/20.0%]
00_Bubble[32/45/71.1%]
00_Detritus[371/376/98.7%]
00_Fiber[61/103/59.2%]
03_Copepoda_large[281/323/87.0%]
03_Copepoda_small[345/360/95.8%]
04_Malacostraca_large[35/133/26.3%]
05_Cladocera[17/38/44.7%]
13_Engraulis_egg_1[65/67/97.0%]
13_Engraulis_egg_2_3[26/35/74.3%]
13_Engraulis_egg_4_6[8/10/80.0%]
13_Engraulis_egg_7_8[4/4/100.0%]
13_Fish_egg[17/27/63.0%]
Groupe 00 : 87.90 %
Groupe 03 : 91.65 %
Groupe 04 : 26.32 %
Groupe 05 : 44.74 %
Groupe 13 : 83.92 %

Total : 82.77 %

DNN produit ici un résultat semblable à RF sur les oeufs.

4. Avec RF variante A

00_Artefact[0/0/NaN%]
00_Bubble[32/45/71.1%]
00_Detritus[347/351/98.9%]
00_Fiber[62/131/47.3%]
03_Copepoda_large[250/269/92.9%]
03_Copepoda_small[353/379/93.1%]
04_Malacostraca_large[40/160/25.0%]
05_Cladocera[17/53/32.1%]
13_Engraulis_egg_1[68/68/100.0%]
13_Engraulis_egg_2_3[24/31/77.4%]
13_Engraulis_egg_4_6[6/8/75.0%]
13_Engraulis_egg_7_8[4/6/66.7%]
13_Fish_egg[16/25/64.0%]
Groupe 00 : 83.68 %
Groupe 03 : 93.06 %
Groupe 04 : 25.00 %
Groupe 05 : 32.08 %
Groupe 13 : 85.51 %
Total : 79.88 %

C'est le meilleur score, mais le temps de calcul est bien allongé.

5. Avec RF variante B

00_Artefact[0/0/NaN%]
00_Bubble[32/45/71.1%]
00_Detritus[359/361/99.4%]
00_Fiber[64/124/51.6%]
03_Copepoda_large[236/264/89.4%]
03_Copepoda_small[342/364/94.0%]
04_Malacostraca_large[38/181/21.0%]
05_Cladocera[18/49/36.7%]
13_Engraulis_egg_1[63/64/98.4%]
13_Engraulis_egg_2_3[23/33/69.7%]
13_Engraulis_egg_4_6[6/8/75.0%]
13_Engraulis_egg_7_8[4/7/57.1%]
13_Fish_egg[16/26/61.5%]
Groupe 00 : 85.85 %
Groupe 03 : 92.04 %
Groupe 04 : 20.99 %
Groupe 05 : 36.73 %

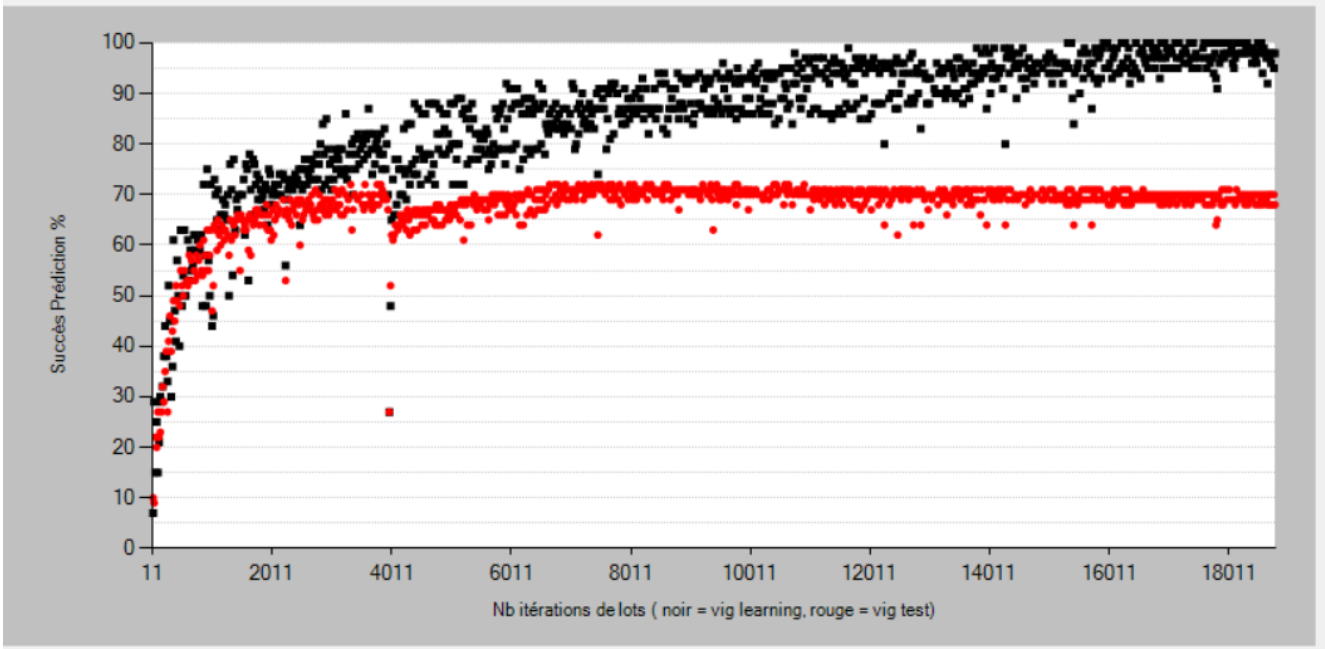
Groupe 13 : 81.16 %
Total : 78.70 %

C'est un bon résultat, mais pas le meilleur. En plus du temps de calcul de RF, il faut environ 8 secondes supplémentaires pour que CNN pour redimensionne les 2200 vignettes de l'échantillon 308.
Le cas est ici idéal : beaucoup de vignettes par catégories et peu de catégories. Les classifieurs sont à leur plus au niveau de performance.

6. Utilisation du Learning Pelgas 2018 Martin Huret

1. Avec le CNN

Les paramètres sont identiques avec ceux du Megalearning
Nombre total de vignettes : 6788, dont 5871 pour le learning
Temps de calcul : 8780 secondes (2h25)



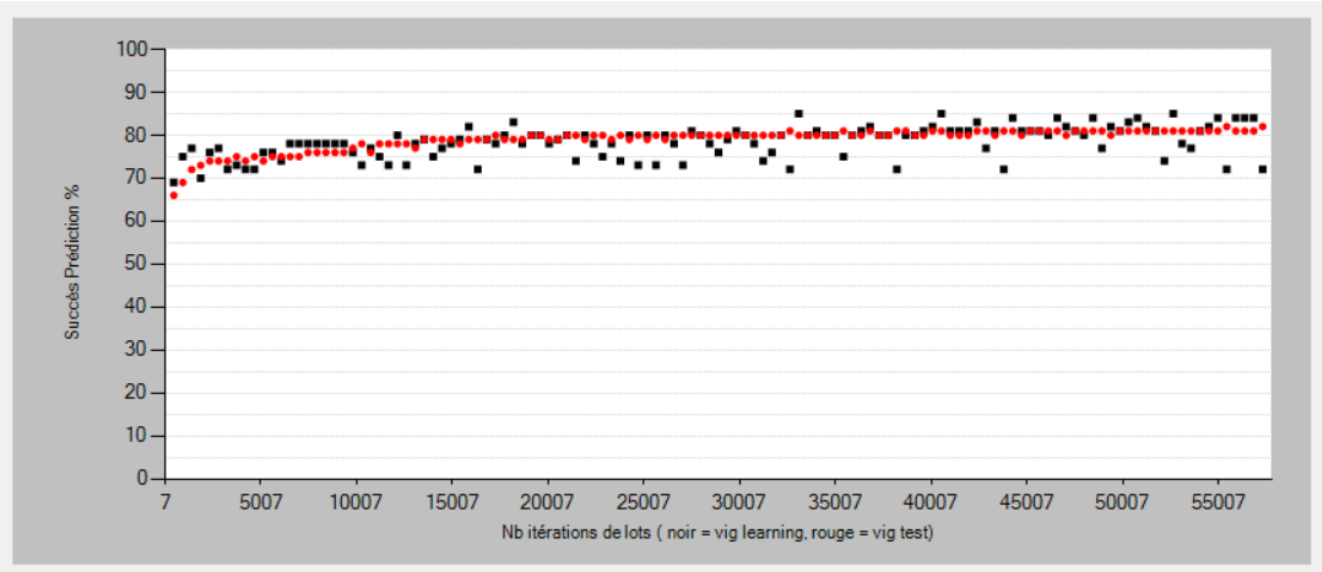
CNN de E:\Zoocam_learning\learn_cufes_PELGAS 2018\Learning

00_Artefact [9/9/100.00%]
00_Bubble [48/48/100.00%]
00_Bubbles_large [29/30/96.67%]
00_Detritus [79/100/79.00%]
00_Fiber [29/48/60.42%]
00_Halosphera [13/18/72.22%]
00_Paint_rust [1/7/14.29%]
03_Copepoda_large [73/100/73.00%]
03_Copepoda_small [80/100/80.00%]
04_Malacostraca_large [18/43/41.86%]

05_Cladocera [37/64/57.81%]
13_Engraulis_egg_1 [44/48/91.67%]
13_Engraulis_egg_2_3 [18/44/40.91%]
13_Engraulis_egg_4_6 [29/50/58.00%]
13_Engraulis_egg_7_8 [19/47/40.43%]
13_Engraulis_egg_9_11 [4/13/30.77%]
13_Fish_egg [34/45/75.56%]
13_Sardine_egg_1 [0/1/0.00%]
13_Sardine_egg_2_3 [41/52/78.85%]
13_Sardine_egg_4_6 [29/45/64.44%]
13_Sardine_egg_7_8 [0/2/0.00%]
13_Sardine_egg_9_11 [0/3/0.00%]
13_Sardine_egg_dam [18/32/56.25%]
Groupe 00 : 80.00 %
Groupe 03 : 76.50 %
Groupe 04 : 41.86 %
Groupe 05 : 57.81 %
Groupe 13 : 61.78 %
Total : 68.70 %

2. En DNN et utilisation du fichier learning.pid

Temps de Calcul : 16 minutes

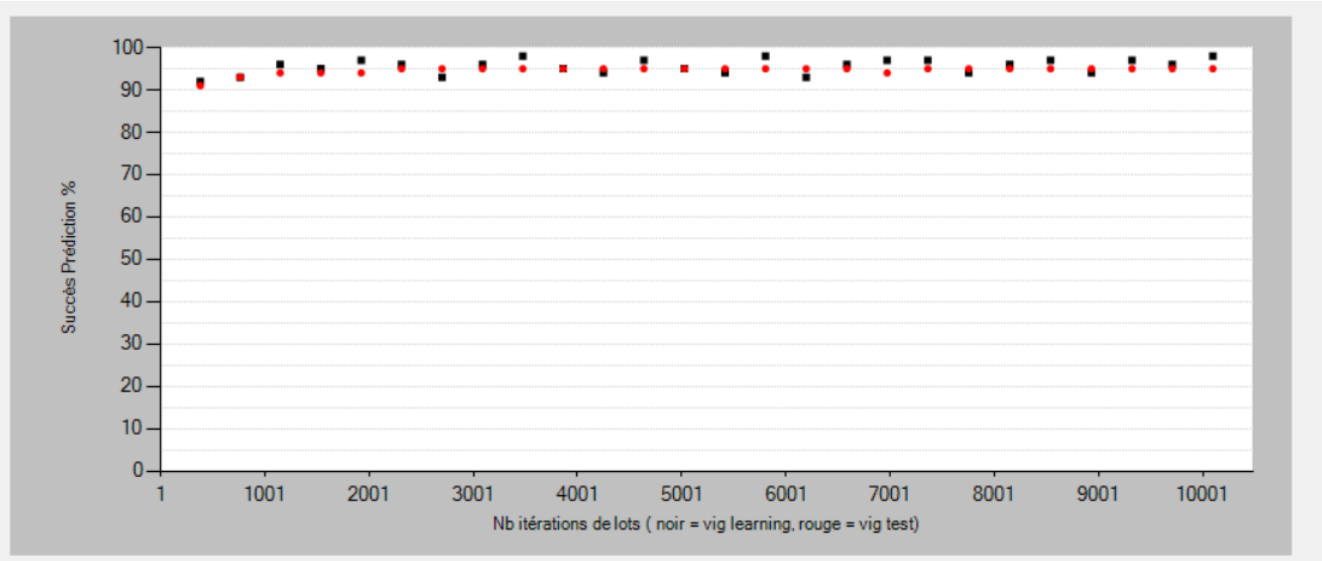


00_Artefact [5/5/100.00%]
00_Artefact_2 [1/4/25.00%]
00_Bubble [48/48/100.00%]
00_Bubbles_large [29/30/96.67%]
00_Detritus [102/114/89.47%]

00_Fiber [44/48/91.67%]
00_Halosphera [9/14/64.29%]
00_Paint_rust [6/7/85.71%]
03_Copepoda_large [118/137/86.13%]
03_Copepoda_small [107/122/87.70%]
04_Malacostraca_large [23/43/53.49%]
05_Cladocera [57/64/89.06%]
13_Engraulis_egg_1 [46/47/97.87%]
13_Engraulis_egg_2_3 [25/44/56.82%]
13_Engraulis_egg_4_6 [29/50/58.00%]
13_Engraulis_egg_7_8 [29/47/61.70%]
13_Engraulis_egg_9_11 [4/13/30.77%]
13_Fish_egg [42/44/95.45%]
13_Sardine_egg_1 [0/1/0.00%]
13_Sardine_egg_2_3 [45/52/86.54%]
13_Sardine_egg_4_6 [39/45/86.67%]
13_Sardine_egg_7_8 [0/2/0.00%]
13_Sardine_egg_9_11 [0/3/0.00%]
13_Sardine_egg_dam [27/32/84.38%]
Groupe 00 : 90.37 %
Groupe 03 : 86.87 %
Groupe 04 : 53.49 %
Groupe 05 : 89.06 %
Groupe 13 : 75.26 %
Total : 82.19 %

3. En DNN et utilisation du fichier learning_cnn.txt

Temps de calculs : 3 minutes



00_Artefact [5/5/100.00%]
00_Bubble [48/48/100.00%]
00_Bubbles_large [30/30/100.00%]
00_Detritus [105/114/92.11%]
00_Fiber [41/48/85.42%]
00_Halosphera [11/14/78.57%]
00_Paint_rust [4/7/57.14%]
03_Copepoda_large [134/137/97.81%]
03_Copepoda_small [117/122/95.90%]
04_Malacostraca_large [42/43/97.67%]
05_Cladocera [64/64/100.00%]
13_Engraulis_egg_1 [47/47/100.00%]
13_Engraulis_egg_2_3 [43/44/97.73%]
13_Engraulis_egg_4_6 [45/50/90.00%]
13_Engraulis_egg_7_8 [44/47/93.62%]
13_Engraulis_egg_9_11 [12/13/92.31%]
13_Fish_egg [42/44/95.45%]
13_Sardine_egg_1 [0/1/0.00%]
13_Sardine_egg_2_3 [51/52/98.08%]
13_Sardine_egg_4_6 [45/45/100.00%]
13_Sardine_egg_7_8 [2/2/100.00%]
13_Sardine_egg_9_11 [3/3/100.00%]
13_Sardine_egg_dam [31/32/96.88%]
Groupe 00 : 91.73 %
Groupe 03 : 96.91 %
Groupe 04 : 97.67 %
Groupe 05 : 100.00 %
Groupe 13 : 96.05 %
Total : 95.45 %

Est-ce la bonne combinaison ? Nous le verrons dans le test avec l'échantillon 308

4. Avec RF et utilisation du fichier learning.pid

1016 individus test, 6788 individus Learning sélectionnés 24 catégories

00_Artefact [1/5/20.00%]
00_Artefact_2 [1/4/25.00%]
00_Bubble [48/48/100.00%]
00_Bubbles_large [29/30/96.67%]
00_Detritus [96/114/84.21%]
00_Fiber [43/48/89.58%]
00_Halosphera [8/14/57.14%]
00_Paint_rust [5/7/71.43%]

03_Copepoda_large [126/137/91.97%]
03_Copepoda_small [110/122/90.16%]
04_Malacostraca_large [16/43/37.21%]
05_Cladocera [61/64/95.31%]
13_Engraulis_egg_1 [44/47/93.62%]
13_Engraulis_egg_2_3 [12/44/27.27%]
13_Engraulis_egg_4_6 [32/50/64.00%]
13_Engraulis_egg_7_8 [34/47/72.34%]
13_Engraulis_egg_9_11 [1/13/7.69%]
13_Fish_egg [40/44/90.91%]
13_Sardine_egg_1 [0/1/0.00%]
13_Sardine_egg_2_3 [46/52/88.46%]
13_Sardine_egg_4_6 [34/45/75.56%]
13_Sardine_egg_7_8 [0/2/0.00%]
13_Sardine_egg_9_11 [0/3/0.00%]
13_Sardine_egg_dam [26/32/81.25%]
Bonnes prédictions = 80.02%
Groupe 00 : 85.56 %
Groupe 03 : 91.12 %
Groupe 04 : 37.21 %
Groupe 05 : 95.31 %
Groupe 13 : 70.79 %
Total : 80.02 %

5. Avec RF et utilisation du fichier learning_cnn.txt

Random Forest
1012 individus test, 6758 individus Learning sélectionnés 23 catégories
00_Artefact [3/5/60.00%]
00_Bubble [48/48/100.00%]
00_Bubbles_large [30/30/100.00%]
00_Detritus [113/114/99.12%]
00_Fiber [45/48/93.75%]
00_Halosphera [13/14/92.86%]
00_Paint_rust [7/7/100.00%]
03_Copepoda_large [135/137/98.54%]
03_Copepoda_small [117/122/95.90%]
04_Malacostraca_large [40/43/93.02%]
05_Cladocera [63/64/98.44%]
13_Engraulis_egg_1 [45/47/95.74%]
13_Engraulis_egg_2_3 [37/44/84.09%]
13_Engraulis_egg_4_6 [46/50/92.00%]
13_Engraulis_egg_7_8 [46/47/97.87%]
13_Engraulis_egg_9_11 [12/13/92.31%]

13_Fish_egg [44/44/100.00%]
13_Sardine_egg_1 [1/1/100.00%]
13_Sardine_egg_2_3 [52/52/100.00%]
13_Sardine_egg_4_6 [41/45/91.11%]
13_Sardine_egg_7_8 [0/2/0.00%]
13_Sardine_egg_9_11 [3/3/100.00%]
13_Sardine_egg_dam [32/32/100.00%]
Bonnes prédictions = 96.15%
Groupe 00 : 97.37 %
Groupe 03 : 97.30 %
Groupe 04 : 93.02 %
Groupe 05 : 98.44 %
Groupe 13 : 94.47 %
Total : 96.15 %

Cela prouve que les individus tous biens prédits par CNN sont aussi bien prédits par DNN et RF. La bonne correspondance entre les paramètres morphologiques et l'agencement des pixels des vignettes est vérifiée.

7. Prédiction de l'échantillon 308 avec le learning Pelgas 2018 Martin Huret

1. Avec CNN

00_Artefact[0/1/.0%]
00_Bubble[29/42/69.0%]
00_Bubbles_large[5/10/50.0%]
00_Detritus[342/383/89.3%]
00_Fiber[44/98/44.9%]
00_Halosphera[9/11/81.8%]
00_Paint_rust[0/5/.0%]
03_Copepoda_large[250/299/83.6%]
03_Copepoda_small[336/423/79.4%]
04_Malacostraca_large[16/79/20.3%]
05_Cladocera[10/31/32.3%]
13_Engraulis_egg_1[67/77/87.0%]
13_Engraulis_egg_2_3[13/21/61.9%]
13_Engraulis_egg_4_6[3/8/37.5%]
13_Engraulis_egg_7_8[3/10/30.0%]
13_Engraulis_egg_9_11[0/4/.0%]
13_Fish_egg[12/18/66.7%]

13_Sardine_egg_1[0/0/NaN%]
13_Sardine_egg_2_3[26/32/81.3%]
13_Sardine_egg_4_6[10/16/62.5%]
13_Sardine_egg_7_8[0/3/.0%]
13_Sardine_egg_9_11[0/2/.0%]
13_Sardine_egg_dam[5/29/17.2%]
_Autres[0/0/NaN%]
Groupe 00 : 78.00 %
Groupe 03 : 81.16 %
Groupe 04 : 20.25 %
Groupe 05 : 32.26 %
Groupe 13 : 63.18 %
Total : 73.66 %

Les prédictions sont semblables à l'autotest, et les performances sont insuffisantes.

2. Avec RF

00_Artefact[0/0/NaN%]
00_Artefact_2[0/0/NaN%]
00_Bubble[32/44/72.7%]
00_Bubbles_large[6/7/85.7%]
00_Detritus[364/375/97.1%]
00_Fiber[57/86/66.3%]
00_Halosphera[9/9/100.0%]
00_Paint_rust[0/0/NaN%]
03_Copepoda_large[317/337/94.1%]
03_Copepoda_small[392/440/89.1%]
04_Malacostraca_large[33/51/64.7%]
05_Cladocera[18/28/64.3%]
13_Engraulis_egg_1[68/73/93.2%]
13_Engraulis_egg_2_3[14/16/87.5%]
13_Engraulis_egg_4_6[9/17/52.9%]
13_Engraulis_egg_7_8[4/7/57.1%]
13_Engraulis_egg_9_11[0/0/NaN%]
13_Fish_egg[15/18/83.3%]
13_Sardine_egg_1[0/0/NaN%]
13_Sardine_egg_2_3[29/36/80.6%]
13_Sardine_egg_4_6[10/15/66.7%]
13_Sardine_egg_7_8[0/0/NaN%]
13_Sardine_egg_9_11[0/0/NaN%]
13_Sardine_egg_dam[9/43/20.9%]
_Autres[0/0/NaN%]

Groupe 00 : 89.83 %
Groupe 03 : 91.25 %
Groupe 04 : 64.71 %
Groupe 05 : 64.29 %
Groupe 13 : 70.22 %
Total : 86.52 %

RF prédit ici mieux que CNN.

3. Avec DNN

00_Artefact[0/0/NaN%]
00_Artefact_2[0/0/NaN%]
00_Bubble[32/44/72.7%]
00_Bubbles_large[6/6/100.0%]
00_Detritus[386/417/92.6%]
00_Fiber[53/74/71.6%]
00_Halosphera[8/9/88.9%]
00_Paint_rust[0/1/.0%]
03_Copepoda_large[314/337/93.2%]
03_Copepoda_small[382/416/91.8%]
04_Malacostraca_large[30/64/46.9%]
05_Cladocera[17/22/77.3%]
13_Engraulis_egg_1[68/68/100.0%]
13_Engraulis_egg_2_3[23/27/85.2%]
13_Engraulis_egg_4_6[8/11/72.7%]
13_Engraulis_egg_7_8[4/4/100.0%]
13_Engraulis_egg_9_11[0/0/NaN%]
13_Fish_egg[16/17/94.1%]
13_Sardine_egg_1[0/0/NaN%]
13_Sardine_egg_2_3[31/38/81.6%]
13_Sardine_egg_4_6[9/17/52.9%]
13_Sardine_egg_7_8[0/0/NaN%]
13_Sardine_egg_9_11[0/0/NaN%]
13_Sardine_egg_dam[5/30/16.7%]
_Autres[0/0/NaN%]
Groupe 00 : 88.02 %
Groupe 03 : 92.43 %
Groupe 04 : 46.88 %
Groupe 05 : 77.27 %
Groupe 13 : 77.36 %
Total : 86.89 %

Mais DNN fait encore mieux.

4. Avec RF Variante A

00_Artefact[0/0/NaN%]
00_Artefact_2[0/0/NaN%]
00_Bubble[30/42/71.4%]
00_Bubbles_large[6/7/85.7%]
00_Detritus[365/385/94.8%]
00_Fiber[50/80/62.5%]
00_Halosphera[8/9/88.9%]
00_Paint_rust[0/0/NaN%]
03_Copepoda_large[320/349/91.7%]
03_Copepoda_small[387/437/88.6%]
04_Malacostraca_large[26/43/60.5%]
05_Cladocera[16/27/59.3%]
13_Engraulis_egg_1[68/74/91.9%]
13_Engraulis_egg_2_3[14/20/70.0%]
13_Engraulis_egg_4_6[5/12/41.7%]
13_Engraulis_egg_7_8[4/7/57.1%]
13_Engraulis_egg_9_11[0/0/NaN%]
13_Fish_egg[14/17/82.4%]
13_Sardine_egg_1[0/0/NaN%]
13_Sardine_egg_2_3[29/36/80.6%]
13_Sardine_egg_4_6[12/16/75.0%]
13_Sardine_egg_7_8[0/0/NaN%]
13_Sardine_egg_9_11[0/0/NaN%]
13_Sardine_egg_dam[7/41/17.1%]
_Autres[0/0/NaN%]
Groupe 00 : 87.76 %
Groupe 03 : 89.95 %
Groupe 04 : 60.47 %
Groupe 05 : 59.26 %
Groupe 13 : 68.61 %
Total : 84.96 %

Résultats insuffisants pour les œufs.

5. Avec RF Variante B

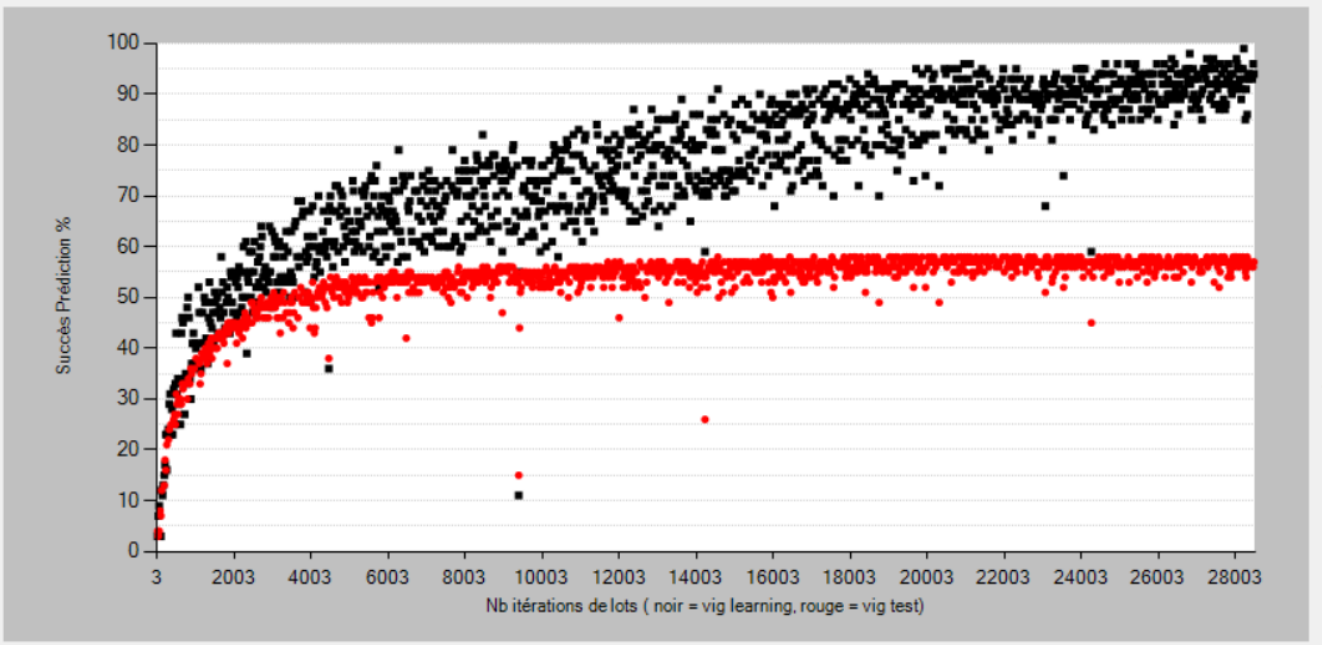
00_Artefact[0/0/NaN%]
00_Artefact_2[0/0/NaN%]

00_Bubble[31/42/73.8%]
00_Bubbles_large[6/7/85.7%]
00_Detritus[370/386/95.9%]
00_Fiber[53/77/68.8%]
00_Halosphera[9/10/90.0%]
00_Paint_rust[0/0/NaN%]
03_Copepoda_large[323/348/92.8%]
03_Copepoda_small[391/436/89.7%]
04_Malacostraca_large[30/51/58.8%]
05_Cladocera[17/22/77.3%]
13_Engraulis_egg_1[67/70/95.7%]
13_Engraulis_egg_2_3[12/16/75.0%]
13_Engraulis_egg_4_6[10/22/45.5%]
13_Engraulis_egg_7_8[3/5/60.0%]
13_Engraulis_egg_9_11[0/0/NaN%]
13_Fish_egg[14/15/93.3%]
13_Sardine_egg_1[0/0/NaN%]
13_Sardine_egg_2_3[29/35/82.9%]
13_Sardine_egg_4_6[10/15/66.7%]
13_Sardine_egg_7_8[0/0/NaN%]
13_Sardine_egg_9_11[0/1/.0%]
13_Sardine_egg_dam[8/44/18.2%]
Groupe 00 : 89.85 %
Groupe 03 : 91.07 %
Groupe 04 : 58.82 %
Groupe 05 : 77.27 %
Groupe 13 : 68.61 %
Total : 86.33 %
Résultats insuffisants pour les œufs

Quelque soit le classifieur, le taux de bonne prédiction est très moyen. On peut supposer que le nombre d'individus qui compose ce learning est insuffisant.

8. Utilisation du Learning Ecotaxa Pelgas 2016

Tous les individus sont sélectionnés. Rappel : il n'y a pas de fichier PID, donc on ne peut faire que du CNN après avoir « nettoyé » les vignettes JPEG.



Après 4 heures de calculs :

Actinopterygii_egg [73/100/73.00%]
artefact_bubble [94/100/94.00%]
Bacillariophyceae_Diatoma [82/100/82.00%]
Brachyura_megalopa [49/76/64.47%]
Calanoida_Acartiidae [42/100/42.00%]
Calanoida_Calanidae [39/100/39.00%]
Calanoida_Centropagidae [46/100/46.00%]
Calanoida_Euchaetidae [0/8/0.00%]
Calanoida_Pontellidae [16/40/40.00%]
Calanoida_Temoridae [58/100/58.00%]
Centropagidae_Centropages [45/100/45.00%]
Centropagidae_Isias [38/69/55.07%]
Cirripedia_nauplii [36/80/45.00%]
Copepoda_dead [33/100/33.00%]
Copepoda_Harpacticoida [0/7/0.00%]
Copepoda_multiple [50/100/50.00%]
Copepoda_Poecilostomatoida [82/100/82.00%]
Crustacea_larvae [82/100/82.00%]
Crustacea_nauplii [16/61/26.23%]
Cyclopoida_Oithonidae [91/100/91.00%]
Decapoda_zoea [39/100/39.00%]
detritus_fiber [75/98/76.53%]
Diplostraca_Cladocera [40/100/40.00%]
Engraulidae temp_egg 1 temp [86/100/86.00%]
Engraulidae temp_egg 2 3 temp [57/100/57.00%]

Engraulidae temp_egg 4 6 temp [48/98/48.98%]
Engraulidae temp_egg 7 8 temp [31/80/38.75%]
Engraulidae temp_egg 9 11 temp [41/58/70.69%]
Engraulidae temp_egg unkn temp [0/2/0.00%]
Eumalacostraca_Amphipoda [15/28/53.57%]
Eumalacostraca_Decapoda [2/14/14.29%]
Eumalacostraca_Euphausiacea [64/100/64.00%]
Gnathostomata_Actinopterygii [4/13/30.77%]
Harosa_Rhizaria [72/89/80.90%]
Limacinidae_Limacina [32/73/43.84%]
Maxillopoda_Copepoda [48/100/48.00%]
Metazoa_Chaetognatha [28/52/53.85%]
Metazoa_Echinodermata [4/12/33.33%]
Mollusca_egg [15/23/65.22%]
Noctilucaceae_Noctiluca [51/58/87.93%]
not-living_artefact [64/69/92.75%]
not-living_detritus [52/100/52.00%]
other_egg [32/69/46.38%]
other_gelatinous [33/66/50.00%]
other_multiple [22/100/22.00%]
Sardina temp_egg 1 temp [19/33/57.58%]
Sardina temp_egg 2 3 temp [62/100/62.00%]
Sardina temp_egg 4 6 temp [55/93/59.14%]
Sardina temp_egg 7 8 temp [19/48/39.58%]
Sardina temp_egg 9 11 temp [3/23/13.04%]
Sardina temp_egg unkn temp [54/100/54.00%]
Tunicata_Appendicularia [2/12/16.67%]

Loss: 0.252
Train accuracy: 94%
Test accuracy :57%

Le taux de bonnes prédictions est insuffisant dans les individus test. Comment interpréter ce plafond à 60% ?

9. Prédiction échantillon 308 avec Learning Ecotaxa Pelgas 2016

Je désactive l'effacement des coins puisque les vignettes de l'échantillon 308 n'en ont pas, et une fois la prédiction faite, je renomme manuellement les répertoires des

œufs pour pouvoir effectuer la performance: *Engraulidae temp_egg 1 temp* devient *13_Engraulis_egg_1*, etc…

Voici les résultats par rapport au répertoire validé:
13_Engraulis_egg_1[53/65/81.5%]
13_Engraulis_egg_2_3[10/22/45.5%]
13_Engraulis_egg_4_6[2/3/66.7%]
13_Engraulis_egg_7_8[3/3/100.0%]
13_Engraulis_egg_9_11[0/0/NaN%]
13_Sardine_egg_1[0/6/.0%]
13_Sardine_egg_2_3[18/22/81.8%]
13_Sardine_egg_4_6[4/7/57.1%]
13_Sardine_egg_7_8[0/2/.0%]
13_Sardine_egg_9_11[0/0/NaN%]
13_Sardine_egg_dam[4/24/16.7%]

Groupe 13 : 61.04 %

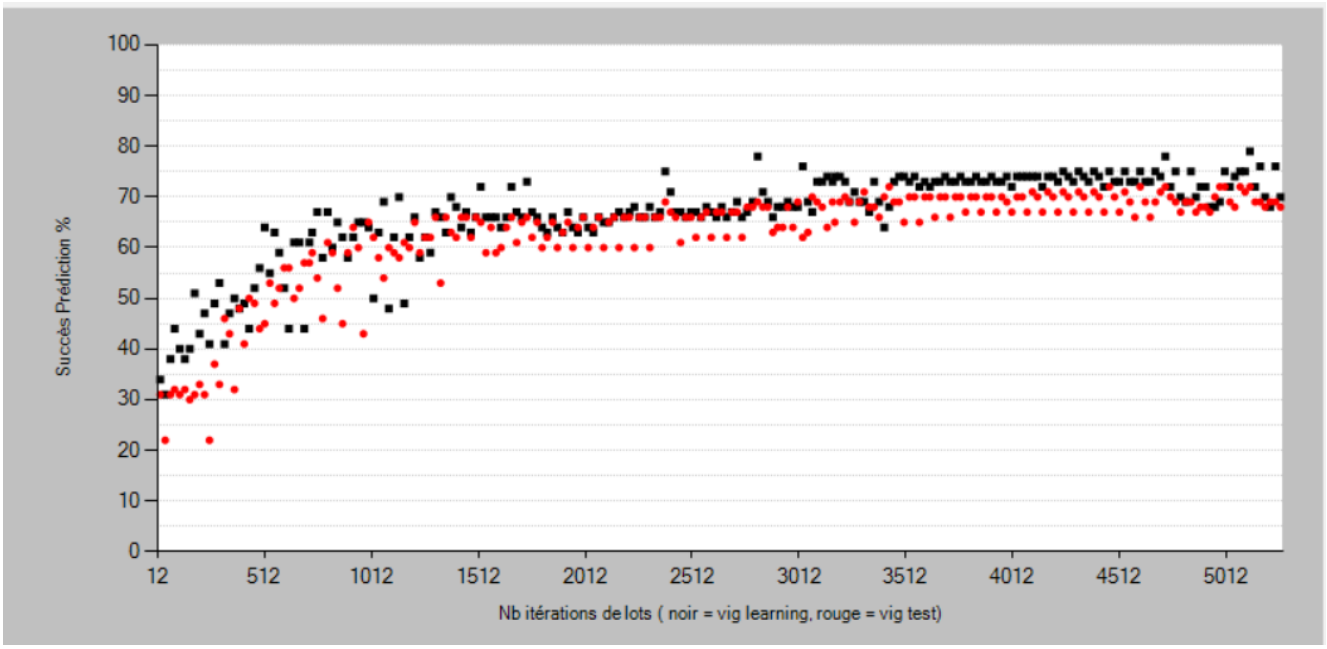
Le résultat est mitigé. Le nombre insuffisant de vignettes par catégorie et la forte compression des images rendent CNN peu efficace.

10. Etude avec tous les œufs d'anchois de Pelgas 2018 Leg 1

Nous allons déterminer si CNN est performant quand il s'agit de classer correctement les stades dans un learning composé uniquement d'œufs d'anchois, en réglant certains de ses paramètres. Puis nous allons comparer avec RF et DNN. Comme pour le Megalearning réduit, nous nous attendons à de très bonnes performances, mais les catégories se ressemblent plus, il est donc utile d'effectuer cette série de tests.

1. CNN Paramètres A

Max taille vignettes : 300, taille matrice CNN : 56*56, temps de calcul : 2000 secondes (33 minutes)



13_Engraulis_egg_1 [97/100/97.00%]
13_Engraulis_egg_2_3 [85/100/85.00%]
13_Engraulis_egg_4_6 [40/100/40.00%]
13_Engraulis_egg_7_8 [81/100/81.00%]
13_Engraulis_egg_9_11 [3/48/6.25%]
Total : 68.30 %

Matrice de confusion :

Matrice de confusion

REAL\PRED	13 Engraulis egg 1	13 Engraulis egg 2 3	13 Engraulis egg 4 6	13 Engraulis egg 7 8	13 Engraulis egg 9 11
13 Engraulis egg 1	**97 / 100**	3	0	0	0
13 Engraulis egg 2 3	11	**85 / 100**	3	1	0
13 Engraulis egg 4 6	1	42	**40 / 100**	17	0
13 Engraulis egg 7 8	0	10	9	**81 / 100**	0
13 Engraulis egg 9 11	0	4	3	38	**3 / 48**

Le premier stade est très bien classé, pour les autres, j'aurai personnellement des difficultés à classer moi même. Quand au 9-11, le manque d'effectifs (320) est pénalisé.

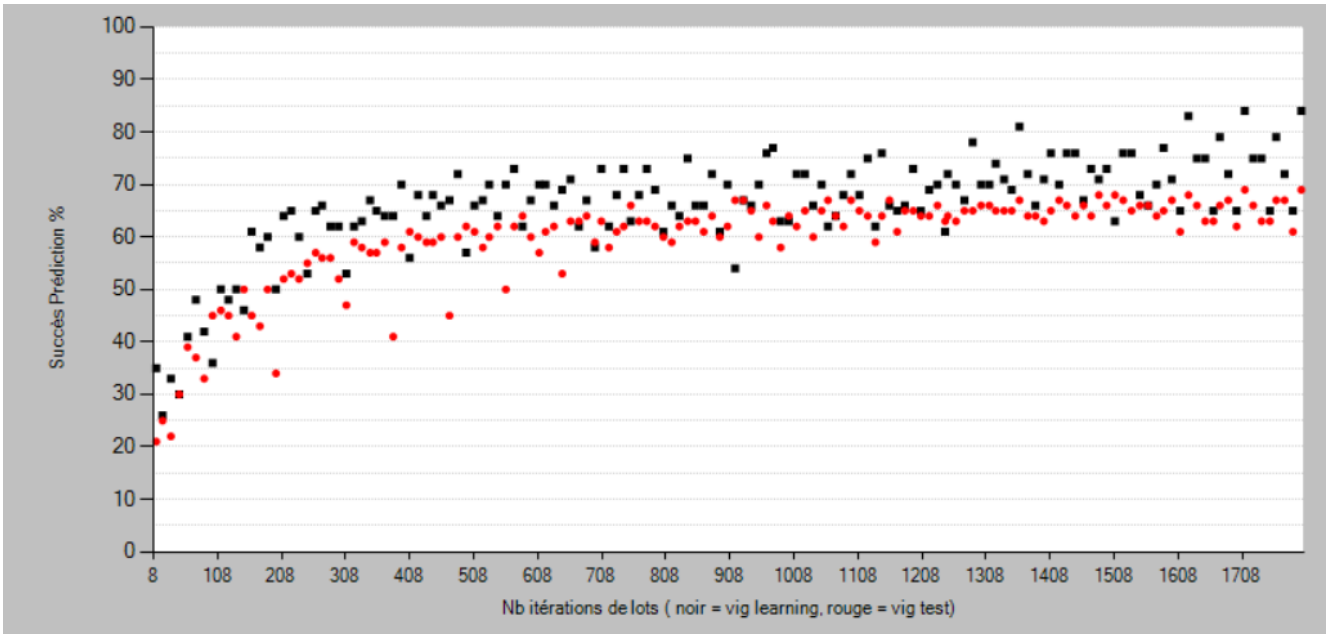
2. CNN Paramètres B

Max taille vignettes : 200, taille matrice CNN : 120*120, batchsize = 200 (500 plante le programme par manque de mémoire GPU. Avec 200, 4Go de ram GPU sont utilisées)
Temps de calcul : 25 minutes
13_Engraulis_egg_1 [100/100/100.00%]

13_Engraulis_egg_2_3 [0/100/0.00%]
13_Engraulis_egg_4_6 [0/100/0.00%]
13_Engraulis_egg_7_8 [0/100/0.00%]
13_Engraulis_egg_9_11 [0/48/0.00%]
ça ne fonctionne pas du tout !

3. CNN Paramètre C

Pas de mise à l'échelle du redimensionnement, taille CNN : 120*120, temps de calcul : 25 minutes



13_Engraulis_egg_1 [95/100/95.00%]
13_Engraulis_egg_2_3 [80/100/80.00%]
13_Engraulis_egg_4_6 [58/100/58.00%]
13_Engraulis_egg_7_8 [73/100/73.00%]
13_Engraulis_egg_9_11 [4/48/8.33%]
Groupe 13 : 69.20 %
Total : 69.20 %

REAL\PRED	13 Engraulis egg 1	13 Engraulis egg 2 3	13 Engraulis egg 4 6	13 Engraulis egg 7 8	13 Engraulis egg 9 11
13 Engraulis egg 1	**95 / 100**	5	0	0	0
13 Engraulis egg 2 3	11	**80 / 100**	8	1	0
13 Engraulis egg 4 6	0	27	**58 / 100**	15	0
13 Engraulis egg 7 8	0	8	18	**73 / 100**	1
13 Engraulis egg 9 11	0	2	3	39	**4 / 48**

Le résultat comparable au « Paramètres A »

4. CNN Paramètres D

Max taille vignettes : 300, taille matrice CNN : 56*56, ajout d'une 3eme couche à l'architecture, identique à la 1 et 2. Temps de calcul : 15 minutes

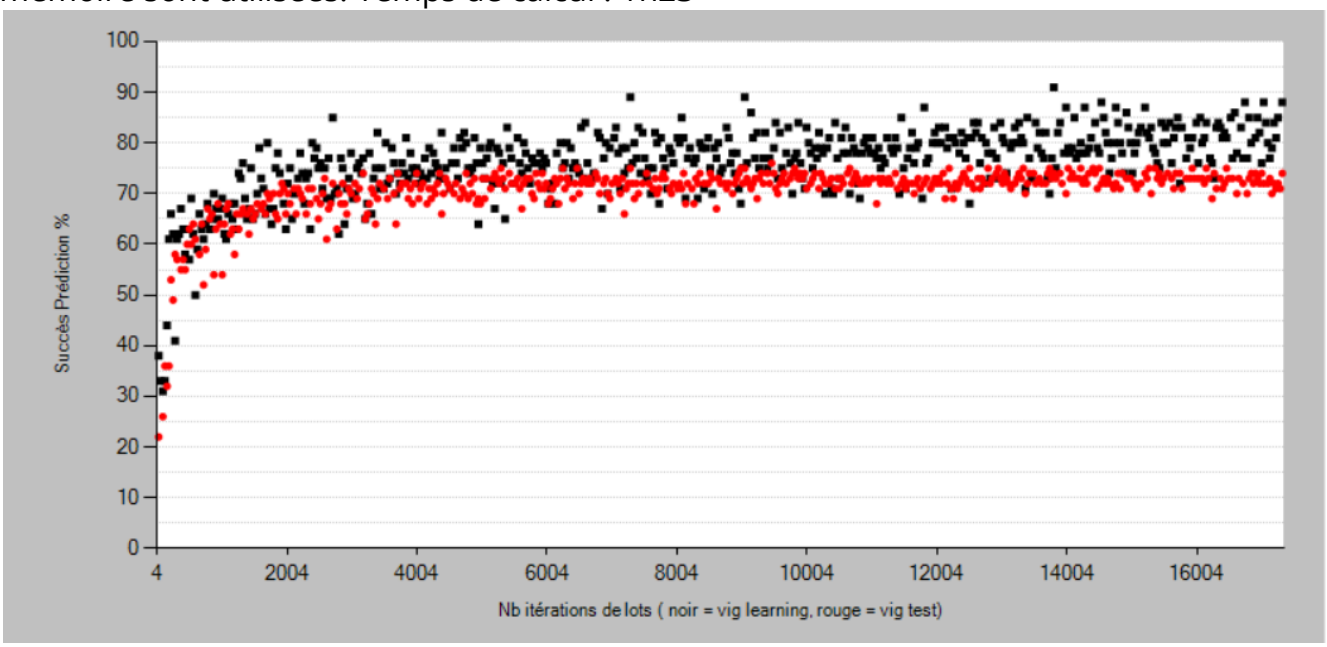
13_Engraulis_egg_1 [97/100/97.00%]
13_Engraulis_egg_2_3 [80/100/80.00%]
13_Engraulis_egg_4_6 [70/100/70.00%]
13_Engraulis_egg_7_8 [82/100/82.00%]
13_Engraulis_egg_9_11 [3/48/6.25%]
CNN
CNN de E:\Zoocam_learning\valid_pelgas2018_mix
Groupe 13 : 74.11 %
Total : 74.11 %

5. CNN Paramètres E

Identique à D (3 couches) mais avec les volumes à 30 au lieu de 16 (le GPU est donc plus sollicité : 20% au lieu de 10%)
Résultat : les deux courbes de prédiction stagnent à 10% pendant plusieurs minutes de calcul.

6. CNN Paramètres F

3 couches volumes à 30, matrice CNN 96*96, batch size 100 (sinon pas assez de mémoire GPU), max taille vignettes 300. Le GPU se charge à 33%, plus de 4 Go GPU de mémoire sont utilisées. Temps de calcul : 1h23



13_Engraulis_egg_1 [98/100/98.00%]
13_Engraulis_egg_2_3 [82/100/82.00%]

13_Engraulis_egg_4_6 [65/100/65.00%]
13_Engraulis_egg_7_8 [77/100/77.00%]
13_Engraulis_egg_9_11 [10/48/20.83%]
Total : 74.11 %
Le résultat est meilleur que « Paramètres A ». On peut donc jouer sur l'architecture du réseau pour augmenter significativement le taux de bonnes prédictions.
Malheureusement, il y a peu de vignettes au stade 9_11

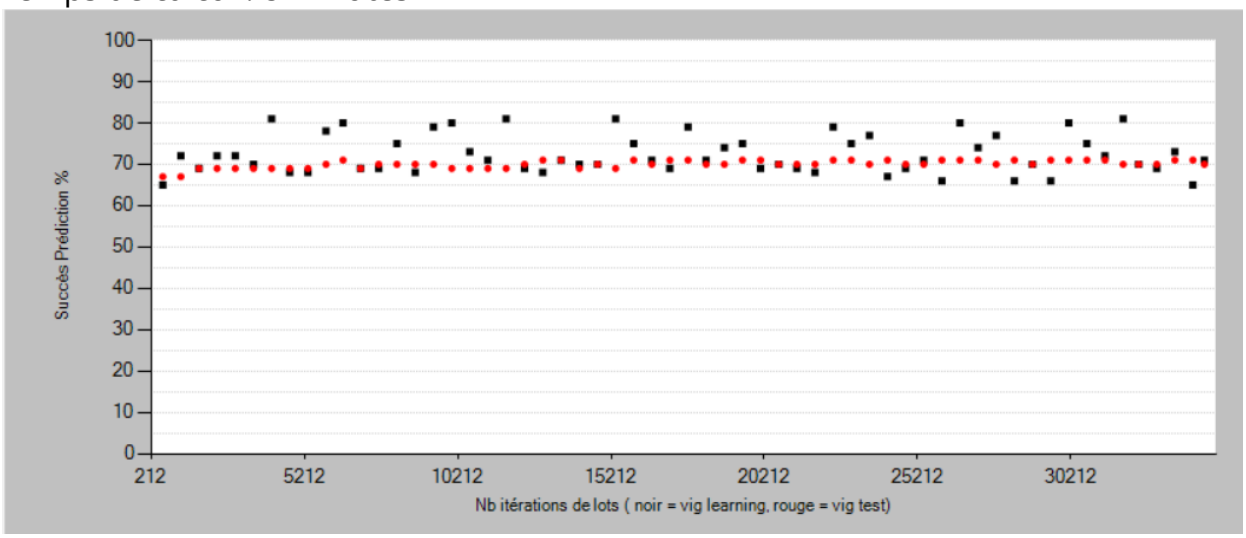
7. RF

Voici ce que cela donne avec le RF sur le Megalearning avec que les œufs d'anchois :
13_Engraulis_egg_1 [404/450/89.78%]
13_Engraulis_egg_2_3 [223/449/49.67%]
13_Engraulis_egg_4_6 [216/444/48.65%]
13_Engraulis_egg_7_8 [381/449/84.86%]
13_Engraulis_egg_9_11 [2/55/3.64%]
Random Forest 1847 individus test, 12312 individus Learning selectionnés 5 catégories
Bonnes prédictions = 66.38%
Groupe 13 : 66.38 %
Total : 66.38 %

REAL\PRED	13 Engraulis egg 1	13 Engraulis egg 2 3	13 Engraulis egg 4 6	13 Engraulis egg 7 8	13 Engraulis egg 9 11
13 Engraulis egg 1	**404 / 450**	38	5	3	0
13 Engraulis egg 2 3	39	**223 / 449**	123	64	0
13 Engraulis egg 4 6	3	65	**216 / 444**	160	0
13 Engraulis egg 7 8	0	5	62	**381 / 449**	1
13 Engraulis egg 9 11	0	2	3	48	**2 / 55**

8. DNN

Temps de calcul : 8 minutes



13_Engraulis_egg_1 [422/450/93.78%]

13_Engraulis_egg_2_3 [316/449/70.38%]
13_Engraulis_egg_4_6 [252/444/56.76%]
13_Engraulis_egg_7_8 [309/449/68.82%]
13_Engraulis_egg_9_11 [8/55/14.55%]
Groupe 13 : 70.76 %
Total : 70.76 %

REAL\PRED	13 Engraulis egg 1	13 Engraulis egg 2 3	13 Engraulis egg 4 6	13 Engraulis egg 7 8	13 Engraulis egg 9 11
13 Engraulis egg 1	**422 / 450**	27	1	0	0
13 Engraulis egg 2 3	45	**316 / 449**	82	6	0
13 Engraulis egg 4 6	1	114	**252 / 444**	77	0
13 Engraulis egg 7 8	1	21	116	**309 / 449**	2
13 Engraulis egg 9 11	0	1	5	41	**8 / 55**

Un peu meilleur que le RF, mais le CNN reste supérieur.

13_Engraulis_egg_9_11 pose encore problème dans cette série de tests : son faible taux de bonnes prédictions (car peu d'individus) à une grande influence sur le résultat général des classifieurs.

11. Synthèse

Mega Learning Tous indiv	Global	Groupe 13
CNN	60.67	59.72
RF learning.pid	75.94	69.58
RF learning_cnn.txt	89.53	87.41
DNN learning.pid	75.8	70.75
DNN learning_cnn.txt	88.68	87.16

Mega Learning +2000 indiv	Global	Groupe 13
CNN	76.62	78.2
RF learning.pid	79.36	72.11
RF learning_cnn.txt	-	-
DNN learning.pid	81.02	75.55
DNN learning_cnn.txt	-	-

Pred 308 Mega Learning Tous	Global	Groupe 13
CNN	73.66	68.2
RF	77.34	76.47
DNN	80.21	75.86
RF variante A	78.09	77.45
RF variante B	-	-

Pred 308 Mega Learning +2000 indiv	Global	Groupe 13
CNN	76.8	79.19
RF	78.96	82.39
DNN learning.pid	82.77	83.72
RF variante A	78.7	81.16
RF variante B	-	-

LearningPelgas 2018L1 Tous	Global	Groupe 13
CNN	68.7	61.78
RF learning.pid	80.02	70.79
RF learning_cnn.txt	96.15	94.47
DNN learning.pid	82.19	75.26
DNN learning_cnn.txt	95.45	96.05

Pred 308 Learning Pelgas2018L1 Tous	Global	Groupe 13
CNN	73.66	63.18
RF	86.52	70.22
DNN	86.89	77.36
RF variante A	84.96	68.61
RF variante B	86.33	68.61

12. Interprétations

- Le gain de prédictions avec le CNN est significatif (+10%), avec un nombre d'individus importants (>=2000) dans chaque catégories mais il s'explique aussi par la réduction du nombre de catégories. Cette variation est moins importante avec les autres classifieurs.
- Le DNN semble un peu plus performant que le RF de 3 à 4%
- RF dopé au CNN n'apporte pas de gain significatif (Variante A et B). Par contre, utiliser le learning généré par le CNN (fichier *learning_cnn.txt*) semble très intéressant.
- Le CNN peut se « planter » même avec un nombre important d'individus dans la catégorie (groupe 4 : Malacostraca). Confusion probable avec le groupe 5 (Cladocéra) , il y a peut être un problème de taille de la matrice CNN qui serait trop petite par rapport aux appendices de ces espèces. Il est probable que le redimensionnement des vignettes détruit d'autant plus d'information quand la variabilité des vignettes est grande. La taille de référence étant de 300*300 pixels, beaucoup de vignettes sont dégradées. CNN fonctionnera surement mieux avec un redimensionnement moins agressif, ce qui signifierait que les vignettes aient une taille semblable entre chaque catégorie. Il peut donc être délicat de choisir la taille de la matrice d'entrée de CNN en fonction des organismes traités.

- Pour les œufs, CNN rejoint les performances des autres classifieurs seulement s'il y a plus de 2000 individus. Moins il y a de catégories, plus CNN semble devenir performant. De plus, une bonne architecture (choix du nombre de couches) peut le rendre particulièrement efficace. Bien évidemment, la bonne qualité de compression de l'image est souhaitable (mauvais résultats avec Ecotaxa).

13. Conclusion

On a maintenant une meilleure idée des outils de classifications : RF est souple d'emploi (ajout de catégories à la volée) et sans surprise, DNN est moins souple, mais un peu plus performant, CNN peut être très bon ou franchement peu efficace suivant son paramétrage. Mon choix de n'avoir sélectionné qu'un seul échantillon test (CUFES 308) est critiquable, en effet ; il en aurait fallu 3 ou 4 mais il aurait fallu « sacrifier » de l'information sur les œufs de sardines si peu nombreux. On peut se pencher aussi sur la validation des échantillons, et surtout le manque d'œufs de sardines et des œufs d'anchois stade 9_11. Il faudrait compléter cette étude avec un jeu d'échantillons conséquent.

En vue de perfectionnement, de nouveaux algorithmes d'amélioration de netteté et contraste d'images émergent (wavelets), il serait intéressant d'étudier quels sont leurs effets sur les performances dans CNN…

Image brute *Image traitée avec Iris / wavelets*

14. Annexes

1. Effectifs de l'échantillon CUFES 308 validé

Taxons	Effectifs
00_Artefact	2
00_Bubble	32
00_Bubbles_large	7
00_Detritus	436
00_Fiber	64

00_Halosphera	11
00_Paint_rust	0
00_Phytoplankton	0
03_Copepoda_large	378
03_Copepoda_small	411
04_Malacostraca_large	42
05_Cladocera	31
13_Engraulis_egg_1	69
13_Engraulis_egg_2_3	26
13_Engraulis_egg_4_6	13
13_Engraulis_egg_7_8	4
13_Engraulis_egg_9_11	0
13_Fish_egg	18
13_Sardine_egg_1	0
13_Sardine_egg_2_3	35
13_Sardine_egg_4_6	14
13_Sardine_egg_7_8	0
13_Sardine_egg_9_11	0
13_Sardine_egg_dam	9
15_Multiple	0
17_Gasteropoda	0
17_Zooplankton	0
Cut_Objects	280
DUPLCT_Objects	541

2. Effectifs de Pelgas 2018 leg 1

Taxons	Effectifs
00_Artefact	2 107
00_Bubble	39 852
00_Bubbles_large	2 332
00_Detritus	120 465
00_Fiber	26 538
00_Halosphera	832
00_Paint_rust	292
00_Phytoplankton	143
03_Copepoda_large	58 046
03_Copepoda_Pontellidae	85
03_Copepoda_small	126 070
04_Malacostraca_large	4 996
05_Cladocera	427
13_Engraulis_egg_1	5 487
13_Engraulis_egg_2_3	5 396
13_Engraulis_egg_4_6	3 569
13_Engraulis_egg_7_8	3 204
13_Engraulis_egg_9_11	320
13_Fish_egg	1 697
13_Sardine_egg_1	45

13_Sardine_egg_2_3	1 947
13_Sardine_egg_4_6	1 082
13_Sardine_egg_7_8	187
13_Sardine_egg_9_11	170
13_Sardine_egg_dam	692
17_Gasteropoda	271

3. Composants du PC :

- Processeur I7 4790K 4 Cores + 4 HT (4.0 Ghz, 4.2 Ghz en turbo) année 2015
- 16 Go de ram DDR3 1600 Mhz
- Carte Nvidia K620 pour l'affichage
- Carte Nvidia 1080 gtx 8Go pour le calcul (variable d'environnement à modifier pour spécifier le GPU à utiliser parmi les 2 cartes)
- Kaspersky désinstallé, Microsoft Essentials installé.