

Comparaison des performances entre les méthodes linéaires et non linéaires pour la prédiction de données transcriptomiques de cancers

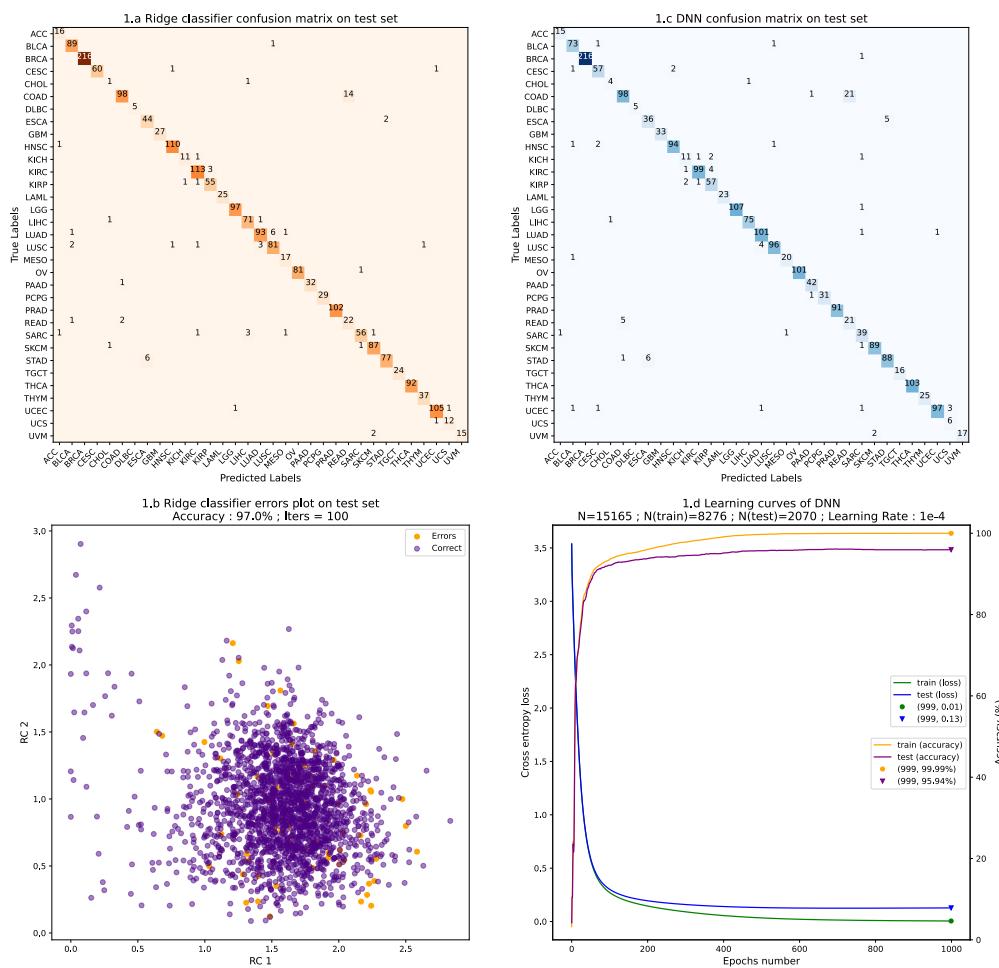
Louis Lalonde¹, Emire Medom Babou¹

¹Département de Biochimie et Médecine Moléculaire, Université de Montréal.

Introduction: Dans l'ère de la biologie computationnelle, l'analyse des données transcriptomiques issues des cancers peut nous permettre de développer des traitements personnalisés et efficaces pour améliorer les soins prodigues aux patients [1]. Le Cancer Genome Atlas (TCGA), un référentiel de données oncologiques, offre des ressources de grande qualité aux membres de la communauté scientifique pour réaliser diverses expériences analytiques. Toutefois, la complexité et la variabilité de ces ensembles de données posent des défis majeurs en termes d'analyse et d'interprétation. En ce sens, notre projet vise à évaluer de manière comparative la performance de différentes méthodes d'apprentissage machine sur les données TCGA pour que les membres de la communauté scientifique puissent prendre des décisions éclairées sur l'architecture de leurs expériences.

Méthodologies: Pour réaliser l'expérience, nous avons utilisé les données TCGA constituées de profils d'expression de gènes ARN-seq. 80% des données ont été réservées à l'entraînement tandis que 20% des données ont été attribuées aux tests. Deux évaluations comparatives (objectifs), évaluant à chacune d'entre elles une approche de résolution de problème distincte, seront effectuées. La première évaluation, de type classification, analysera comparativement un modèle d'apprentissage machine de la famille des méthodes linéaires et non linéaires. La deuxième évaluation, de type réduction de dimensionnalité, analyseront comparativement deux modèles linéaires. Nous évaluerons la performance des modèles de classification à l'aide de la métrique de précision tandis que nous utiliserons la variance moyenne expliquée par composante et le coefficient de Pearson pour évaluer les modèles de réduction de dimensionnalité.

Figure 1. Ridge classifier vs DNN performance analysis on classification of cancer type in TCGA data



que les précisions de nos modèles sont pratiquement équivalentes pour l'ensemble de données tests. 97.0% pour le Ridge Classifier (figure 1.b) et 95.94% pour le DNN (figure 1.d). À la lumière de ces résultats, nous en déduisons que pour les données TCGA, la méthode linéaire utilisée par le Ridge Classifier et la méthode non linéaire utilisée par le DNN

Résultats: Evaluation A. Classification Moléculaire : Régression Linéaire vs DNN. On observe que l'attribution des patients aux différentes étiquettes n'est pas identique pour les matrices de confusion du Ridge Classififer (figure 1.a) et celles du DNN (figure 1.c). Toutefois, on observe une convergence vers les mêmes résultats globaux de la classification. Le groupe de type cellulaire qui contient le plus d'étiquettes non confuses est le même pour les deux modèles : BRCA. Il y a exactement 216 étiquettes associées à ce groupe pour le Ridge Classififer et le DNN. Le groupe de type cellulaire qui contient le moins d'étiquettes non confuses est le même pour les modèles : CHOL. Il y a respectivement 1 et 4 étiquettes associées à ce groupe pour le Ridge Classififer et le DNN. En ce qui a trait aux types cellulaires confus, nous observons que la paire qui contient le plus d'étiquettes est COAD (cancer colon) et PRAD (cancer de la prostate). Il y a respectivement 14 et 21 étiquettes associées à ce groupe pour le Ridge Classififer et le DNN. Finalement, on observe

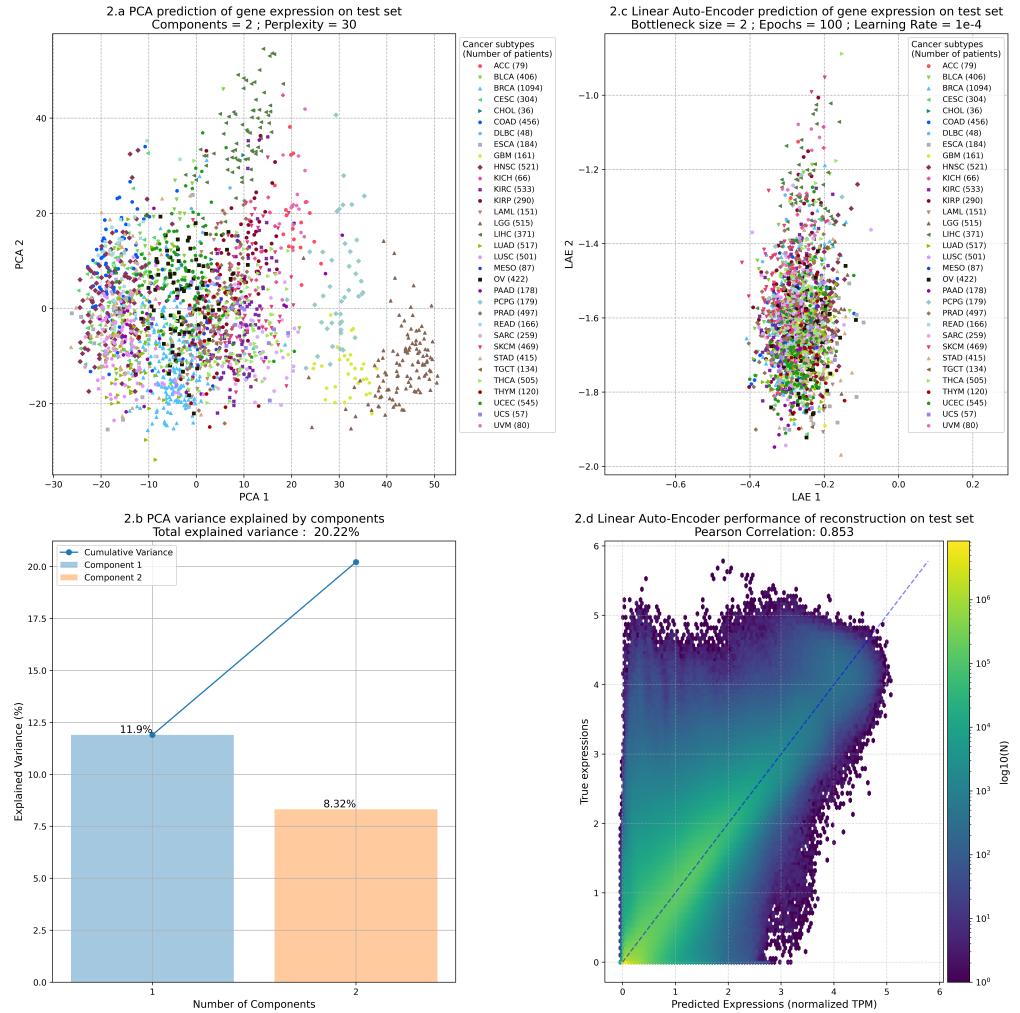
permettent de classifier les groupes de cancers avec une très haute performance de façon similaire, mais pas identique. On peut émettre l'hypothèse que les différences observées sont attribuables aux fonctions internes et aux objectifs de classification distincte des modèles. Davantage, dû à leur très bonne performance sur les données de l'ensemble test, on peut émettre l'hypothèse que les résultats observés seraient généralisables sur d'autres ensembles de données.

Évaluation B. Équivalence entre PCA

2D et Bottleneck d'Auto-Encodeur Linéaire. Pour l'analyse en composante principale (PCA), nous observons que trois groupes de cancers tendent à se distinguer du groupe principal (figure 2.a). On y retrouve les groupes LGG, GBM et LIHC. On observe que les valeurs de TMP normalisées ont une plage approximative de [-30, 50] pour la PCA 1 et de [-30, 63] pour la PCA 2 (figure 2.a). Pour la réduction de dimensionnalité effectuée par l'Auto-Encodeur Linéaire (LAE), on observe que les groupes de cancers LIHC, ACC et SKCM tendent à se distinguer du groupe principal, mais de façon très limitée (figure 2.c). En ce qui a trait aux métriques quantitatives, on observe que la première composante a une variance expliquée de 11.97% alors que la deuxième composante a une variance expliquée de 8.38% (figure 2.b). La variance totale expliquée des données par les composantes s'élève à 20.35% (figure 2.b). On observe que les valeurs de TMP normalisées ont une plage approximative de [-0.5, -0.1] pour la LAE 1 et de [-0.2, -0.995] pour la LAE 2 (figure 2.b). En ce qui a trait aux métriques quantitatives, on observe que le coefficient de corrélation Pearson est de 0.853 (figure 2.d). La divergence observée au niveau des cancers les plus différemment exprimés et la plage des valeurs TMP normalisées mettent en lumière la différence fonctionnelle interne des deux modèles. La PCA cherche à maximiser la variance expliquée des données en identifiant les composantes principales, alors que le LAE cherche à minimiser la perte d'information lorsque les données sont réduites par le goulot d'étranglement du réseau et maximiser la fidélité des données lorsqu'elles sont reconstruites. Finalement, il est intéressant de noter que les deux composantes de la PCA ne capturent que 20.22% de la variance totale expliquée. En revanche, le LAE reconstitue fidèlement 85.3% des données après compression. On peut émettre l'hypothèse que pour cet ensemble de données, le LAE est plus performant que la PCA à deux dimensions pour l'atteinte de son objectif.

Conclusion: Notre projet sur les méthodes d'apprentissage machine appliquées aux données TCGA permet de préciser et nuancer les performances de différentes méthodes couramment utilisées en biologie computationnelle. En comparant les méthodes d'apprentissage machine pour la classification, nous avons observé que le modèle linéaire (Ridge Classifier) et le modèle non linéaire (DNN) affichent tous deux des performances très élevées pour la classification des cancers. Pour les problèmes de réduction de dimensionnalité, l'Auto-Encodeur Linéaire a démontré une capacité supérieure à capturer les données et atteindre son objectif principal par rapport à la PCA. Toutefois, il est important de rappeler que malgré que ces deux méthodes d'analyses permettent de réduire les dimensions du jeu de données, leur objectif global reste différent et contexte dépendant. Ces résultats soulignent l'importance de choisir les bonnes méthodes analytiques dépendamment des besoins et des objectifs de l'expérience. Pour aller plus loin, il serait intéressant d'étendre le travail qui a été fait par Sauvé. L et Lemieux S. dans une expérience menée sur le développement d'auto-encodeur pour le pronostic du cancer à partir de données d'expression génique [1] et d'évaluer la performance et la généralisation de leur pipeline d'apprentissage machine Auto-Encoder Classifier-Deep-Neural network (AE+C) sur notre jeu de données TCGA.

Figure 2. PCA vs Linear Auto-Encoder analysis on classification of cancer type in TCGA data



Annexe A

Pour le jeu de donnée de BRCA: On a 11802 gènes, 1023 échantillons et 1023 classes d'étiquettes.

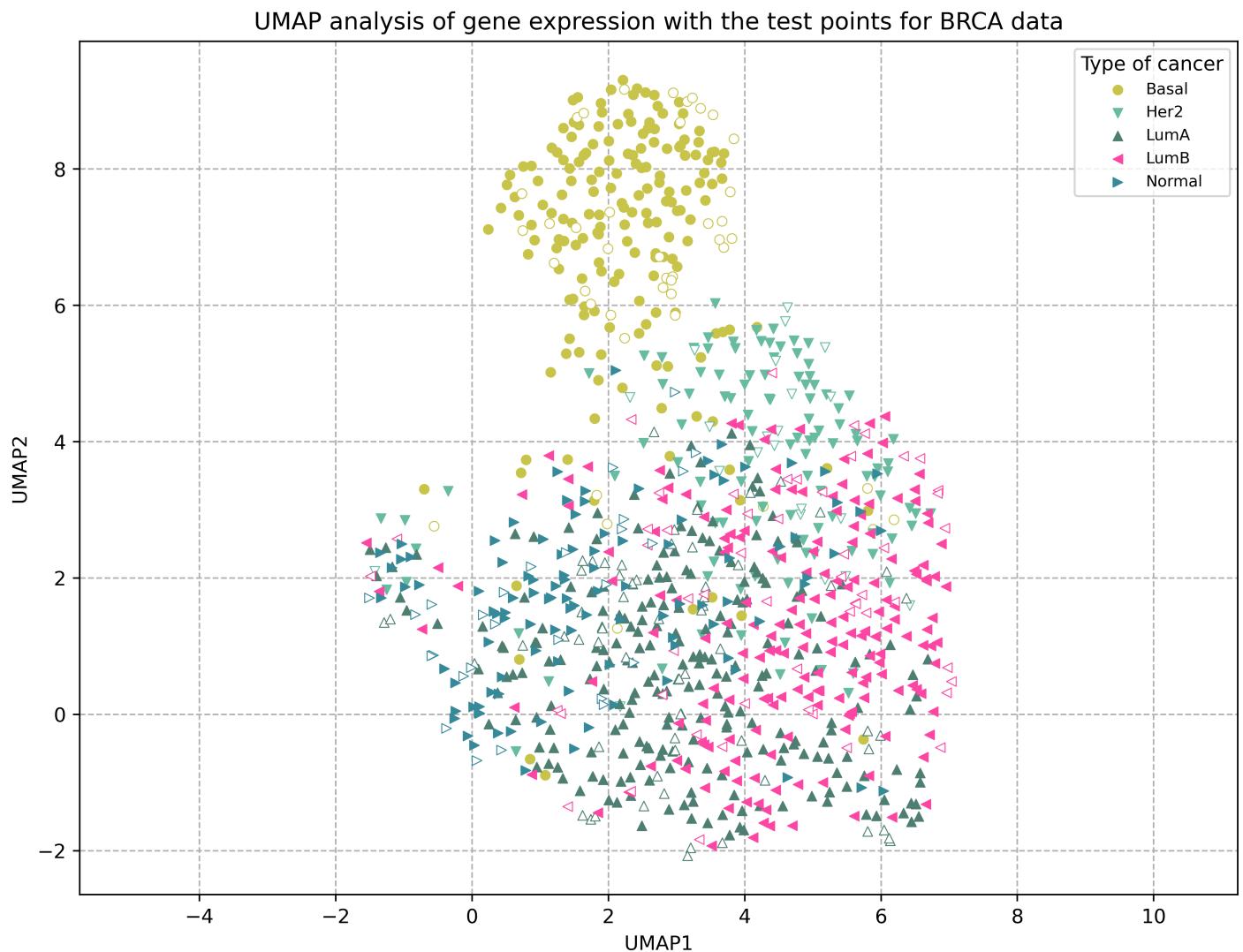


Figure 3: Composantes principales 1 et 2 de l'analyse UMAP de l'expression différentielle de gène de différents groupes de types cellulaires pour les données de BRCA.

Annexe B

Pour le jeu de donnée de TCGA: On a 15165 gènes, 10346 échantillons et 10346 classes d'étiquettes.

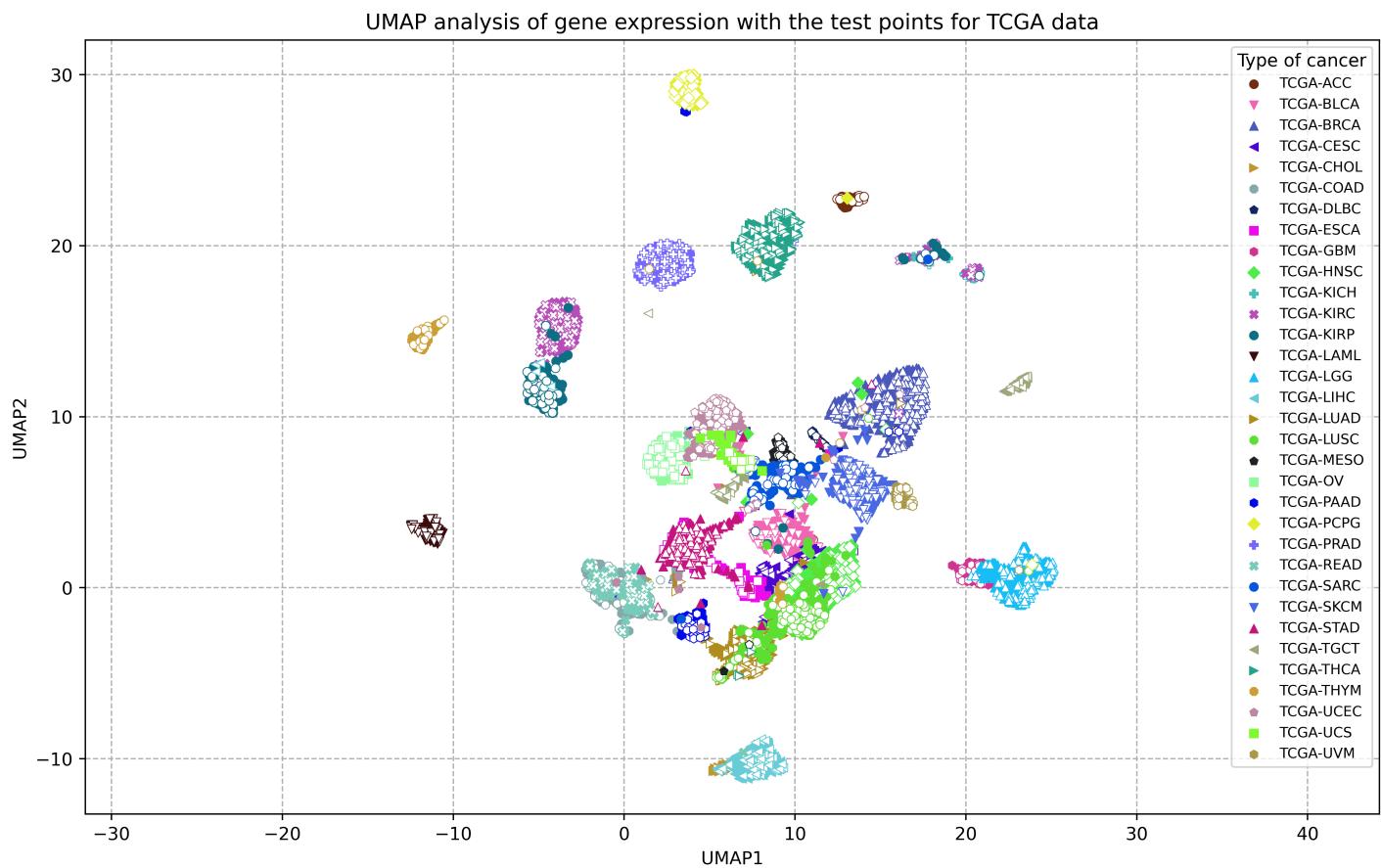


Figure 4: Composantes principales 1 et 2 de l'analyse UMAP de l'expression différentielle de gène de différents groupes de types cellulaires pour les données de TCGA.

Annexe C

Pour le jeu de donnée de TALL : On a 60660 gènes, 264 échantillons et 264 classes d'étiquettes.



Figure 5: Composantes principales 1 et 2 de l'analyse UMAP de l'expression différentielle de gène de différents groupes de types cellulaires pour les données de TALL.

Annexe D

Pour le jeu de donnée de LAML : On a 19597 gènes, 300 échantillons et 300 classes d'étiquettes.

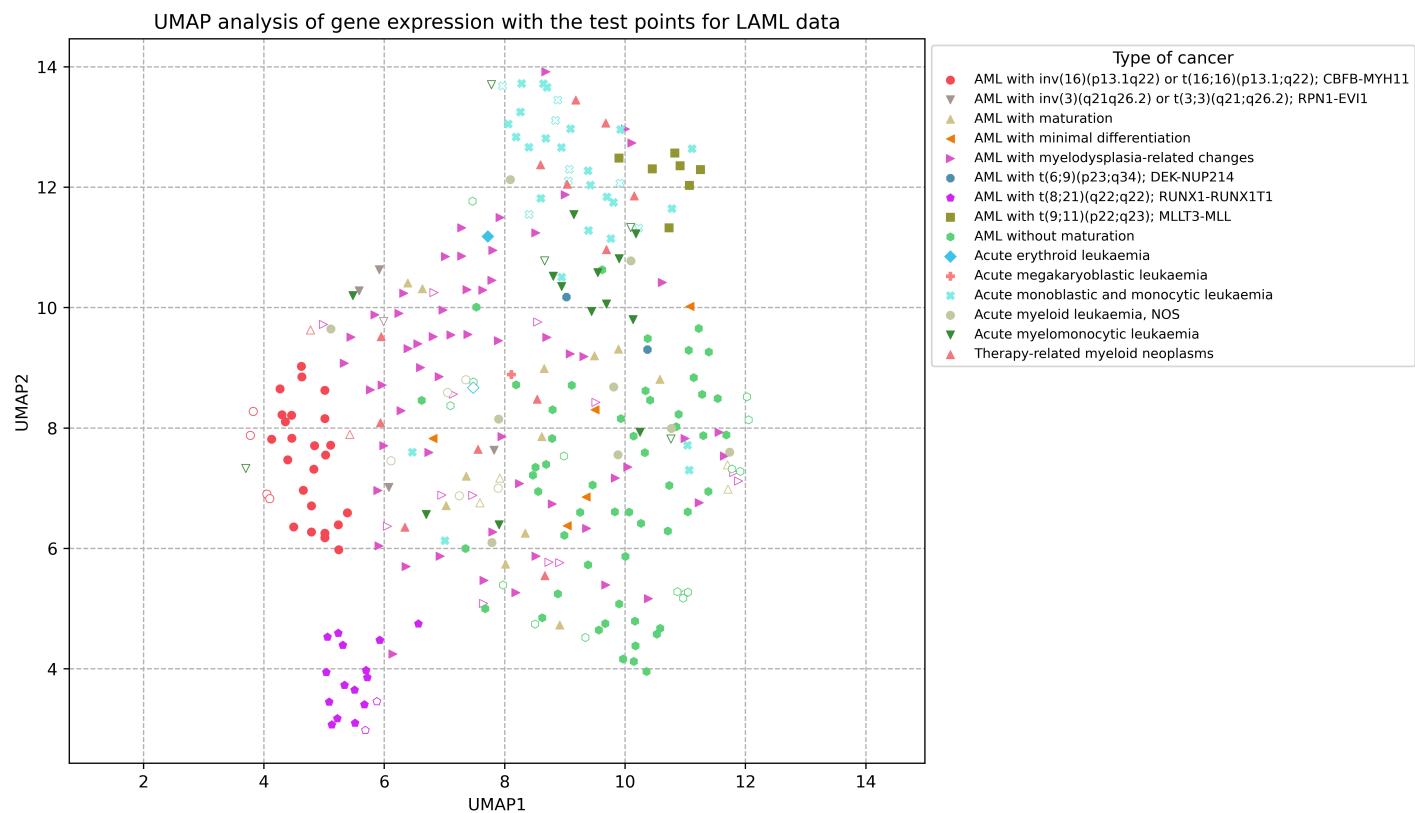


Figure 6: Composantes principales 1 et 2 de l'analyse UMAP de l'expression différentielle de gène de différents groupes de types cellulaires pour les données de LAML.

Annexe E : Création d'un réseau de neurones artificiel pour un problème de régression avec un auto-encodeur pour le jeu de données BRCA pour 1000 itérations.

L'auto-encodeur est un type d'algorithme d'apprentissage non supervisé utilisé pour régénérer la représentation d'un ensemble de données d'entrée. Son utilité est d'apprendre une représentation d'un ensemble de données, généralement dans le but de réduire la dimension de cet ensemble [2]. L'auto-encodeur a d'abord été proposé comme une généralisation non linéaire de l'analyse en composantes principales (ACP) par Kramer. Toutefois, les premiers travaux notables sont attribués à Geoffrey Hinton et à son équipe, qui ont introduit l'auto-encodeur pour la réduction dimensionnelle dans les années 1990. [3].

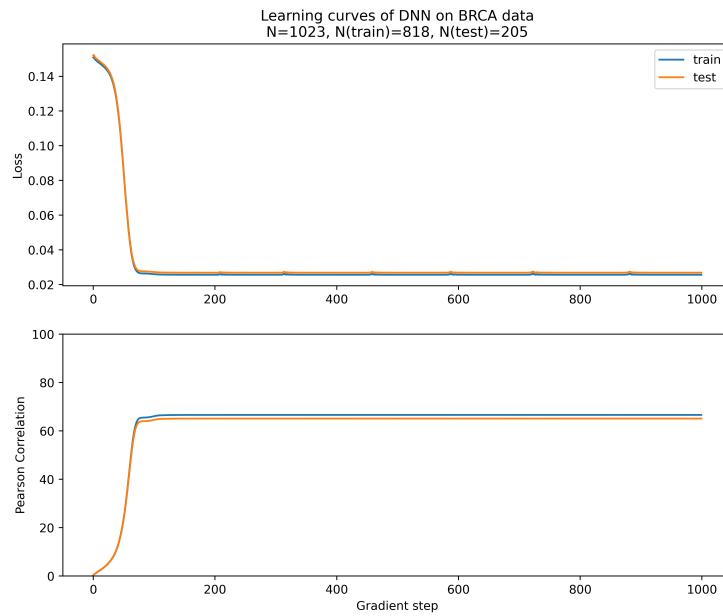


Figure 7: Courbes des pertes de la fonction de cross entropy et de la précision de l'entraînement du réseau de neurones en fonction du nombre d'epochs avec un auto-encodeur

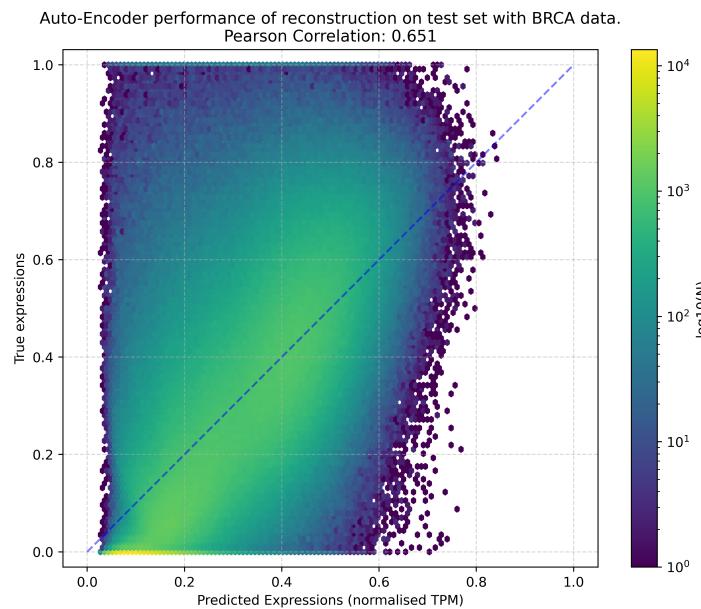


Figure 8: Visualisation de la corrélation sur les données de test prédites et vraie pour un auto-encodeur sur des données BRCA

Annexe F : Implémentation d'un Variational Auto-Encoder pour 1000 itérations

L'auto-encodeur variationnel est une extension novatrice des auto-encodeurs traditionnels. Il est utilisé dans le but de fournir une manière probabiliste pour décrire une observation dans un espace latent, ainsi il vise à introduire un cadre probabiliste pour générer la représentation compressée des données d'entrée au lieu d'un cadre déterministe comme un auto-encodeur [4]. L'auto-encodeur variationnel a été développé par Diederik P. Kingma et Max Welling [5].

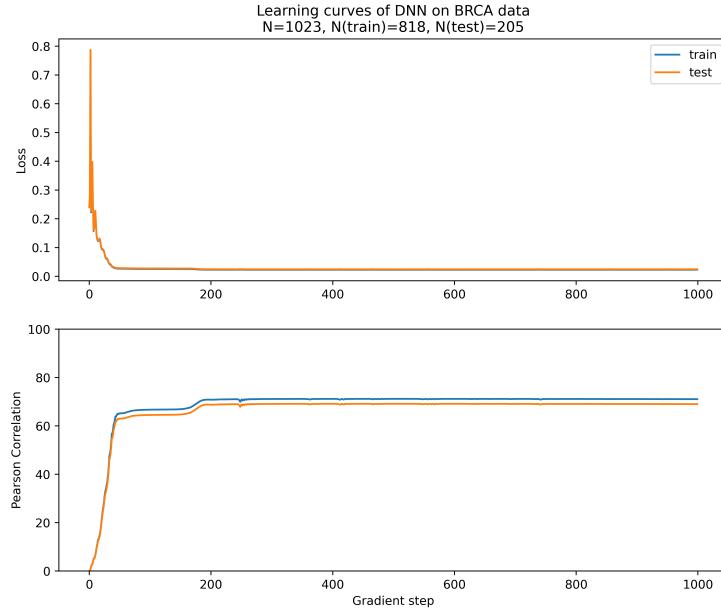


Figure 9: Courbes des pertes de la fonction de cross entropy et de la précision de l'entraînement du réseau de neurones en fonction du nombre d'epochs avec un auto-encodeur variationnel

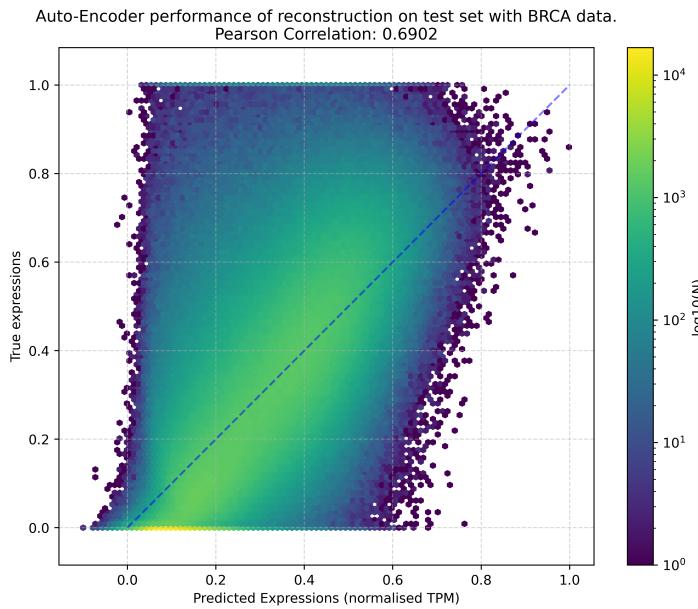


Figure 10: Visualisation de la corrélation sur les données de test prédites et vraie pour un auto-encodeur variationnel sur des données BRCA

Annexe G : UMAP supervisé pour 1000 itérations

L'UMAP supervisé est un algorithme de réduction de dimension basé utilisant des étiquettes cibles. Il permet une meilleure visualisation en conservant les propriétés structurelles importantes des données tout en extrayant proprement les classes connues [6]. Il a été développé par a été développé par Leland McInnes, John Healy et James Melville [7].

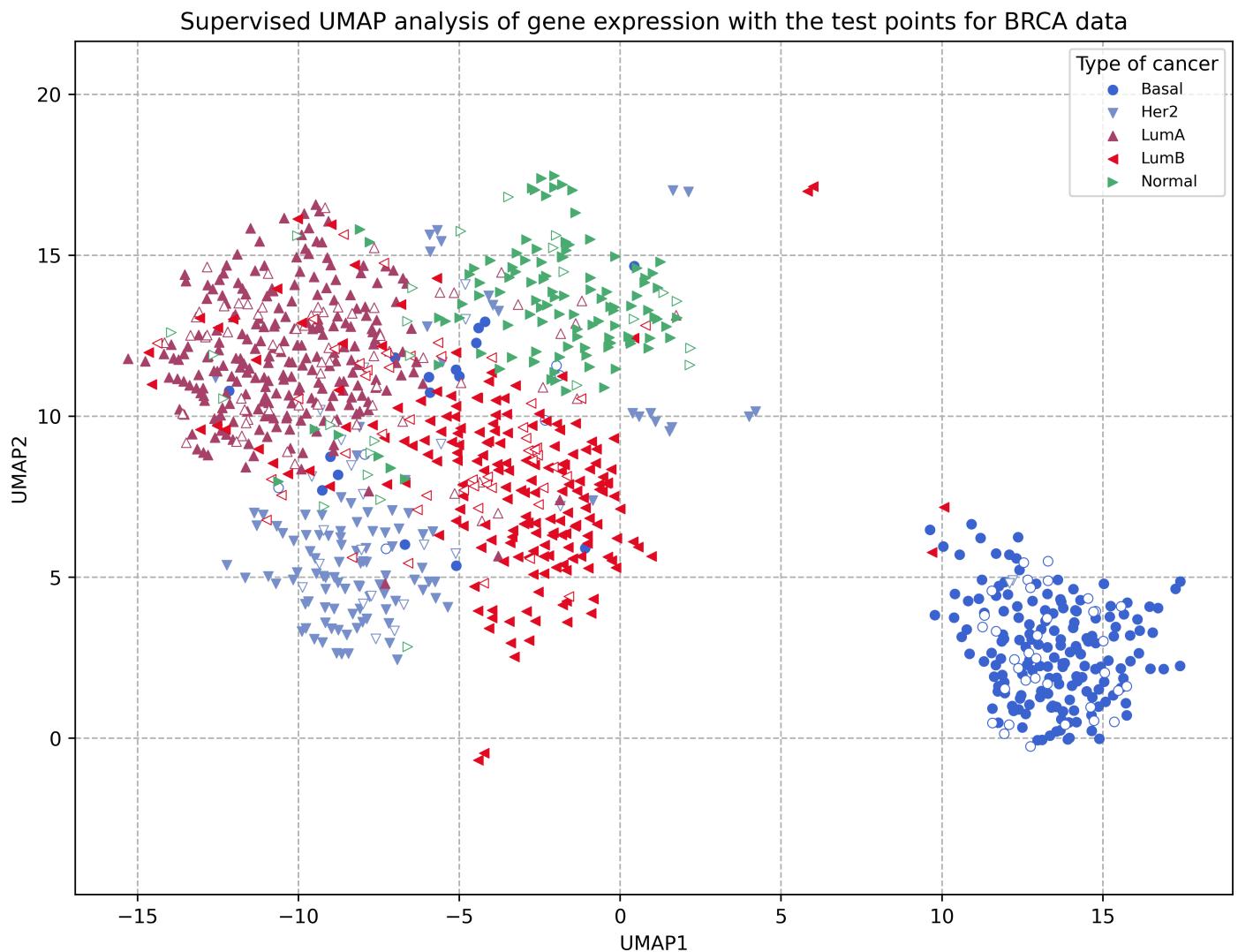


Figure 11: Composantes principales 1 et 2 de l'analyse UMAP supervisée de l'expression différentielle de gène de différents groupes de types cellulaires pour le jeu de données de BRCA

Références

- [1] L. Sauvé and S. Lemieux. "Development and Tuning of Auto-Encoder Approaches for Cancer Prognosis from Gene Expression Data".
- [2] G. E. Hinton et r. R. Salakhutdinov. "Réduire la dimensionnalité des données avec les réseaux de neurones".
- [3] Vonintsoa R. "Qu'est-ce qu'un auto-encodeur ?"
- [4] "Encodeurs automatiques variationnels".
- [5] Max Welling (2013) Diederik P Kingma. "Auto-Encoding Variational Bayes".
- [6] "<https://umap-learn.readthedocs.io/en/latest/supervised.html>".
- [7] McInnes et Healy. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction".