Préparation efficace des données d'apprentissage. Application à la classification d'images pour la détection du cancer du sein

Mouna Sabrine Mayouf, Florence Dupin de Saint-Cyr

Préparation efficace des données d'apprentissage
Application à la classification d'images pour la détection du cancer du sein

Mona Mayouf et Florence Dupin de Saint-Cyr

Université de Toulouse, IRIT-CNRS

13 avril 2021

Résumé

Mesurer "l'informativité" d'un dataset dans le cadre d'une tâche de classification est une question difficile. Dans cet article, nous essayons de circonscrire cette notion en introduisant de nouvelles mesures et en énonçant quelques principes sur la préparation des données. Nous expérimentons l'intérêt de ces mesures et la validité de ces principes en introduisant plusieurs protocoles destinés à comparer les différentes manières de préparer les données. Nous concluons en mettant en relation l'efficacité de la préparation des données et sa diversité théorique.

Mots-clef : mesures de l'information, apprentissage profond, réseaux de neurones convolutionnels, préparation des données d'apprentissage, Breakhis.

1 Introduction

Affectant une femme sur huit dans le monde, le cancer du sein est l'un des types de cancer les plus répandus auprès de la gente feminine, avec l'un des taux de mortalité les plus élevés [SOPH16]. Dans le contexte de la détection des cancers du sein, les réseaux de neurones convolutifs (CNNs) ont fait preuve d'une précision remarquable et d'une fiabilité compétitive par rapport aux méthodes traditionnelles. Cependant, selon [GBC16], les approches CNN nécessitent que le réseau soit entraîné sur une grande quantité de données. L'un des principaux problèmes est que cette quantité n'est pas toujours disponible (données manquantes, non accessibles ou trop coûteuses). L'augmentation des données a été introduite pour résoudre ce problème et est devenue l'une des meilleures pratiques pour améliorer les performances du CNN. La contribution de cette génération artificielle de données sur le processus d'apprentissage est encore mal comprise. En effet, en raison de l'aspect boîte noire du CNN, il est difficile d'identifier comment la structure des données guide l'apprentissage. L'augmentation des données est-elle efficace simplement parce qu'elle donne une redondance qui aide l'apprentissage ? Est-il nécessaire de fournir de nouvelles données issues du monde réel ou suffit-il de générer des données à partir des données dont on dispose déjà ? Comment pouvons-nous quantifier l'information contenue dans un ensemble de données pour une tâche de classification ? Quel est l'impact de la technique d'augmentation sur le processus d'apprentissage ?

L'augmentation des données est souvent vue comme une étape standard obligatoire de la préparation des données. L'augmentation traditionnelle dans le cas de données images est basée sur des transformations basiques qui génèrent des images extrêmement proches par rapport à la distribution des données initiales [VDM01]. D'autres transformations (telles que le découpage, le bruitage gaussien, le mélange, le chevauchement) s'avérent utiles pour certaines tâches de classification [SK19]. Avec le succès des réseaux adverses génératifs (GANs), des images de synthèse sont générées artificiellement [GBC16]. Cependant, dans des domaines critiques comme le domaine médical, où les transformations doivent préserver l'étiquette initiale, les transformations possibles sont restreintes. De plus, le manque de données qui rend le processus d'apprentissage infructueux peut être cumulé avec un déséquilibre du dataset, dans lequel il existe une différence flagrante dans le nombre d'échantillons pour une catégorie par rapport à une autre. Ce déséquilibre peut induire des biais d'apprentissage : [SK19].

Dans cet article, nous étudions d'abord comment estimer la quantité d'information dans un dataset en proposant plusieurs nouvelles mesures. Nous énonçons ensuite un ensemble de principes qui permettraient de régir la préparation des données (et aider à répondre

à certaines des questions introduites ci-dessus). Puis nous présentons plusieurs protocoles expérimentaux afin de vérifier la validité de notre ensemble de principes. Enfin nous les expérimentons sur le dataset BreakHis (base d'images histopathologiques du cancer du sein).

2 Contexte

2.1 Mesures classiques de l'information

Dans cette section, nous rappelons d'abord les définitions de quelques métriques présentes dans la littérature. Pour cela, nous considérons un dataset D composé de n éléments, $D = \{s_1, \ldots, s_n\}$, chaque élément s étant associé à une classe unique c qui est appelée l'étiquette de s et notée $s.label = c$. L'ensemble des classes possibles est noté $\mathcal{C}$. Classiquement, l'*abondance* $a_D(c)$ d'une classe c étant donné un dataset D est le nombre d'éléments de cette classe dans D, et l'*abondance proportionnelle* d'une classe c dans D, $P_D(c)$, est le pourcentage de représentation d'une classe parmi toutes les classes :

$$a_D(c) = |\{s \in D : s.label = c\}|$$
$$p_D(c) = \frac{a_D(c)}{n}$$

Selon [MLPFS+20], trois indicateurs ont été définis dans la littérature pour estimer la *diversité* d'un ensemble de données, à savoir la *variété*, l' *équilibre de la répartition* et la *disparité*. Nous listons ici les mesures associées à ces indicateurs :

- *Variété :* La richesse $\mathcal{R}$ est une métrique liée à la *variété* et représente le nombre de classes effectivement considérées pour la tâche de classification [Sti98, Mac65] :

$$\mathcal{R}(D) = |\{c \in C : p_D(c) > 0\}|$$

- *Équilibre de la répartition :* Le rapport de déséquilibre $\mathcal{IR}$ est le rapport entre la classe majoritaire et la classe minoritaire (dans une classification binaire) [OPBM06] :

$$\mathcal{IR}(D) = \frac{a_D(\text{Classe majoritaire})}{a_D(\text{Classe minoritaire})}$$

  Selon [OPBM09], le dataset est peu déséquilibré lorsque $1,5 < \mathcal{IR} < 3$, moyennement déséquilibré pour $3 < \mathcal{IR} < 9$ et très déséquilibré lorsque $\mathcal{IR} > 9$. Notons qu'il existe plusieurs autres mesures qui caractérisent la répartition des données (voir [MLPFS+20]). Cependant, étant donné qu'elles sont toutes basées sur l'abondance proportionnelle p, cela signifie qu'elles ne prennent en compte que le nombre d'éléments par classe sans tenir compte des différentes natures de ces éléments.

- *Disparité :* La disparité $\mathcal{D}$ quantifie la variété des données en fonction d'une distance par paire d entre les classes.

$$\mathcal{D} = \sum_{c \in C} \sum_{c' \in C} d(c, c')$$

  Cependant, fournir la distance d entre deux classes nécessite des connaissances supplémentaires (provenant par exemple du contexte de la tâche de classification).

Puisque la variété et l'équilibre ne sont basées que sur l'abondance, les métriques associées nous aideront peu à caractériser la quantité d'information. C'est pourquoi dans la section 4, nous proposons d'introduire plusieurs nouvelles métriques basées soit sur la Disparité soit sur le diamètre du dataset, en y intégrant une distance d plus appropriée aux images.

2.2 Le dataset BreakHis

"BreakHis" pour "Breast Cancer Histopathological Database") est un dataset public composé de 7909 images de biopsies histopathologiques de tumeurs du sein observées selon quatre grossissements microscopiques : 40X, 100X, 200X et 400X [SOPH16]. Parmi les étiquettes qui caractérisent les images, nous nous concentrons sur le type de tumeur qui est soit bénin, soit malin. Ce dataset est déséquilibré avec $\mathcal{IR} = 2.19$ (voir Tableau 1). Pour faire face à ce déséquilibre, nous proposons plusieurs techniques d'augmentation des données et comparons leur impact sur le processus d'apprentissage.

3 Préparation des données

L'étape classique de préparation des données consiste en un rééquilibrage et une augmentation. Notons que selon [SK19], il existe trois façons de rééquilibrer : 1) *sur-échantillonner la classe minoritaire*, ce qui revient à augmenter sa taille ; 2) *sous-échantillonner la classe majoritaire*, c'est-à-dire en retirer des éléments ; 3) *choisir de façon ré-équilibrante* en augmentant la chance de sélectionner un élément dans la catégorie marginalisée. (Cette dernière technique n'entre pas dans le cadre de cet article).

Cette section présente d'abord une liste de principes qui devraient être respectés lors de toute préparation de données, puis nous introduisons le formalisme

adopté et la signature des fonctions de transformation que nous allons utiliser pour l'augmentation des données. Afin de valider les principes que nous présentons ici, nous avons conçu dans la section 3.4 un ensemble discriminant de protocoles expérimentaux dont les résultats permettent de confirmer ou d'infirmer ces hypothèses.

3.1 Principes généraux pour une préparation efficace des données

Les pratiques de préparation des données les plus utilisées sont celles qui ont obtenu les meilleurs résultats pour la classification, certaines pratiques sont connues pour fonctionner mieux que d'autres, mais les caractéristiques de ces bonnes pratiques de préparation ne sont pas toujours explicites. De plus, il n'est pas évident de savoir si certaines pratiques sont bonnes ou pas, par exemple, l'augmentation crée parfois des doublons de certains échantillons, est-ce efficace de le faire ? Ci-dessous, nous proposons une liste de principes inspirés des usages courants afin de donner une meilleure idée de ce que devrait être une "préparation rationnelle des données". Certains de ces principes sont bien connus, et semblent évidents, mais en les explicitant, nous montrons que plus d'expériences sont nécessaires pour les valider. Cela souligne également le besoin de métriques qui pourraient mieux caractériser les datasets, justifiant ainsi le travail effectué dans la section 4.

- *Dataset équilibré* (BD) (pour *Balanced Dataset*)
- *Taille suffisante de l'ensemble de données* (SDS) (pour Sufficient Dataset Size)
- *Pas de doublons* (NDI) (pour No Duplication of Items)
- *Opérateurs de transformation bien choisis* (WCTO) (pour Well Chosen Transformation Operators)
- *Operateurs de transformation variés* (VTO) (pour Variety of Transformation Operators)
- *Ajout de Données Externes* (FED) (pour Fresh External Data)

En effet, nos expériences présentées dans la section 5 confirment ces principes rationnels. Plus précisément, (BD) stipule qu'un dataset avec une répartition équilibrée se comporte mieux pour une tâche de classification. (SDS) implique qu'un trop petit ensemble de données peut avoir un impact négatif (inefficacité et convergence lente) sur le processus d'entrainement, même lorsque les données sont équilibrées. De plus, par (NDI), nous postulons que la duplication à l'identique ne permet pas de compenser le manque de données (elle n'améliore ni l'efficacité ni la convergence). En outre, l'utilisation de l'augmentation de données sans données "fraîches" mais avec des éléments transformés a souvent un impact positif sur le processus d'apprentissage, en particulier lorsque la transformation est "conservatrice de l'étiquette", ce qui est le sens de (WCTO). (WCTO) est aussi utile pour équilibrer un dataset de taille suffisante puisqu'en ajoutant des éléments transformés à bon escient à la classe minoritaire on peut obtenir un impact positif sur l'apprentissage. (VTO) exprime que l'utilisation de différents opérateurs de transformation a un impact positif. Enfin, l'ajout de données réelles externes (FED) est plus performant que l'ajout de données générées mais peut nécessiter plus de temps d'apprentissage.

3.2 Transformation d'images BreakHis

En plus de l'opérateur d'identité noté *Id*, le processus d'augmentation des données est basé sur deux types d'opérateurs *élémentaires* qui sont conservateurs d'étiquettes (voir [AAC+17] pour les opérateurs géométriques et [TLB+19] pour les opérateurs de couleur) :

- *Opérateurs géométriques :* Étant donné que les images BreakHis sont des rectangles de 460x700, toute opération géométrique non-miroir produirait une forme différente qui devrait être remodelée ou recadrée afin d'alimenter le CNN. Pour éviter cette opération supplémentaire qui pourrait diminuer la précision, nous optons pour les deux seuls opérateurs qui préservent la forme rectangulaire : les renversements horizontaux et verticaux. Ces deux opérateurs sont notés respectivement **H** et **V**.
- *Opérateurs de couleur :* Afin d'augmenter le nombre d'images, nous considérons également la possibilité de jouer sur les couleurs. Nous utilisons deux opérateurs : une inversion de couleur RGB et une transformation de l'encodage RGB de l'image en encodage de couleur HSV. Ils sont notés respectivement **c** et **C**.

Afin d'effectuer plus de quatre augmentations distinctes, il est nécessaire de combiner les opérateurs élémentaires en les appliquant séquentiellement. Cependant, certaines combinaisons pourraient créer des doublons des mêmes images (par exemple, HV=VH, Hc=cH, ...). Finalement, en raison des symétries, seules 15 combinaisons distinctes sont possibles, à savoir : H, V, c, C, HV, Hc, HC, Vc, VC, cC, HVc, HVC, HcC, VcC, HVcC.

3.3 Signature d'une transformation

Dans cette section, nous introduisons la signature d'un type particulier de transformation de données pour laquelle les données sont divisées en parties égales d'échantillons et où la même transformation est appliquée à tous les éléments de chaque partie.

Définition 1 (Signature)**.** *La signature d'une* transformation équitablement répartie *d'un ensemble de données, désignée par $tr(D, ops, ratio)$, est une fonction des paramètres suivants :*

- *D : le dataset à transformer*
- *ops : la liste des opérateurs à appliquer aux différentes parties*
- *r : le ratio de division du dataset en parties (sur lesquelles les opérateurs s'appliqueront)*

où $tr(D, (op_1, \dots op_p), r) = (op_1(D_1), \dots, op_p(D_p))$ est une partition de l'ensemble de données D en $D_1, \dots D_p$ (où $p = 100/r$) sur lesquels les opérateurs $(op_1, \dots op_p)$ sont appliqués respectivement et $op(D)$ est une abréviation pour : $op(D) = \{op(s)|s \in D\}$

Par exemple, nous pouvons considérer l'augmentation réalisée en appliquant une opération élémentaire parmi (H, V, c, C) à chaque 25% de l'ensemble de données D : cette augmentation a la signature $tr(D, (H, V, c, C), 25)$. Elle consiste à partitionner D en quatre parties (D_1, D_2, D_3, D_4) et à appliquer H à D_1, V à D_2, c à D_3 et C à D_4 pour obtenir un nouveau dataset $D' = tr(D, (H, V, c, C), 25) = (H(D_1), V(D_2), c(D_3), C(D_4))$.

Notons qu'une augmentation qui applique le même opérateur op à l'ensemble des données D a la signature suivante $tr(D, (op), 100)$.

3.4 Protocoles expérimentaux

Dans cette section, nous proposons 13 protocoles différents de préparation des données pour le dataset BreakHis, conçus dans le but de nous permettre de valider les principes généraux énoncés ci-dessus. D désigne la partie des éléments du BreakHis affectée à l'entraînement (nous avons pris 2/3 du dataset initial) et D_i désigne le nouveau dataset d'entraînement après préparation avec le protocole P_i. Dans ce qui suit, tous les échantillons s' de D_i sont tels que $s'.label = s.label$ où s est l'échantillon original dans D qui donne s' dans D_i (par une transformation appartenant à $\{Id, H, V, C, c, HC, Vc, CV, cH\}$). En d'autres termes, les protocoles créent de nouveaux échantillons qui sont étiquetés en fonction de leur étiquette initiale dans le dataset original.

Le dataset d'entraînement de BreakHis D est composé de deux classes, la classe minoritaire appelée m, c'est la catégorie bénigne : $m = \{s \in D|s.label = benign\}$. La classe majoritaire notée M est la catégorie maligne : $M = \{s \in D|s.label = malignant\}$. Donc $D = M \cup m$. Notons que dans BreakHis m a une taille égale à la moitié de la taille de la classe majoritaire M, les protocoles utilisent cette caractéristique pour équilibrer les données.

1. Protocoles 1 (le dataset brut) : P_{1a} est un protocole témoin dans lequel aucun équilibrage ni aucune augmentation ne sont effectués. $D_{1a} = D$; P_{1b} est un second protocole de contrôle où seule une augmentation est effectuée, sans apporter de "nouvelles" informations, par une simple duplication à l'identique des items de la classe majoritaire (donc en augmentant le déséquilibre) : $D_{1b} = D \cup tr(M, (Id), 100)$; P_{1c} est un troisième protocole témoin n'apporte toujours aucune "nouvelle" information mais augmentant la taille du dataset par simple duplication des items afin d'équilibrer et d'augmenter les données $D_{1c} = D \cup m \cup tr(D \cup m, (Id), 100)$.
2. Protocoles 2 (données équilibrées) : P_{2a} et P_{2b} doublent la taille de la classe minoritaire avec un seul opérateur. P_{2a} utilise un opérateur géométrique : $D_{2a} = D \cup tr(m, (H), 100)$; P_{2b} utilise un opérateur couleur : $D_{2b} = D \cup tr(m, (C), 100)$. P_{2c} : équilibre par sous-échantillonnage. $D_{2c} = m \cup Sample(M, |m|)$ où $Sample(X, n)$ est une fonction qui sélectionne aléatoirement n éléments parmi l'ensemble X ;
3. Protocole P_3 (données déséquilibrées, mais de taille augmentée) utilise un opérateur de couleur pour augmenter la taille de la classe majoritaire : $D_3 = D \cup tr(M/2, (C), 100)$.
4. Protocoles 4 (données équilibrées et augmentées) : avec deux opérateurs successifs distincts. P_{4a} utilise les opérateurs géométriques H et V : $m' = m \cup tr(m, (H), 100)$ (double la taille de la minorité), $D_{4a} = M \cup m' \cup tr(M \cup m', (V), 100)$ (augmente l'ensemble des données) ; P_{4b} est similaire à P_{4a} mais utilise les opérateurs de couleur C et c ; P_{4c} utilise les opérateurs H et C ; P_{4d} utilise les opérateurs V et c. P_{4e} utilise les quatre opérateurs appliqués sur différentes parties de l'ensemble de données : $m' = m \cup tr(m, (H, V, C, c), 25)$ (doublement de la taille de la minorité), $D_{4e} = M \cup m' \cup tr(M \cup m', (C, c, V, H), 25)$ (augmentation de la taille de la minorité), $D_{4e} = M \cup m \cup tr(m, (H, V, C, c), 25)$

$\cup$ $tr(M,(C,c,V,H),25)$ $\cup$ $tr(m,(C,c,V,H),25)$ $\cup$ $tr(m,(HC,Vc,CV,cH),25)$. P_{4f} comble le manque de données en ajoutant des échantillons provenant d'un autre dataset externe[1] : $D_{4f} = M \cup m \cup m\_extra \cup M\_extra$ où m_extra (resp. M_extra) est un ensemble de $3|m|$ (resp.$|M|$) images de la catégorie minoritaire (resp. majoritaire) de l'autre dataset.

Dans ce qui suit, le protocole i est abrégé P_i, la variante j d'un protocole i est abrégée P_{ij} et le dataset obtenu à partir de D par P_{ij} est abrégé D_{ij}.

4 Mesures de diversité

D'après les définitions rappelées dans la section 2.1, pour calculer la *disparité* $\mathcal{D}$ d'un dataset, nous sommes en mesure de fournir un moyen de calculer la distance entre les différentes classes. Nous proposons de définir la distance entre deux classes en introduisant d'abord la distance entre deux images. Ensuite, nous définissons la distance entre deux classes par la distance entre les images moyennes de chaque classe. Il existe plusieurs façons de calculer la distance entre deux images, par exemple la distance euclidienne est basée sur une comparaison point à point des pixels de chaque image (c'est la norme de la différence matricielle).

Un autre aspect est de prendre en compte des informations supplémentaires afin d'intégrer dans la distance le fait que les symétries horizontales et verticales ne doivent pas augmenter la distance entre les images, car pour une tâche de classification d'images histopathologiques, ces symétries n'ont pas d'importance. C'est pourquoi nous choisissons d'utiliser une mesure standard appelée SSIM (structural similarity index measure) [Wan03] qui estime la similarité de deux images en se basant sur une sorte de contraction des images en fonction de leur luminance, leur contraste et leur structure.

Définition 2 (SSI [Wan03])**.** *s_1 et s_2 étant deux items,*

$$SSI(s_1,s_2) = \frac{(2\mu_1\mu_2+\alpha_1)(2\sigma_{12}+\alpha_2)}{(\mu_1^2+\mu_2^2+\alpha_1)(\sigma_1^2+\sigma_2^2+\alpha_2)}$$

. où μ_1, μ_2, σ_1^2, σ_2^2, σ_{12}, α_1, α_2 sont respectivement les moyennes et la variance de $s_1.image$ et de $s_2.image$, la covariance de $s_1.image$ et de $s_2.image$, et deux constantes[2]*.*

1. `https://iciar2018-challenge.grand-challenge.org/Dataset/`

2. Ces constantes ont été introduites par [Zho04] pour éviter l'instabilité lorsque le dénominateur est proche de 0 en fixant $\alpha_1 = 0,01 \times L$ et $\alpha_2 = 0,03 \times L$ où L est la plage dynamique des valeurs de pixel. Pour BreakHis, les images sont codées avec 8 bits/pixel donc $L = 255$.

Notons que cette mesure de similarité est invariante par rapport aux flips verticaux et horizontaux, puisqu'une image et son flip ont la même moyenne et la même variance. Cependant, ceci n'est pas valable pour les opérations sur les couleurs. Nous proposons donc une variante de *SSI*, appelée *SSIC*, qui conserve les étiquettes pour toutes les opérations présentées en section 2.2, c'est-à-dire invariante pour les opérations de couleur **c** et **C**.

Définition 3. *Si s_1, s_2 étant deux items,*

$$SSIC(s_1,s_2) = \min_{op\in\{Id,c,C\}} SSI(s_1, op^{-1}(s_2))$$

.

Remarque 1. *Si s est un échantillon à comparer à son échantillon transformé par inversion de couleur RGB : $c(s)$, alors la distance $SSIC(s,c(s))$ est de 0 puisque $c^{-1}(c(s))= s$.*

Proposition 1. *Pour toute combinaison cop des opérateurs élémentaires $\{Id, H, V, C, c\}$, il s'avère que pour tout échantillon s, $SSIC(s,cop(s)) = 0$.*

Démonstration. La preuve concernant les opérateurs géométriques est due à la définition de *SSI* 2. *SSIC* étant construit sur *SSI* pour ignorer les transformations de couleur d'où le résultat. □

Nous sommes maintenant en mesure de définir l'item le plus représentatif parmi un ensemble X, appelé $\mu_I(X)$. C'est l'item qui est le plus similaire aux autres items de X :

$$\mu_I(X) = \text{argmax}_{x\in X} \sum_{y\in X\setminus\{x\}} SSIC(x,y)$$

Nous proposons d'évaluer la diversité d'un ensemble de données D par sa disparité et son diamètre. La disparité a déjà été rappelée ci-dessus et est liée à la distinction entre les différentes classes. Le diamètre est une mesure générale de l'étendue de l'ensemble indépendamment des classes, c'est la distance maximale entre deux items quelconques du dataset. Ces deux mesures peuvent être définies soit sur la distance euclidienne d, soit sur la mesure de similarité plus informée *SSIC*, ce qui donne quatre mesures $diam$, $disp$, $diam_I$ et $disp_I$ où I représente la "mesure informée". Nous normalisons la distance euclidienne par la matrice de distance la plus grande possible notée $|im255|$, c'est-à-dire l'image composée de 255 sur les trois canaux RGB (vu que les images sont de taille 460×700, d'où $|im_{255}| = \sqrt{460 \times 700 \times 3 \times 255^2} = 250627.5125$).

Définition 4. *Etant donné un dataset D, à deux classes D_1 et D_2 ($D = D_1 \cup D_2$),*

- $diam(D) = \frac{\max_{s_1,s_2 \in D} d(s_1.image, s_2.image)}{|im_{255}|}$
- $disp(D) = \frac{d(\mu(D_1),\mu(D_2))}{|im_{255}|}$
- $diam_I(D) = \max_{s_1,s_2 \in D}(1 - SSIC(s_1.image, s_2.image))$
- $disp_I(D) = (1 - SSIC(\mu_I(D_1), \mu_I(D_2))$

Notons qu'en ce qui concerne les disparités, les définitions sont données pour une classification binaire ($|\mathcal{C}| = 2$) où D_1 est la partie de l'ensemble D contenant la première classe et D_2 est la partie de l'ensemble contenant la seconde classe [3].

La proposition suivante montre que tous les protocoles que nous fournissons, à l'exception de P_{4f}, n'apportent aucune information "nouvelle" au dataset.

Proposition 2. *Pour tous les datasets D_{ij} obtenus par les protocoles sauf D_{4f} :*

$$diam_I(D_{ij}) = diam_I(D_{1a})$$
$$disp_I(D_{ij}) = disp_I(D_{1a})$$

Démonstration. La preuve est basée sur l'invariance de *SSIC* vis à vis des opérateurs de couleur et de géométrie. □

5 Résultats et discussion

Dans cette partie, nous essayons d'estimer la quantité d'information contenue dans les différents datasets obtenus par les protocoles précédents. Comme indiqué dans la section 2.1, la variété et l'équilibre de la répartition peuvent être estimés respectivement par la mesure de richesse $\mathcal{R}$ et le ratio de déséquilibre $\mathcal{IR}$. Cette section évalue les différents protocoles sur deux aspects, d'abord la quantité d'information présente dans le dataset produit par le protocole, ensuite la précision de la classification obtenue par un CNN entraîné sur ces jeux de données. Le tableau 1 donne les différentes tailles D_{ij} des datasets obtenus par les différents protocoles P_{ij}. Notons que la *richesse* du dataset obtenu avec l'un ou l'autre des protocoles reste la même puisque le nombre de classes reste constant : $\mathcal{R}(D_{ij}) = 2$ pour tous les protocoles (deux classes : bénigne et maligne), car aucun des protocoles proposés n'ajoute un élément d'une nouvelle classe ou ne supprime tous les éléments d'une classe existante.

3. Si il y avait plus de deux classes, la disparité serait $\frac{2\sum_{c \in C}\sum_{c' \in C \setminus \{c\}} d(\mu(D_c),\mu(D_{c'}))}{|C| \times (|C|-1) \times |im_{255}|}$

En ce qui concerne l'*équilibre de la répartition*, le rapport de déséquilibre $\mathcal{IR}$ des ensembles de données D_{ij} obtenus par les différents protocoles est toujours de 1 (en raison du doublement de la taille de la classe minoritaire qui a une taille égale à la moitié de celle de la classe majoritaire), pour tout protocole P_{ij} sauf P_{1a}, P_{1b} et P_3. Notons qu'en raison de la proposition 2, les disparités $disp_I$ et les diamètres informés $diam_I$ sont les mêmes pour tous les datasets sauf D_{4f}.

P	$\lvert D_{ij}\rvert$	$\mathcal{R}$	$\mathcal{IR}$	$disp$	$diam$	$disp_I$	$diam_I$
1a	5271	2	2,19	0.0254	0.1299	0.0975	0.0157
1b	8785	2	4,38	0.0254	0.1299	0.0975	0.0157
1c	14056	2	1	0.0254	0.1299	0.0975	0.0157
2a	7028	2	1	0.0528	0.1299	0.0975	0.0157
2b	7028	2	1	0.0826	0.2453	0.0975	0.0157
2c	3514	2	1	0.0265	0.1045	0.0975	0.0157
3	7028	2	3,28	0.0654	0.4213	0.0975	0.0157
4a	14056	2	1	0.038	0.1465	0.0975	0.0157
4b	14056	2	1	0.1168	0.5812	0.0975	0.0157
4c	14056	2	1	0.2051	0.3489	0.0975	0.0157
4d	14056	2	1	0.1030	0.4731	0.0975	0.0157
4e	14056	2	1	0.4361	0.8312	0.0975	0.0157
4f	14056	2	1	0.4673	0.4369	0.1385	0.2413

Tableau 1 – Resulats des metriques d'informativité

Le tableau 2 décrit les résultats obtenus par le réseau entraîné sur les jeux de données produits par les différents protocoles. *Acc* représente le taux d'exactitude du réseau, Acc est le taux d'items correctement classés parmi tous les items :

$$Acc = \frac{TM + TB}{TM + TB + FM + FB}$$

avec TB (respectivement TM, FB, FM) dénote le nombre d'items correctement affectés à la catégorie bénigne (resp. correctement à la catégorie maligne, mal étiquetée bénigne, mal étiquetée maligne). *Prec* est la précision, elle indique la proportion d'éléments correctement attribués parmi ceux qui sont prédits comme étant malin :

$$Prec = \frac{TM}{TM + FM}$$

Rec est le rappel, il indique la proportion d'échantillons correctement affectés à la classe maligne parmi tous les échantillons qui sont malins dans la vérité terrain :

$$Rec = \frac{TM}{TM + FB}$$

Nous donnons également une indication du comportement de l'apprentissage en mentionnant l'époque de

stabilisation *StbE* qui est calculée grâce à la technique de régularisation par arrêt précoce (*early stopping*) [Pre98]. Il s'agit de l'époque à partir de laquelle l'erreur de l'apprentissage est presque constante.

P	StbE	Acc (%)	Prec (%)	Rec (%)
1a	inf	47.23	53.22	48.59
1b	inf	49.08	46.59	49.03
1c	inf	50.01	48.23	47.71
2a	1966	64.12	65.27	67.02
2b	2133	69.43	66.15	68.13
2c	inf	50.03	46.02	49.93
3	inf	55.79	52.03	56.46
4a	2146	88.63	75.10	70.02
4b	2369	85.36	71.36	69.04
4c	1967	90.02	85.03	88.52
4d	2513	84.29	72.13	78.96
4e	2719	95.63	78.49	75.16
4f	2861	96.03	89.46	91.75

Tableau 2 – Résultats de classification

Le tableau 3 montre les principes qui semblent confirmer la performance du protocole. Pour WCTO et VTO nous précisons respectivement la liste des opérateurs et le nombre d'opérateurs distincts utilisés.

P	BD	SDS	NDI	(ops)	(nb ops)	FED
1a	no	no	yes	no	0	no
1b	no	no	no	no	0	no
1c	yes	yes	no	no	0	no
2a	yes	no	yes	H	1	no
2b	yes	no	yes	C	1	no
2c	yes	no	yes	no	0	no
3	no	no	yes	C	1	no
4a	yes	yes	yes	V + VH	2	no
4b	yes	yes	yes	c + cC	2	no
4c	yes	yes	yes	C + CH	2	no
4d	yes	yes	yes	c + cV	2	no
4e	yes	yes	yes	C+c+V+ H+CH+cV +VC+Hc	8	no
4f	yes	yes	yes	no	0	yes

Tableau 3 – Principes suivis par les protocoles

Dans le tableau 2, les mauvais résultats de P_{1a} et P_{2c} soulignent qu'*un trop petit ensemble de données a un fort impact négatif sur le processus d'apprentissage, même lorsque les données sont équilibrées* confirmant le principe (SDS). De plus, ces deux datasets avaient la plus petite disparité et les plus petits diamètres (absolus et informés). Une petite disparité signifie que les images des deux classes sont proches l'une de l'autre, ce qui rend la tâche de discrimination plus difficile.

La duplication ne compense pas l'insuffisance du dataset. De plus, compenser le manque de données en dupliquant à l'identique les mêmes images rend l'apprentissage encore plus difficile et fait basculer le CNN dans le sur-apprentissage (P_{1b} et P_{1c}), car pour ces derniers protocoles le CNN est instable et bloqué dans un régime transitoire avec une précision inférieure à 50%, confirmant le principe (NDI).

L'utilisation de l'augmentation des données sans nouvelles données extérieures mais avec des éléments transformés a un impact positif sur l'apprentissage puisque P_{4a} donne de meilleurs résultats que P_{1c} et que P_{2b} est meilleur que P_{2c}. Ceci confirme les principes (WCTO) et (BD).

L'équilibrage d'un ensemble de données de taille suffisante en ajoutant des éléments transformés à la classe minoritaire a un impact positif. En outre, on peut vérifier qu'*augmenter un ensemble de données équilibré accroit les performances* (voir $P_4abcdef$ par rapport à P_{2abc}) ce qui soutient à nouveau les principes (SDS) et (BD). Notons que les transformations de couleurs ont un meilleur impact que les transformations géométriques (P_{2b} étant meilleur que P_{2a} et P_{4c} que P_{4a}) ce qui confirme le principe (WCTO). En parallèle, nous observons que la disparité et le diamètre sont augmentés par l'ajout d'échantillons transformés, ce qui relie ces mesures aux principes (WCTO) et (VTO). De plus, nous concluons que le fait de varier les opérateurs en les utilisant sur différentes parties de l'ensemble de données augmente la précision : la meilleure précision 95.63% est obtenue dans ce cas (P_{4e}) avec l'utilisation de 8 opérateurs différents démontrant l'importance du principe (VTO) qui est à nouveau corrélé avec une disparité et un diamètre élevés.

Enfin, nous voyons que P_{4f} a les meilleures performances avec l'ajout de données externes fraîches (DEF) mais ce protocole nécessite plus de temps d'apprentissage. Il est clair qu'avoir la possibilité d'ajouter des données externes est idéal mais pas toujours réalisable, c'est pourquoi nous pouvons considérer que P_{4e} et P_{4c} sont les meilleures préparations de données.

Contrairement à ce qu'on aurait pu attendre, plusieurs ensembles de données qui ont des valeurs égales avec $disp_I$ et $diam_I$ peuvent avoir une efficacité très différente. Ces mesures capturent une sorte de richesse brute similaire à celle qu'un expert humain aurait pu donner en comprenant les équivalences entre les échantillons. Il semble que le réseau bénéficie de la création d'échantillons équivalents qui n'augmentent pourtant pas ce que nous appelons la disparité et le

diamètre “informés” mais augmentent la disparité et le diamètre non informés.

6 Détails techniques

Nous avons utilisé le CNN pré-entraîné VGG19 comme classificateur pour comparer les différents protocoles. Afin d'optimiser l'entraînement du réseau, nous avons utilisé plusieurs techniques de régularisation telles que la régularisation *L2* avec $\alpha = 0,01$, les techniques d'arrêt précoce (*early stopping*) et de décrochage (*dropout*).

Le modèle a été implémenté sur la bibliothèque Keras. Nous avons entraîné notre modèle avec 3000 époques. Nous avons opté pour l'optimisation Adam avec un taux d'apprentissage fixé initialement à 0,0001.

Le dataset initial a été divisé en 2/3 pour l'ensemble d'apprentissage $D = D_{1a}$, 1/6 pour la validation et 1/6 pour le test. $|D_{1a}| = 5271, p(M) = 2/3, p(B) = 1/3$.

7 Conclusion

Cet article fournit une étude du processus de préparation des données à travers l'idée qu'il est nécessaire d'évaluer la quantité d'information présente dans un dataset en termes d'efficacité relativement à la tâche de classification à effectuer. Dans ce but, nous avons défini quatre nouvelles métriques pour évaluer la diversité du dataset et nous avons formalisé six principes rationnels pour la préparation des données. Ensuite, nous avons expérimenté 13 protocoles de préparation de données et identifié parmi eux les plus appropriés pour la classification des images BreakHis.

Nous envisageons de poursuivre cette étude par l'utilisation des cartes de saillance *Saliency map* afin de visualiser ce que le CNN considère comme informatif sur les données de base et les données transformées. Cela nous permettra de proposer des métriques complémentaires à celles proposées ici.

Références

[AAC+17] Teresa Araújo, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, and Aurélio Campilho. Classification of breast cancer histology images using convolutional neural networks. *PloS one*, 12(6), 2017.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[Mac65] Robert H MacArthur. Patterns of species diversity. *Biological reviews*, 40(4) :510–533, 1965.

[MLPFS+20] Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphael Fournier-S'niehotta, Remy Poulain, Lionel Tabourier, and Fabien Tarissan. Measuring diversity in heterogeneous information networks. *arXiv preprint arXiv :2001.01296*, 2020.

[OPBM06] Albert Orriols-Puig and Ester Bernadó-Mansilla. Bounding xcs's parameters for unbalanced datasets. In *Proc. of the 8th annual conference on Genetic and evolutionary computation*, pages 1561–1568, 2006.

[OPBM09] Albert Orriols-Puig and Ester Bernadó-Mansilla. Evolutionary rule-based systems for imbalanced data sets. *Soft Computing*, 13(3) :213, 2009.

[Pre98] Lutz Prechelt. Early stopping-but when ? In *Neural Networks : Tricks of the trade*, pages 55–69. Springer, 1998.

[SK19] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1) :1–48, 2019.

[SOPH16] Fabio Alexandre Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *2016 international joint conference on neural networks (IJCNN)*, pages 2560–2567. IEEE, 2016.

[Sti98] Andrew Stirling. On the economics and analysis of diversity. *Science Policy Research Unit (SPRU), Electronic Working Papers Series, Paper*, 28 :1–156, 1998.

[TLB+19] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58 :101544, 2019.

[VDM01] David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *J. of Comp. and Graphical Statistics*, 10(1) :1–50, 2001.

[Wan03] Zhou et al. Wang. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[Zho04] Wang Zhou. Image quality assessment : from error measurement to structural similarity. *IEEE Trans. image proc.*, pages 600–613, 2004.