Data Challenge D-IA-GNO-DENT

Association EISBM

Albert Saporta, Johann Pellet, Bertrand De Meulder

Health Data Hub, PariSanté Campus, 14 Décembre 2023

Contexte clinique

L'amélogenèse imparfaite (AI) est une dysplasie de l'émail d'origine génétique caractérisée par des anomalies du développement affectant la structure et l'apparence clinique de l'émail de toutes ou de quasiment toutes les dents temporaires et/ou permanentes (altération de la couleur, de la forme, de la surface et/ou de la structure de l'émail).

L'AI peut exister sous forme isolée ou syndromique lorsqu'elle est associée à d'autres manifestations.

Il existe quatre types d'AI classiquement définis par Witkop *et al* en 1988:

1. Type I: **Hypoplastic** (faible épaisseur voire absence d'émail, présence de puits ou stries)
2. Type II: **Hypomature** (émail d'épaisseur normale, dur, pas ou peu de contraste avec la dentine, coloration blanc crayeux à jaune brun
3. Type III: **Hypocalcifié/Hypominéralisé** (émail d'épaisseur normale, mou, de couleur jaune brun et se clivant rapidement
4. Type IV: **Hypoplastique/hypomature** avec **taurodontisme** (émail d'apparence mixte hypoplastique ou hypomature avec allongement de la chambre pulpaire

Type I

Type II

Type III

Taurodontisme

Contexte clinique (cont.)

La dentinogenèse imparfaite (DI) est une maladie d'origine génétique du développement des dents. Les dents sont décolorées (bleu gris à jaune-brun) et translucides.

Ces deux maladies sont liées à des mutations génétiques héréditaires.

L'AI est causée par des perturbations des processus de développement de la dent, souvent des mutations des gènes AMELX, ENAM, MMP20, FAM83H, KLK4, FAM20A.

La DI est le résultat de mutations dans les gènes codants pour la sialophosphoprotéine de la dentine (COL1A1 et COLI1A, DSPP) (Januarti *et al*, 2023, Simancas-Escorcia *et al*, 2018, Goldberg *et al*, 2019, An *et al*, 2015, Jaureguiberry *et al*, 2012).

D-IA-GNO-DENT challenge

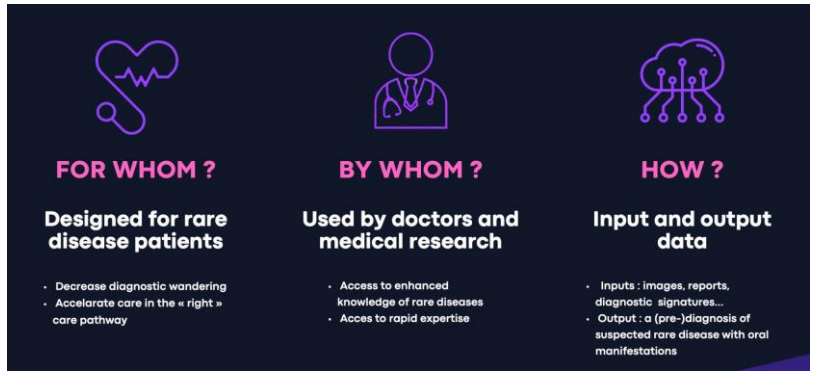
Challenge ouvert le 6 Avril 2023, fermé le 09 Juillet 2023.

Objectifs:

- Développer un algorithme de deep learning permettant d'accélérer le diagnostic de la maladie, en utilisant des données cliniques et visuelles (photos, radiographies)
- Le modèle devra être capable de distinguer:
  - Quelle maladie est présentée ?
    - Absence,
    - Amélogénèse Imparfaite,
    - Dentinogenèse imparfaite
  - Quels symptômes sont présents ?
    - Hypomature
    - Hypoplastic
    - Hypocalcification/Hypomineralisation
    - Taurodontisme
  - Quel gène est principalement responsable ?
  - Si le statut est syndromique ou isolé

Contraintes:

- L'algorithme devra être entraîné en moins de 6 heures et être capable d'inférer une nouvelle image en moins d'une minute
- Il devra fournir des SHAP values ou équivalent pour expliquer ses décisions



Données disponibles pour l'entrainement

254 patients

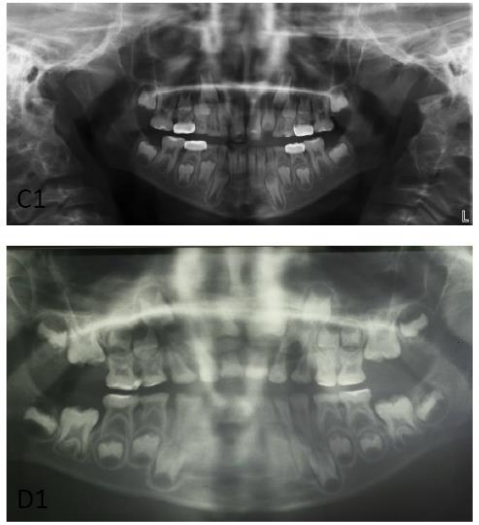
	A	B	C	D	E	F
1	Patient_ID	Patient_gender	Cohort	AI_Type	Responsible_Gene_Name	Is_Isolated_Syndromic
2	PAT_0	M	Amelogenesis Imperfecta	AI_Hypomature	MMP20	Isolated
3	PAT_1	F	Amelogenesis Imperfecta	AI_Hypomature	ENAM	Isolated
4	PAT_2	F	Amelogenesis Imperfecta	AI_Hypomature	WDR72	Isolated

- 46 patients AI type 1 (hypoplastic)
- 46 patients AI type II (hypomature)
- 16 patients AI type III (hypocalcification)
- 7 patients AI type IV (hypomature/hypoplastic/taurodontism)
- 30 patients Dentinogenese imparfaite
- 109 patients Control

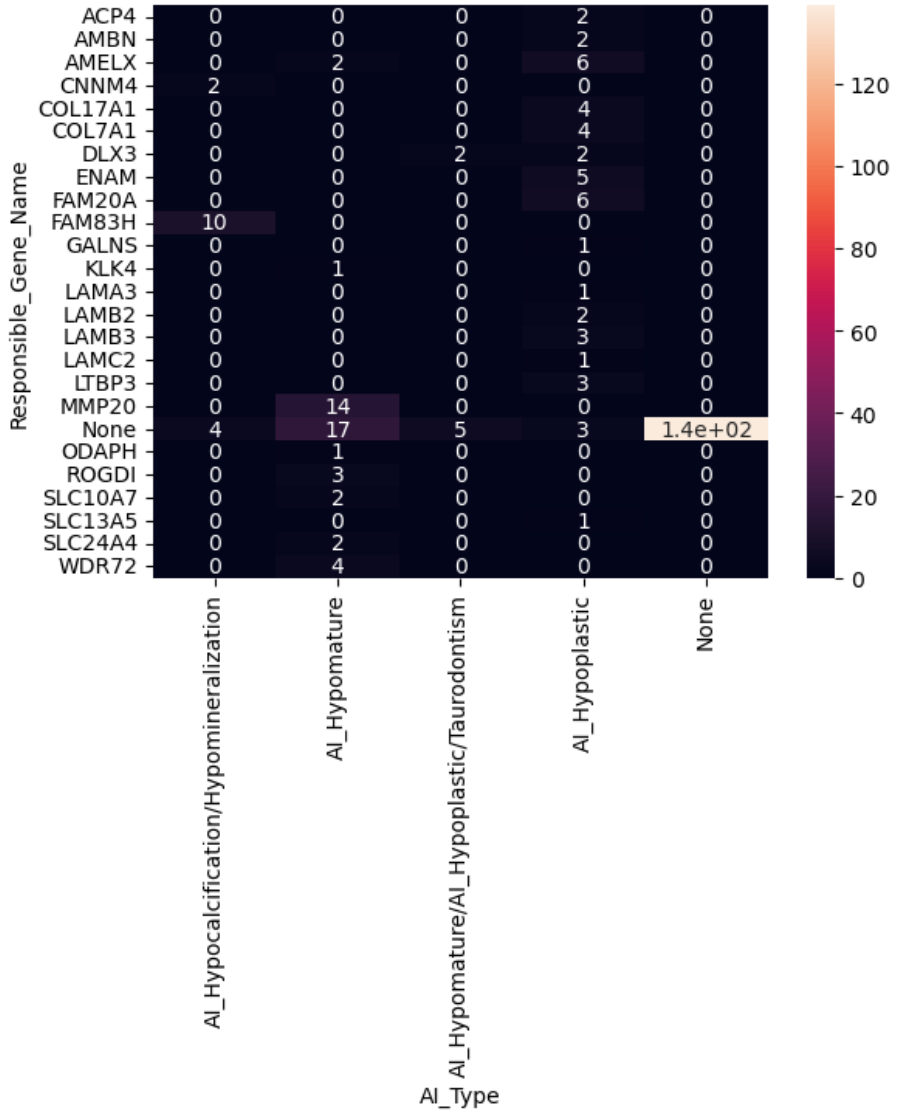
- Plusieurs images par patient (nombre variable).
- Images de types photographies couleurs et de radiographies panoramiques

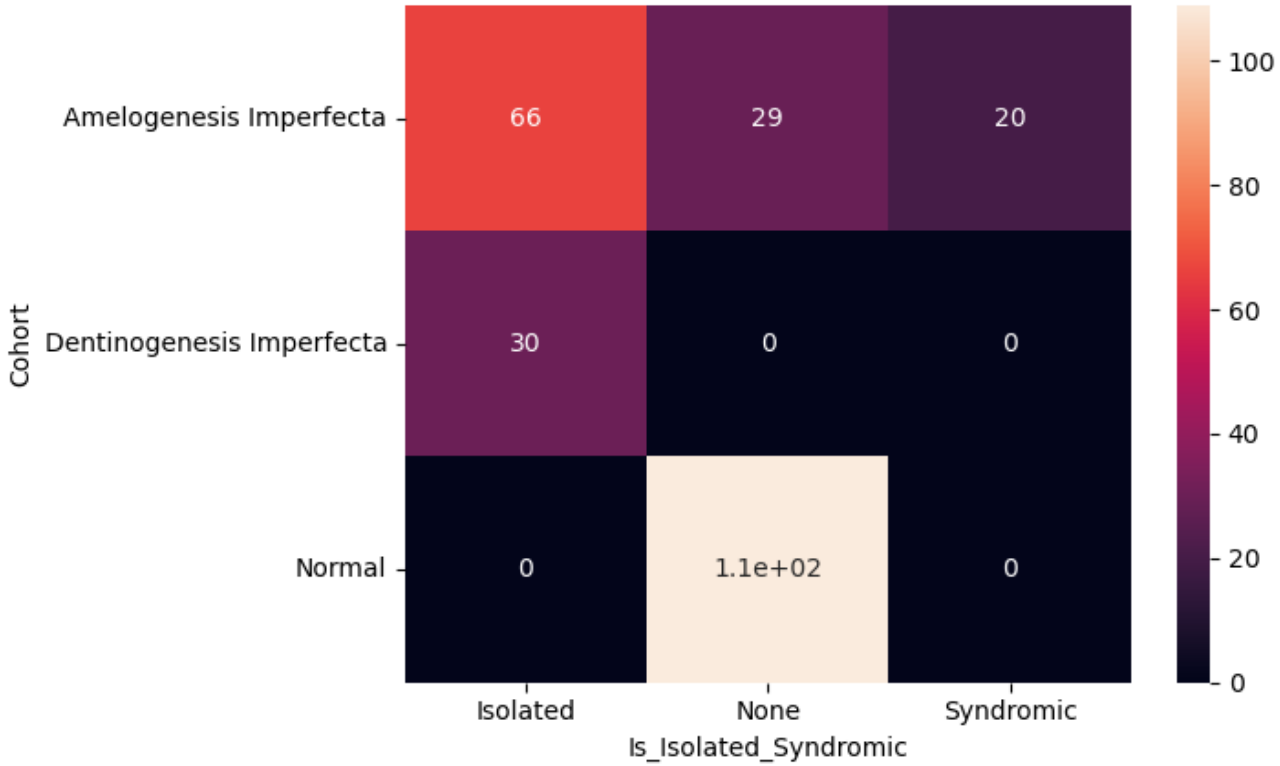
Intrabuccal images :
(Multiple photos per patient)

Radiographic images :
(Multiple photos per patient, at different age)



Données disponibles pour l’entrainement (cont.)





Données disponibles pour l’évaluation

- **110 patients** dans le jeu de données pour l’évaluation publique
- Total de **757 images**

	A	B	C	D	E
1	Patient_ID	Image_ID	Patient_Age_In_Image	Image_Type	Dentition_Cycle
2	PAT_0	IMAGE_0341.jpg	4	Panoramic	Primary
3	PAT_0	IMAGE_0468.jpg	6	Panoramic	Mixte
4	PAT_0	IMAGE_0414.jpg	7	Panoramic	Mixte

- Nombre inconnu dans le jeu de données pour l’évaluation privée

- Exemple de format attendu pour la prédiction:

	trustii_id	Patient_ID	Patient_g	Cohort	AI_Type	Responsib	Is_Isolated_Syndrom
0	1	PAT_209	M	Dentinoge	None	None	Isolated
1	2	PAT_338	F	Normal	None	None	None
2	3	PAT_236	M	Normal	None	None	None
3	4	PAT_87	M	Ameloger	AI_Hypoca	None	Isolated
4	5	PAT_278	M	Normal	None	None	None
5	6	PAT_30	F	Ameloger	AI_Hypop	FAM20A	Syndromic
6	7	PAT_250	F	Normal	None	None	None
7	8	PAT_0	M	Ameloger	AI_Hypom	MMP20	Isolated
8	9	PAT_168	M	Dentinoge	None	None	Isolated
9	10	PAT_97	F	Ameloger	AI_Hypoca	None	None
10	11	PAT_114	F	Ameloger	AI_Hypom	None	None

Exploration des données et conclusions

Pas de données manquantes

! Label '**None**' pour le gène responsable peut vouloir dire '**Pas de gènes identifié**' ou '**Pas de test fait**'

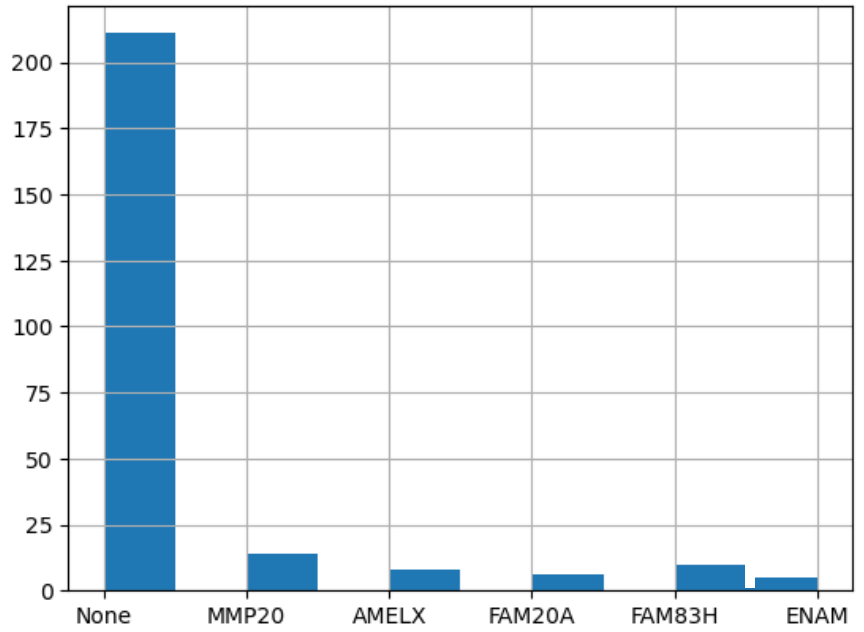
Nombres pour chaque maladie

Amelogenesis Imperfecta	115
Normal	109
Dentinogenesis Imperfecta	30

Nombres pour les types d'AI

None	139
AI_Hypomature	46
AI_Hypoplastic	46
AI_Hypocalcification/Hypomineralization	16
AI_Hypomature/AI_Hypoplastic/Taurodontism	7

Nombres pour les données génétiques



Conclusions:

- Nous devons faire de l'augmentation de données
- Nous décidons de réduire les gènes à prédire à 'MMP20, FAM83H, AMELX, FAM20A, ENAM ou NONE', au vu des faibles fréquences des autres gènes dans le jeu de données

Data processing

1. One-Hot encoding

Nous encodons les données texte sous forme numérique

2. Image preprocessing

Nous redimensionnons les images à 400X400 pixels

Nous transformons les radiographies noir et blanc en couleurs artificielles (RGB)

Nous normalisons les intensités de chaque couleur RGB [0,1]

3. Data augmentation

Nous créons des nouvelles images par augmentation, en introduisant de manière aléatoire du bruit (ColorJitter) et/ou une rotation horizontale.

4. Nous séparons les données en jeu d'entrainement et jeu de validation interne

Cohort	AI_Type	Responsible_Gene_Name	Is_Isolated_Syndromic
Normal	None	None	None
Amelogenesis Imperfecta	AI_Hypomature	MMP20	Isolated
Dentinogenesis Imperfecta	None	None	Isolated
`LabelBinarizer`	`LabelBinarizer`	`LabelBinarizer`	`LabelBinarizer`

One hot encoding

Cohort	AI_Type	Responsible_Gene_Name	Is_Isolated_Syndromic
[0,0,1]	[0,0,0,0,1]	[0,0,0,0,0,1]	[1,0,0]
[1,0,0]	[1,0,0,0,0]	[1,0,0,0,0,0]	[0,1,0]
[0,1,0]	[0,0,0,0,1]	[0,0,0,0,0,1]	[0,1,0]

Patients per Cohort without augmentation

Normal	Amelogenesis Imperfecta	Dentinogenesis Imperfecta
109	115	30

Patients per Cohort with augmentation

Normal	Amelogenesis Imperfecta	Dentinogenesis Imperfecta
109	230	60

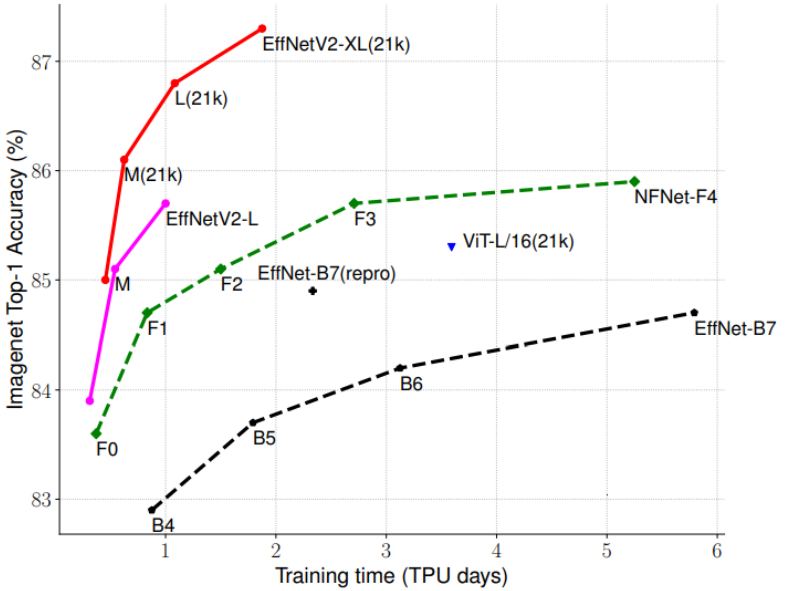
Dataset splitting	Training (N = 358)	Validation (N = 41)
Training dataset (N = 254)	90% (N = 228)	10% (N = 26)
Augmented dataset (N = 145)	80% (N = 130)	20% (N = 15)

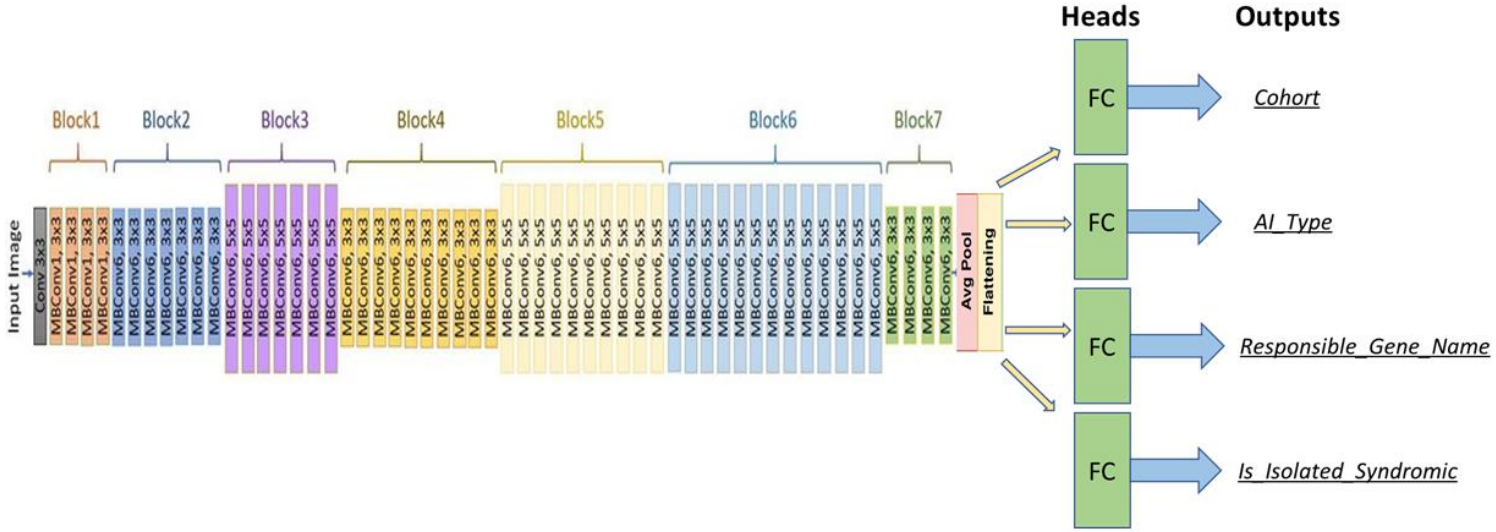
Modelling parameters

Nous avons testé deux modèles de deep learning: ResNet (He *et al*, 2015) et EfficientNet (Tan *et al, 2019)*

Nous avons choisi comme modèle final EfficientNet v2 l (Tan & Le, 2021), qui avait les meilleures performances et restait utilisable selon les contraintes imposées dans le challenge.

Par simplicité, nous avons adapté le modèle en ajoutant quatre ‘Têtes de prédiction’ totalement connectées, une par classe que nous cherchons à prédire.



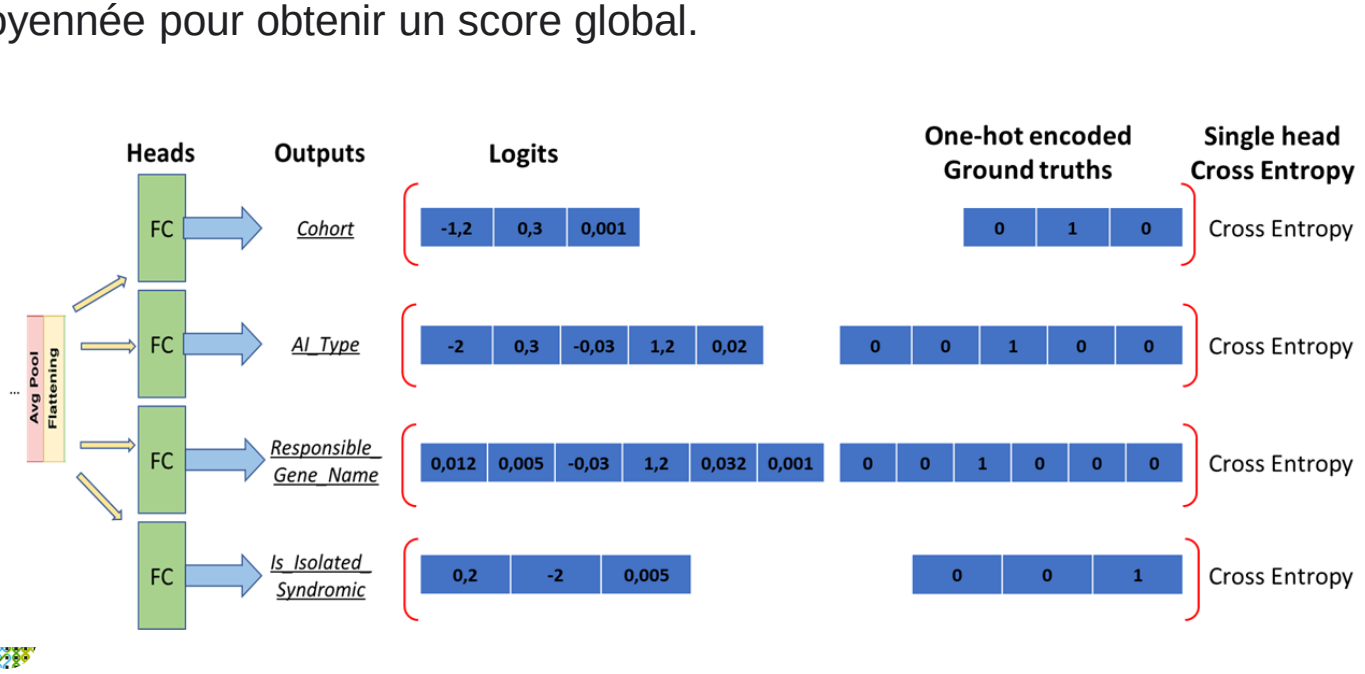


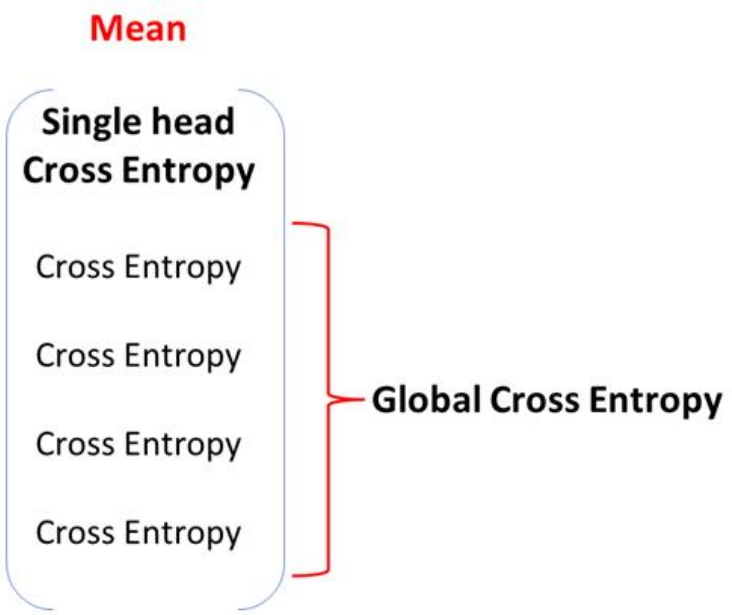
Fonction coût

La fonction coût est une part essentielle d'un modèle de deep learning. Elle permet de mesurer l'erreur d'un modèle en calculant la différence entre les prédictions d'un modèle et la vérité.

L'optimisation du modèle de deep learning passe par la minimisation de la fonction coût à travers toutes les classes.

Nous avons choisi la fonction '**cross entropy**' pour notre fonction coût, calculée sur chaque tête de prédiction, puis moyennée pour obtenir un score global.

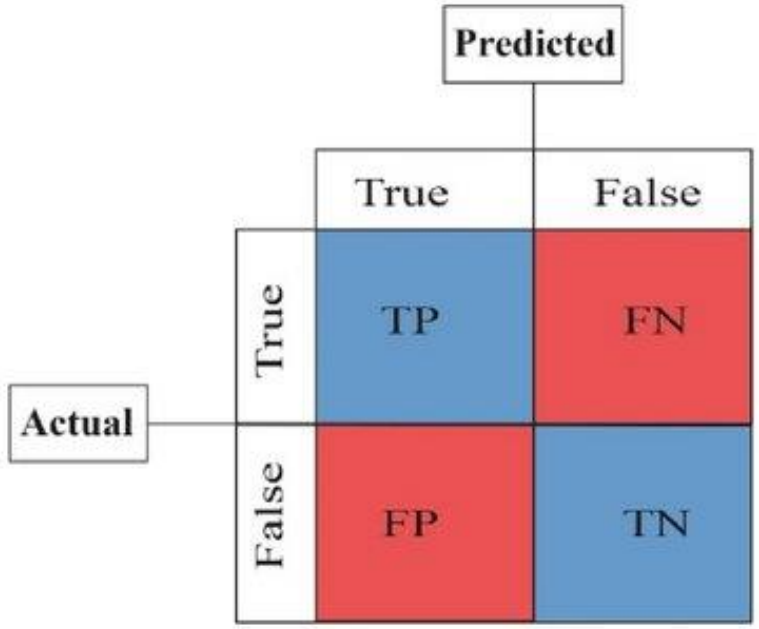




Métrique de performances

Le critère utilisé pour calculer le meilleur modèle à travers toutes les combinaisons de paramètres est le **score F1**.

Ce choix de score permet de ne pas être trop biaisé par le déséquilibre des classes dans le jeu de données (i.e. petit nombre de patients avec Taurodontisme).



$$\textbf{Precision} = \frac{TP}{TP+FP}$$

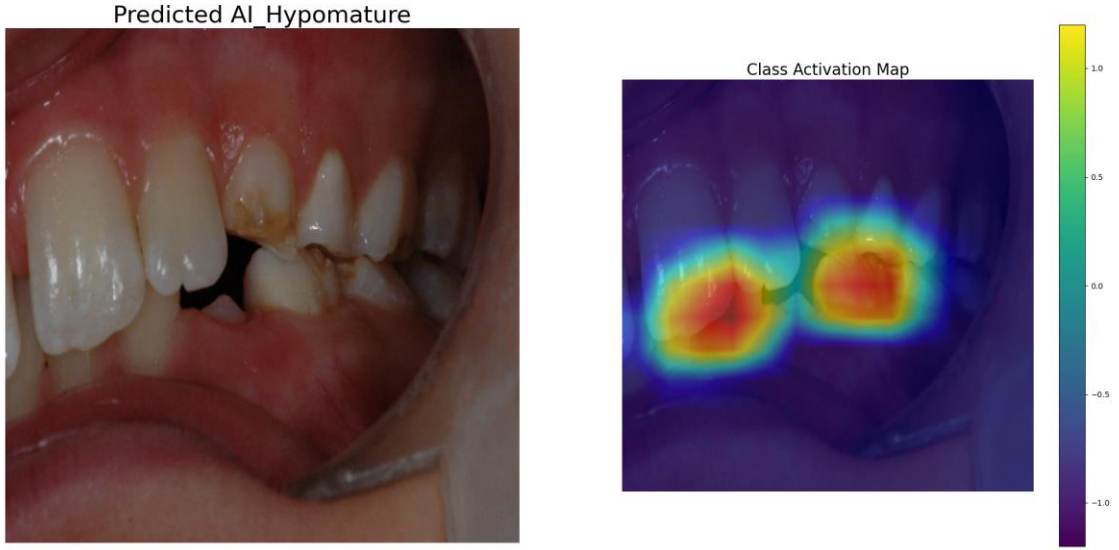
$$\textbf{Recall} = \frac{TP}{TP+FN}$$

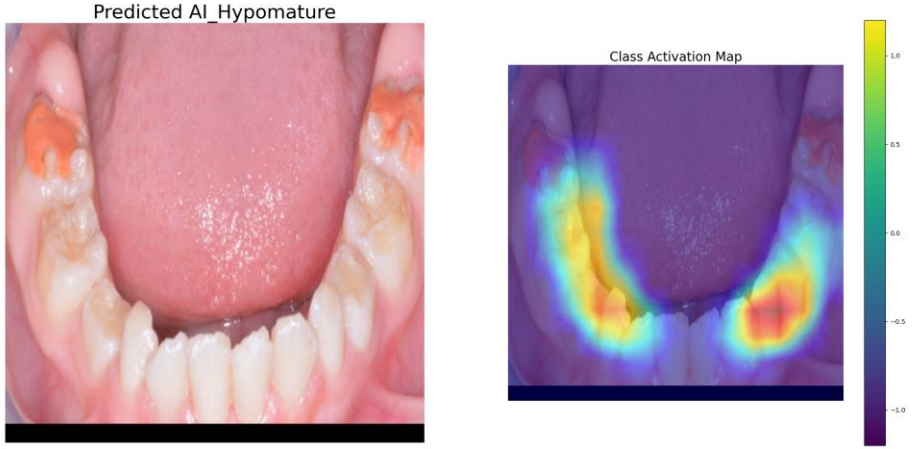
$$\textbf{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

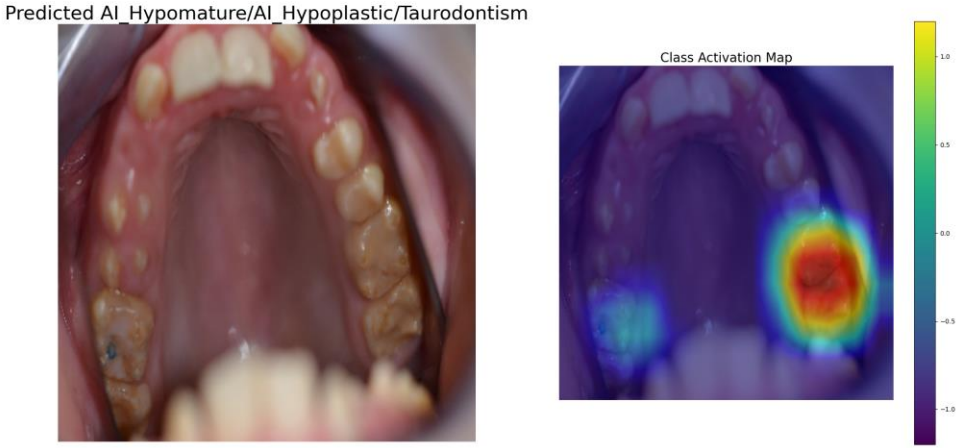
$$\textbf{F-1 Score} = 2 * \frac{precision * recall}{precision + recall}$$

Explainabilité

Pour illustrer les choix de notre modèle, nous avons utilisé la méthode CAM (Class Activation Map, Zhou *et al,* 2015). Voici un échantillon des résultats de notre modèle:



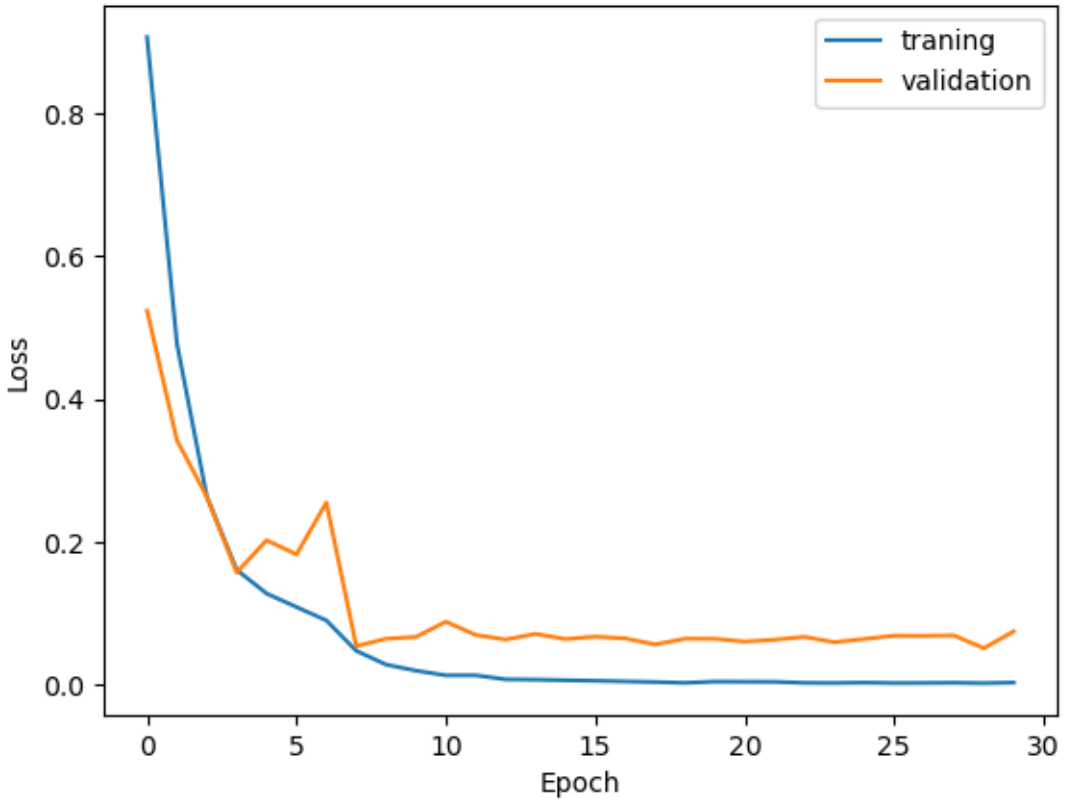




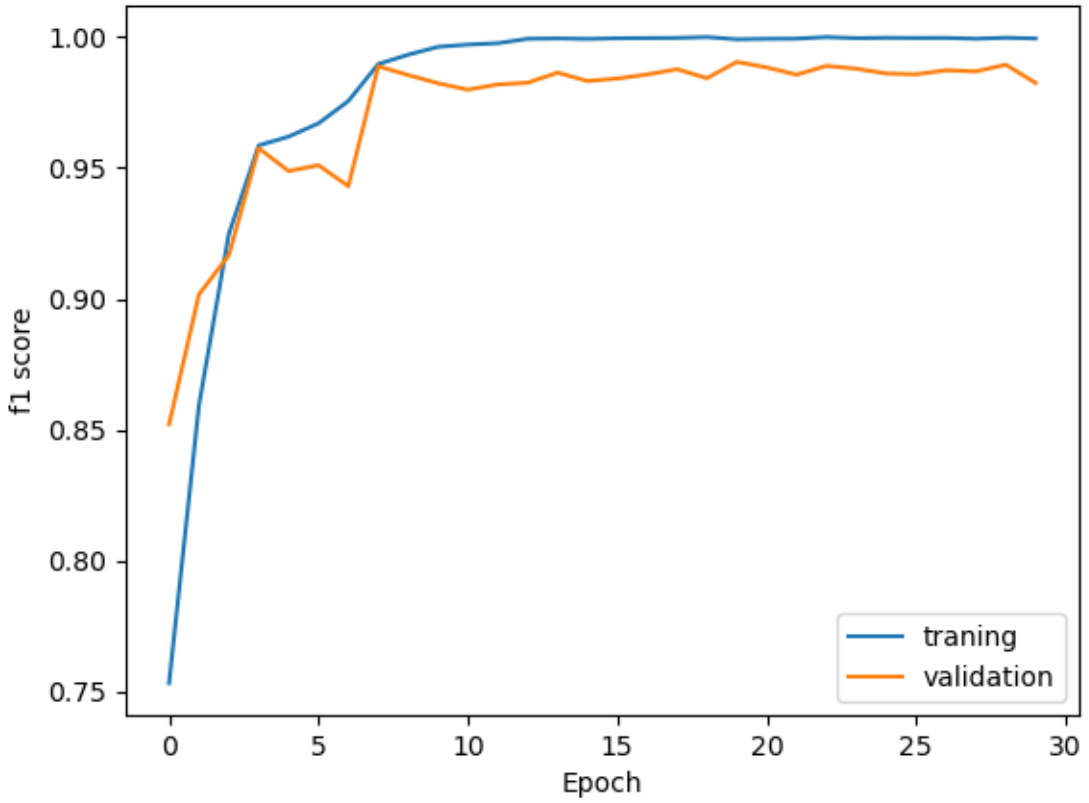
Meilleur modèle

Nous avons choisi notre meilleur modèle sur base de la cross-entropy et du score F1

Global Cross entropy



Global F1 score



Meilleur modèle

Notre choix de meilleur modèle à obtenu un score F1 de 0.72762 sur le jeu de test publique, et 0.8385 sur le jeu de test privé

Rank	Members	Team	Score	Entries	Best	Last
1	1	C Beniac	0.7523522697755766	238	06/07/2023 17:07:02	09/07/2023 21:41:34
5	5	PowerDent	0.7489196165072682	64	09/07/2023 19:24:27	09/07/2023 23:46:10
1	1	jeremy vangansberg	0.7344190503994329	143	09/07/2023 15:18:02	09/07/2023 23:10:28
4	1	Mathurin	0.7299070975369759	65	08/07/2023 06:17:36	09/07/2023 22:37:07
5	3	Association EISBM (team)	0.7276233127825735	147	09/07/2023 18:23:32	09/07/2023 18:23:40
6	2	MRIM	0.7248970534900042	53	05/07/2023 10:37:17	09/07/2023 23:48:27
7	2	Jerome Dauba & Koffi Cornelis	0.723337719758109	135	09/07/2023 22:01:57	09/07/2023 22:24:39
8	2	MontreTesDents	0.7227407890851644	151	09/07/2023 19:22:22	09/07/2023 20:00:10
9	1	_mahcih_	0.6939648449633128	5	09/07/2023 19:26:32	09/07/2023 19:26:40
10	3	SESSTIM - BB, HA, JCD, RU	0.6845381761619156	208	08/07/2023 13:49:52	09/07/2023 23:52:14
11	6	MEDICA	0.6632544568352245	55	09/07/2023 22:39:54	09/07/2023 23:34:13
12	7	Steven Delval	0.6564980365805295	28	09/05/2023 15:36:24	12/05/2023 13:37:20
13	4	Iris Godinot	0.6472870605085798	92	07/07/2023 16:13:11	07/07/2023 23:18:43
14	1	Gaoussou Sanou	0.6178059739250271	35	30/04/2023 19:42:26	05/06/2023 18:07:37
15	1	ai_mtl	0.6145751690562025	6	21/05/2023 20:13:41	24/05/2023 16:23:13

Rank	Members	Team	Score	Entries	Best	Last
1	3	Association EISBM (team)	0.8385693034368635	147	09/07/2023 18:23:32	09/07/2023 18:23:40
5	5	PowerDent	0.8309734635544753	64	09/07/2023 20:04:36	09/07/2023 23:46:10
2	2	Jerome Dauba & Koffi Cornelis	0.748608384469087	135	09/07/2023 22:15:49	09/07/2023 22:24:39
4	2	MRIM	0.7365287562146478	53	05/07/2023 10:37:17	09/07/2023 23:48:27
5	1	C Beniac	0.7351152199692845	238	09/07/2023 21:41:24	09/07/2023 21:41:34
6	1	_mahcih_	0.7350946025667399	5	09/07/2023 18:08:30	09/07/2023 19:26:40
7	2	MontreTesDents	0.7005360018813687	151	08/07/2023 17:42:26	09/07/2023 20:00:10
8	1	Mathurin	0.6808631388165259	65	09/07/2023 22:32:54	09/07/2023 22:37:07
9	1	jeremy vangansberg	0.679576611539215	143	09/07/2023 15:18:02	09/07/2023 23:10:28
10	3	SESSTIM - BB, HA, JCD, RU	0.6784511365382104	208	22/06/2023 11:42:54	09/07/2023 23:52:14
11	4	Iris Godinot	0.6744764326425681	92	29/06/2023 17:07:15	07/07/2023 23:18:43
12	7	Steven Delval	0.645403353968641	28	09/05/2023 15:36:24	12/05/2023 13:37:20
13	6	MEDICA	0.6342894286145531	55	09/07/2023 22:33:05	09/07/2023 23:34:13
14	1	Gaoussou Sanou	0.6160555585563445	35	30/04/2023 19:42:26	05/06/2023 18:07:37
15	1	ai_mtl	0.6159289721865173	6	21/05/2023 20:13:41	24/05/2023 16:23:13

Qui sommes nous ?

EISBM: European Institute for Systems Biology and Medicine, créé en 2017 à Lyon, France

Nous sommes spécialisés dans la recherche médicale et l'utilisation de technologies de pointe pour la recherche scientifique et l'amélioration de la prise en charge des maladies.

Nos activités comprennent le **management et l'analyse de données**, les **approches systémiques** et **l'intelligence artificielle**.

Johann Pellet

Directeur de l'ingénierie

Johann Pellet, MSc est depuis 2022 le directeur de l'EISBM, responsable du développement technologique et innovation. Johann possède une solide expertise en management de données de données scientifiques et médicales et en analyse de données complexes. Il possède un MSc en Bioinformatique (2005) et a plus de 15 ans d'expérience professionnel dans le secteur académique et privée en participant à des projets nationaux et européens.

Bertrand De Meulder

Directeur de la recherche

Bertrand De Meulder, PhD est le directeur scientifique, expert en bioinformatique, analyse de données omiques et biostatistiques. Il a obtenu son doctorat en sciences biologiques à l'Université de Namur (Belgique) en 2013 et à rejoint l'EISBM à sa création en 2017. Il a plus de 10 ans d'expérience dans le secteur académique en participant à des projets européens.

Albert Saporta

Chercheur

Albert Saporta, PhD. Albert est un chercheur, expert en intelligence artificielle et vision par ordinateur. Il a obtenu un doctorat en physique théorique à l'université Claude Bernard Lyon 1 en 2019 et a travaillé en tant que chercheur postdoctoral en physique médicale et intelligence artificielle générative au Centre Léon Bérard. Il a rejoint l'EISBM en 2022.

Merci de votre attention!

www.eisbm.org

Contact us: contact@eisbm.org