Quantum Machine Learning for Cancer Subtype Classification from Multiomics Data

Théo Berthet - 2024848162

May 23, 2025

Abstract

The analysis of omics data in oncology plays a key role to understand biological mechanism under cancer origins and to develop personalized treatments. However this task is limited by the high dimensionality of the data, their complex relationships et the limitations of traditional approaches. The Quantum Machine Learning (QML) is a promising solution to overcome those challenges by exploiting qubits' properties such as superposition and entanglement to treat complex data in an efficient way. This project aims to provide a pipeline to classify cancer subtypes from multiomics data. Using Quantum Support Vector Machine (QSVM) and Quantum Approximate Optimization Algorithm (QAOA), this pipeline aims to enhance diagnostic precision and identify biomarkers. Despite the current limitations of quantum computers, the fast progress in this field gives a ray of hope in the future applications of quantum computing.

1 Introduction

What is the problem ? Why is this problem interesting and important ? Why should it be solved now ? Why do I expect quantum computing to improve existing solutions to the problem ?

Genomic advances completely transformed our comprehension of disease, particularly in oncology, offering new approaches such as personnalized medecine and target therapy. One important step in cancer study is the classification of cancer subtypes from genomic data (expression profile from RNA-seq data or DNA methylation). This classification allows a better understanding of the tumors, identifying biomarkers, and so to develop adapted treatment for each individual. However, this kind of study is quite challenging due to the high dimensionality of the data, their noise and the limitation of actual methods.

Genomic data are usually subject to the curse of dimensionality since the number of features can be quite large (number of genes) compare to the number of samples (cells). This complicates the use of classical machine learning methods. Moreover, biological data are noisy because of technical or biological variations and the classical methods struggle to identify complex and often non linear relationship in multiomics data. Also, another limitation is faced due to the lack of scalability of the current approaches while the actual database continue to increase. So the classic approaches are facing their limitations to manage such big datasets as multiomics. However the current improvement in cancer treatment depend of the development of models able to treat these data in an efficient and precised way.

Quantum Machine Learning (QML) is an emerging solution to tackle those issues. In this way, quantum computing could be an opportunity to overcome the current limitations of classical computing. In fact, QML is a promising way to analyse complex data efficiently ([SSP15]). Using superposition and entanglement enables to treat large data efficiently. Moreover, some quantum methods such as QSVM and QAOA enable to apply quantum computing for classifying tasks and present some promising results that could even outperform classical approaches for complex tasks, as the work of Havlicek et al [HCTea19] suggested it. QAOA can explore complex solution space efficiently which can be crucial to select informative characteristics from the multiomics data. QSVM can capture subtle relationships between genes by projecting the data in non linear space. It should also be mentionned that the quantum technologies are developing quite quickly and so quantum algorithms become more and more feasible for real application.

With this project, we aim to propose a way to use the last quantum computing advances in order to enhance cancer subtypes classification using genomic data. This subject is very interesting since according to the World Health Organization, cancer is one of the principal causes of death in the world with almost 10 millions of death in 2020. A precise classification of cancer subtypes is essential in order to find an appropriate diagnostic and provide a target drug to patients. This is further supported by the explosion of omic databases creating a need in the development of tools able to exploit massive dataset. Since its creation, the TCGA database represents more than 2.5 petabytes of omics data, about 33 types of cancers and more than 20,000 samples ([NWC+13]). Also, even though quantum technologies are still limited by their noise and scale, technologic progress allows the experiment of some QML algorithms and maybe in the future allows valid applications.

2 State of the art

What are some related works to the problem ?

The application of QML for cancer genomics is an emerging field so the number of related works is quite limited. However, some studies have already made some steps into the crossing between oncology and quantum computing.

One of the most noticeable work in this field is the one done by Saggi et al. [SBI+24] which explore the integration of QML for multi omics data to classify efficiently subtypes in lung cancer such as adenocarcinoma or carcinoma. Using Quantum Neural Network, the study shows that it can outperform classical neural network and make it possible to identify complex biomarkers and improving the classification of subtypes by 15%. This underligns the potential of QML to transform omic data analysis for oncology. An hybrid method was adopted, using classical dimensionnal reduction methods combined with quantum classification algorithms such as a Quantum Neural Network with a Quantum Layer Ansatz. But even though this is one of the most promising paper, integrating multiomics data in an hybrid pipeline to perform the lung cancer subtypes classification, we can underlign some possible limitations since the model only classified into 2 modalities and was tested only in the frame of lung cancer.

As it is stated in the work of Benedetti et al. [BSB+21], QML can enhance cancer genomic data analysis to classify subtype and even identify mutations. The focus was given on quantum parallelism in order to process large genomic datasets exploring the use of variationnal quantum circuit (VQC). This parallel processing enable the detection of complex pattern and relationships within the data often missed by classical computing. Employing VQC it's a way to reduce the circuit depth in order to obtain more feasibile algorithms. However, this study use quantum simulators and not real hardwares so we can have some reservation about the biases that a real implementation could introduce.

In the study of Li et al. [LCZ+20], quantum-inspired algorithms are explored for features selection in cancer genomics. This work highlights the potential of quantum computing to identify the most important features in large genomic datasets, improving prediction accuracy where handling this kind of data is quite challenging for classical approaches. This suggests that quantum inspired methods such as Quantum Approximate Optimization Algorithm

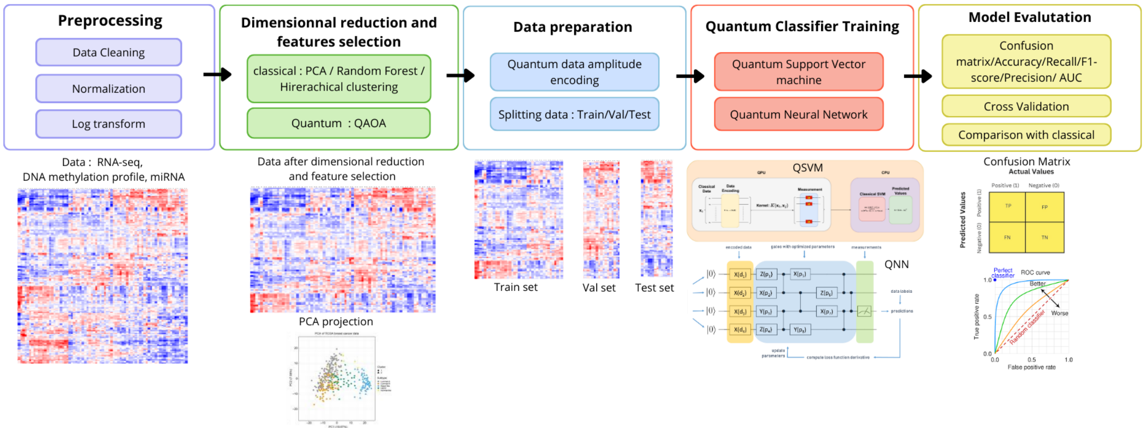
Preprocessing
Data Cleaning
Normalization
Log transform
Dimensionnal reduction and features selection
classical : PCA / Random Forest / Hireachical clustering
Quantum : QAOA
Data preparation
Quantum data amplitude encoding
Splitting data : Train/Val/Test
Quantum Classifier Training
Quantum Support Vector machine
Quantum Neural Network
Model Evalutation
Confusion matrix/Accuracy/Recall/F1-score/Precision/ AUC
Cross Validation
Comparison with classical
Data : RNA-seq, DNA methylation profile, miRNA
Data after dimensional reduction and feature selection
PCA projection
Train set
Val set
Test set
QSVM
QNN
Confusion Matrix
Actual Values
Predicted Values
TP
FP
FN
TN
ROC curve
Perfect classifier
Better
Worse
True positive rate
False positive rate

Figure 1: Method workflow

(QAOA) could overcome computation barriers in genomics and enhance cancer diagnosis. By identifying the informative genes or mutations, it opens the way to reduced the dimensionality of the data (which is quite useful for high dimension data such as The Cancer Genome Atlas) and so possibly increasing the accuracy of predictive modeles.

Finally, in terms of cancer diagnostic, we can cite the work of Perez et al. [PGFM20] that tackles the subject of QML to enhance early cancer detection, using a quantum classifier to differentiate between healthy and malignant genetic profiles at early stages. The classifier used is a Quantum Support Vector Machine (QSVM). Once again, the characteristic of quantum computing such as quantum parallelism allows to analyze complex genomic data more efficiently in order to provide accurate and fast diagnostic. The results show that the QSVM is able to identify weak signals in the data and could be a promising approach for early diagnostic.

One more paper should be mentioned even though it's not completely related to quantum computing but on the dataset available for cancer genomics. The work of Weinstein et al. [NWC+13] led to the build of a dataset called The Cancer Genome Atlas (TCGA) which allowed to identify genetic mutations associated with cancer. This dataset is composed of multiomics data including DNA, gene expression profiles, epigenetic variations etc. This dataset is one of the references for cancer studies in bioinformatics as it's a rich dataset offering a lot of possibilities to test QML. But this voluminous dataset is also imbalanced and quite noisy constituing some obstacles to its use.

So we can see that most of the studies use quantum simulation to solve their problems, it's a way to tackle the issue of the limitation of current hardwares but the application on real quantum system can be questionable. Also, the studies usually focus only on 2 types of cancers only.

3 Methods and Approaches

What are the potential methods and/or approaches to solve the problem ?

From the previous works and our class materials, different methods could be used for this problem. I tried to sum up the possible methods and approach in the figure 1.

First, it's important to get data and preprocessed them. Different publicly available dataset are available like the The Cancer Genome Atlas (TCGA) that include RNA sequencing and DNA methylation profiles. The preprocessing of the data is a critical step since it can have a huge impact on the accuracy of the classifier. Those data should be processed by usually applying a log-transforme (to obtain a better distribution) and a normalization of gene expression is done.

The next step would be to reduce the dimensionality of the data using a PCA by classical computing for example. This step allowed to reduced the complexity of the data and also the number of data the quantum algorithm would have to treat. Effectively, since the current hardwares are quite limited it would be better to reduced the dimensionality to make it more feasible. Especially since the TCGA is a quite big dataset. Another way to do so could have been to apply quantum annealing to identify informative features (genes), or we can take inspiration in the work of Li et al. [LCZ+20] who suggest that QAOA could be used to find informative features inside big dataset by studying non linear interaction between genes. This feature selection could be compare to classical PCA but would use a quantum approach but since the size of the dataset is quite large it's unlikely to be feasible right now. That's why in some studies such as the work of Saggi et al. [SBI+24]. classical computing is used to realize the feature selection. This time, they choose to combine different algorithms for this step including PCA, Random Forest etc to find the best one and then applied a hierarchical clustering.

To encode those data into a quantum computing friendly way and efficiently, some technics exist such as amplitude encoding that would allow to encode a significant number of features in a minimum number of qubits. For example, un vector of 2^n features can be encoding in n qubits. It would be a way to optimize quantum ressource allocation and it was use in the work of Saggi et al. [SBI+24]. One step before employing a classifier can be to concatenate the multiomics data for each patient in order to use all the available data.

Using those selected features we can now try to train a classifier to distinguish cancer subtypes. An approach could be to use a Quantum Support Vector Machine (QSVM) using using Quantum Kernel (as Havlicek et al. [HCTea19] and Zarei et al. [ZA24] suggest the potential) to map the data into a high-dimensional Hilbert space to allows an easier separation of data and then use a classifier on this data. This model could be adapted to treat complex/non-linear relationship between genes. The goal of the QSVM is to find the optimal hyperplane to classify the cancer subtypes using quantum kernel matrix. The quantum kernel computes the similarity between data in a quantum space.

Another approach can be to use Quantum Neural Networks (QNN) (method that is used in the study of Saggi et al. [SBI+24]). For example a variational quantum circuit use quantum backpropagation to train this QNN classifier and learn from the omics data to optimize the weight parameters. Those two quantum methods offer quite promising results as highlighted in the introduction, by noticing complex relationships in the data allowing an enhancement of the predictions' accuracy.

Finally, the model need to be evaluated. Using benchmark dataset we can compare some classification metrics with our model, such as accuracy (proportion of correct predictions), precision (the proportion of true positive predictions), recall (the true positive rate or sensibility), F1-score (harmonic mean of precision and recall) and ROC curve + AUC (mesure the ability of a model to distinguish between classes).

$$Accuracy = \frac{TruePositives + TrueNegatives}{Total}$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Those metrics would be a way to estimate the quality of the classifier to distinguish the different subtypes. Also in order to train the model and assure a good generalisation of the prediction we can split datast into 3 sets : train set to train the model, validation set to finetuned the model and finally test set to evaluate generalization. A good practice would be to apply cross validation.

Also, an interesting idea could be to compare the result of the quantum approach with a classical one such as classical SVM or Neural Networks in order to evaluate the real improvement the quantum approach can offer in cancer subtypes classification in contrast with classical methods.

So, the pipeline defined would be an hybrid one combining classical computing and quantum machine learning algorithms. The preprocessing and feature selection would be done by classical methods and the quantum computing would be use only for the classification task. The major argument for this is the limitation of current quantum hardwares (that would be discussed in the next section) but in a near future a quantum preprocessing could be applied and offering different results.

4 Feasibility and Discussion

Do I expect the approach/solution to be feasible and why ?

About the feasability of this approach, different points should be studied.

For the availability of the data, good database are available giving an easy access to multiomics data, such as TCGA. So the availability of the data is quite good. The problem we can face is how to get those data quantum-friendly to gain in efficency. This point is one of the most critical. Also some good environnement are available (like Quiskit or Tensorflow quantum) making available to try quantum algorithm. We can also mention that some quantum computer and servers are available (like IBM Quantum computer) but it faces some limitations. One of the most concerning is the number of qubits that are quite limited and also the computing accuracy. So only a few number of qubits are available and they introduce a lot of errors so the reliability is quite low.

But as we saw in the state of the art, some promising results were obtained ([BSB$^+$21] and [SBI$^+$24]). In order to use QML, Saggi et al. [SBI$^+$24] employed an hybrid method combining classical and quantum machine learning. In particular, they use quantum simulator to bypass current limitation of hardwares. In their work, Benedetti et al. [BSB$^+$21] used variational quantum circuit that is more adapted for the limitation of the quantum hardwares by minimizing the number of operations and the resulting errors. Also, they preprocessed the data using classical computing in order to simplify the processing by quantum algorithms. So thanks to those works, we can see that using hybrid approach between classical and quantum computing can allow us to put in practice our study despite the current limitation of quantum hardware.

We can also envisage the feasability of our project in a near future. The quantum hardware are improving. So in a few years, the number of qubits will increase allowing more complex computing and the error rate will go by decreasing giving more reliability to the final results. Indeed, IBM Quantum and Google are estimating that some systems could reach several thousands of qubits around 2030 (https://www.ibm.com/roadmaps/quantum/), if this rise of number of qubits is associated with a decrease of the error rate, more complex quantum algorithm could be tested to process omics data.

So, current quantum hardwares are quite limited offering not a good feasibility of our project. However, hybrid approaches can be a present solution to integrate QML to omic data analysis in oncology. And, thanks to technological progress, those current limitations could be overcome in a near future, making this project more feasible but also essential to enhance personalized healthcare and cancer research.

5 Conclusion

This project propose an innovative method to analyse omics data for cancer subtypes classification by combining classical and quantum approaches. It's an important problem since the current method struggle to exploit the multiomic datasets which are a key to improve our understanding and treatment of cancers. Using an hybrid approach is a way to avoid quantum hardwares limitations and classical computing problems. Technics such as QSVM and QAOA allow to overcome limitations of traditional methods to treat complex and high dimensionnal data.

The expectation of this project is to enhance cancer subtypes classification, a more accurate identification of biomarkers and an optimization of personnalized treatments, or at least this study could aim to compare classical approach with a quantum one and by do it so giving some foresights about the potential of quantum computing in oncology. This could open the way for future applications following the development of quantum computing. In the future, the augmentation of the number of qubits and the reduction of the error rate the impact of quantum machine learning in the personnalized healthcare or in the oncologic research could be quite ambitious and allow to save many lives.

References

[BSB$^+$21] Marcello Benedetti, Massimiliano Saggio, Montserrat Banuls, Micaela Faccin, and Marco Pistoia. Quantum machine learning for cancer genomic data analysis. *Quantum Science and Technology*, 6(4), 2021.

[HCTea19] V. Havlíček, A. D. Córcoles, K. Temme, and et al. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019.

[LCZ$^+$20] Xiang Li, Yao Chen, Xin Zhang, Ling-Yun Duan, and Jun-Ping Xie. Quantum-inspired feature selection for cancer genomics. *IEEE Transactions on Quantum Engineering*, 1(1), 2020.

[NWC$^+$13] Cancer Genome Atlas Research Network, Joseph N. Weinstein, Emily A. Collisson, Gordon B. Mills, Kenna R. M. Shaw, Bradley A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, Oct 2013.

[PGFM20] Anna Perez, Tomas Garcia, Javier Fernandez, and Luis Martin. Quantum machine learning for early cancer detection. *Journal of Quantum Information Science*, 8(2), 2020.

[SBI$^+$24] Mandeep Kaur Saggi, Amandeep Singh Bhatia, Mensah Isaiah, Humaira Gowher, and Sabre Kais. Multi-omic and quantum machine learning integration for lung subtypes classification. *arXiv:2410.02085*, 2024.

[SSP15] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185, 2015.

[ZA24] Mohadeseh Zarei and Elaheh Afsaneh. Potential of quantum machine learning for solving the real-world problem of cancer classification. *SN Applied Sciences*, 6(10):513, 2024.