Molecular Subtyping of Colorectal Cancer: In the Frontiers of Personalized Diagnostics and Treatment

Author: Susanna Avagyan
BS in Data Science
American University of Armenia
Yerevan, Armenia
susanna_avagyan@edu.aua.am

Supervisor: Hans Binder
Interdisciplinary Center for Bioinformatics
Leipzig University
Leipzig, Germany
binder@izbi.uni-leipzig.de

Abstract—**Being the third most diagnosed and second most deadly cancer worldwide, colorectal cancer is a highly complex, multigenic disease that has very high inter-patient variability in terms of the genetics of the tumor. This raises the need for developing a personalized treatment for CRC patients for better efficacy and reduced toxicity. Molecular subtyping of the disease is a way to define biological subgroups for which targeted treatment can be optimized. Our research aimed to test different Machine Learning and Deep Learning models that combined theoretically and practically tested state-of-the-art concepts to obtain biologically meaningful clusters from somatic mutations and copy number alterations as CRC patient subgroups. Four different methods with different types of inputs were tested, Spectrum, xGeneModel, Kmeans clustering, and Deep Embedded Clustering. Our results proved the most efficient way of obtaining these subgroups to be a Deep Learning clustering model (DEC) applied to prior biologically enriched data using Biological Process genesets from Gene Ontology. The obtained clusters were treated as labels to build classifiers as a predictive tool for incoming patient records, from which Logistic Regression performed the best. Survival Analysis showed that the obtained clusters were not distinct in terms of overall survival patterns. However, we brought forward the hypothesis that these can be significantly different considering specific drug treatments, for which we did not have sufficient data to check the hypothesis. The code and material of the method are available at: https://github.com/susieavagyan/capstone-cancer-subtyping**

Keywords—**cancer, colorectal, subtyping, clustering, personalized medicine**

Introduction

Worldwide, colorectal cancer is the third most diagnosed cancer and the second most deadly cancer. An estimated 1,9 million people were diagnosed with colorectal cancer in 2020, and around 1 million died. Colorectal cancer (CRC) is highly heterogeneous at the genomic and transcriptomic levels. Two important features of CRC are high inter-patient variability and high spatial heterogeneity. These characteristics influence the molecular characterization of tumor tissue, hence challenging the ”one-fits-all” approach of current medicine in terms of disease treatment and progression.

Personalized(precision) medicine uses patient-specific genetic or other biomarker information to make treatment decisions that can make patient care more efficient. Its applications in cancer treatment have been shifting the organ-centric generalized treatment choices towards a more molecular level, personalized analysis, and decision making. So, to obtain subgroups of cancer cases that are molecularly distinct and can be experimentally optimized for more precise, efficient, and less toxic treatment, molecular subtyping using various types of data (somatic mutations of the tumor, gene expression changes in cancer cells, molecular pathway disruptions) can be performed. With the development and improvement of sequencing and gene expression measuring techniques, the data to perform this task have significantly increased, providing more predictive capacity for any proposed methods to perform subtyping. However, there is no highly optimized algorithm or method to perform this task just yet, and much research to define new methods using different types of data is ongoing.

Subtyping using somatic mutations has recently proven to have much potential in terms of this task. Somatic mutations are stable and have critical functions in cancer development and progression (Vural et al., 2016). Moreover, investigating somatic mutation profiles can aid in cancer diagnosis and treatment due to the vast number of clinical guidelines based on single gene mutation (Kuijjer et al., 2018). This project aims to use Machine Learning and Deep Learning methods to obtain molecular subtypes/clusters of CRC, classify new patient data, and evaluate the performance of each proposed model in terms of biological meaning and clinical outcome.

However, as somatic mutational data is very sparse and high dimensional, the common clustering algorithms often lead to biologically meaningless clusters. Hence with this paper, we also aim to prove the hypothesis that using somatic mutations as a baseline and performing an additional data enrichment/stratification step, which aims to integrate more biological meaning into the data and reduce feature space size, will provide better results.

We then apply four different clustering methods: Spectrum, xGeneModel (Zhang et al., 2018) with driver gene focus, Kmeans clustering with enriched data, and Deep Embedded Clustering (Rohani et al., 2020) with enriched data. For the classification tasks, a number of classification algorithms, such as Random Forest, Logistic Regression, Multi-Layered Perceptron, etc., are used. The obtained clusters are further analyzed using biological insights and clinical data about patients. Specifically, Kaplan Meier Survival Analysis is done for validation of cluster distinctiveness in terms of disease progression and survival.

Materials and Methods

A. Dataset

The dataset was obtained from the public database cBioPortal, from a recent study MSK-MET 2O21 (Memorial Sloan Kettering - Metastatic Events and Tropisms 2021) study (Nguyen et al., 2022). It contains tumor genomic and clinical outcome data from a pan-cancer cohort of over 25,000 patients with metastatic diseases, including information about 50 different primary and metastatic cancer types. This paper is going to focus on colon and rectal cancer, which includes 3093 patients from the dataset, from which 2073 have primary and 1020 metastatic cancers (Fig 1). For model building, primary cancer data is used.

Exploratory data analysis (EDA) showed that our data is very sparse (Fig2,3). This means there are many patients with very few mutations, which is typical for cancer mutation profiles. There are also numerous patients with hypermutated tumors (n>15)

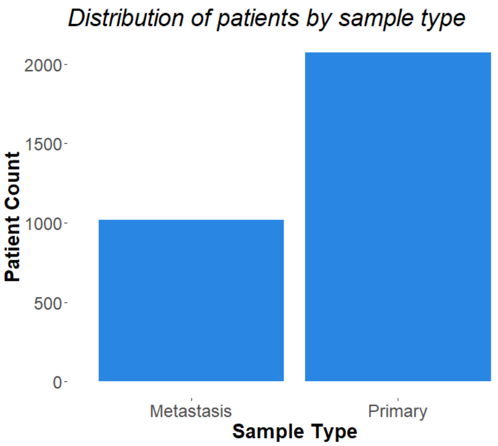

Fig. 1. Primary vs Metastatic cancer patients count

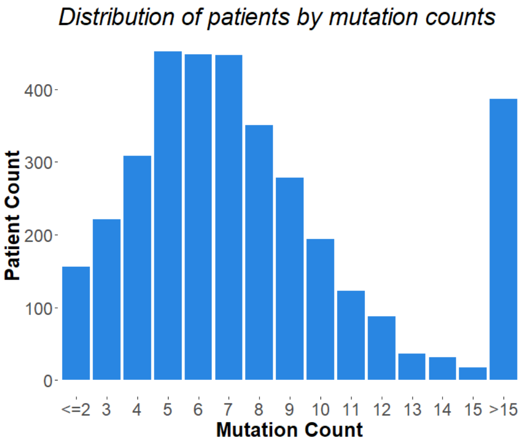

Fig. 2. Distribution of patients based on mutation count. Most patients have 5-8 mutations

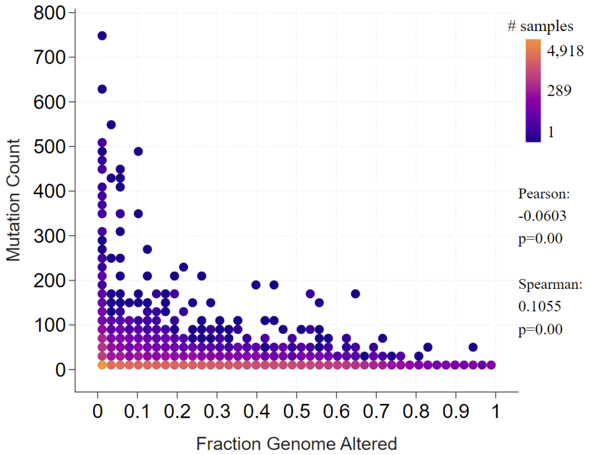

Fig. 3. Fraction of Genome Altered is the percentage of the genome that has been affected by copy number gains or losses. Total Mutations refers to the number of mutations that are found in the tumor genomes

B. *Preprocessing*

After obtaining relevant genomic (including mutations and copy number variations (cnv) and clinical data, preprocessing of the data was performed to

- remove nonrelevant for the task types of mutations, such as silent (no effect) and splice site mutations
- remove cnv's that cause loss of heterozygosity
- map gene names to universal HGNC nomenclature
- transform dataframes to one-hot-encoding for each patient-gene pair

As the model was going to perform clustering, then classification on a single data frame, we needed to combine the mutational and cnv data with some biological logic behind it. Hence, the data were combined by counting cnv data as mutations in case

- cnv is a deletion in a tumor suppressor gene
- cnv is an amplification of an oncogene

EDA of the patient clinical information showed general patterns of our patient sample. In terms of gender, our dataset is balanced (Fig.4), white race group is overrepresented (Fig.5), age is normally distributed within the range <30 - 90. (Fig.6).

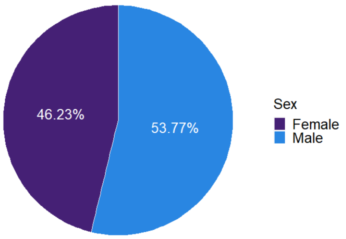

Fig. 4. Distribution of patients by sex

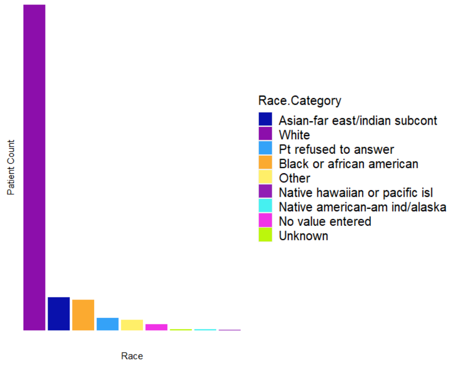

Fig. 5. Distribution of patients by race

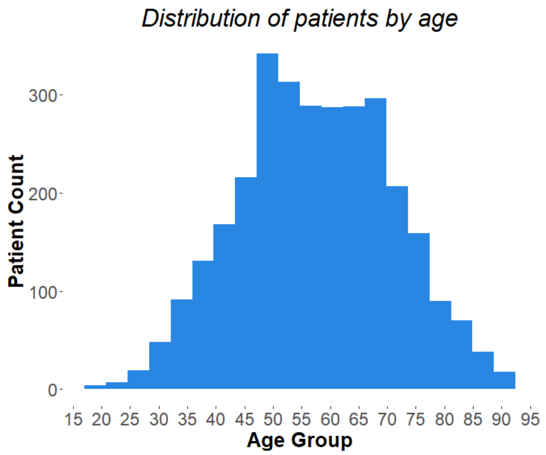

Fig. 6. Age of patients at the time of genetic sequencing of tumors (data collection

Models

Obtaining the processed data, we defined two major steps that it needed to go through - clustering and classification.

C. *Clustering*

The clustering step is required for cancer case subtyping, which can be further used for identifying descriptive genetic signatures for patient groups and use this for developing personalized treatment. Cancer subtyping on somatic mutations has great potential. However, it poses a couple of challenges. Firstly, the mutational data is very sparse. As our EDA also showed, the vast majority of patients have profiles with 5-8 mutations. This data is difficult to cluster, just considering the mathematical operations that happen under the hood. Secondly, the genetic variance is very big. This means even if each patient has very few mutations, those mutations happen in a wide range of different genes and are very different from person to person. This is especially an issue for colorectal cancer; as compared to other cancers like breast cancers, CRC has high spatial heterogeneity and high interpatient variability (Molinari et al., 2018) Having these in mind, we need to implement methods that are optimized in terms of different aspects: dealing with data sparsity, integrating biological meaning, optimizing local density-awareness, etc.

We therefore propose and compare four different clustering methods: Spectrum, xGeneModel (Zhang et al., 2018) with driver gene focus, Kmeans clustering with enriched data, and Deep Embedded Clustering (Rohani et al., 2020) with enriched data. For all methods, clustering is done with k = 5 clusters, which was chosen arbitrarily.

1) Spectrum: The first method, Spectrum, uses processed one-hot encoded mutation matrix and applies a self-tuning density-aware kernel that enhances the similarity between points that share common nearest neighbors. It uses a tensor product graph data integration and diffusion procedure to reduce noise and reveal underlying structures. The algorithm basically clusters eigenvectors derived from a matrix representing the data's graph. The adaptive density-aware kernel, which calculates the similarity matrix between samples is defined as follows:

$$A_{ij} = exp(\frac{-d^2_{(s_i s_j)}}{\sigma^i \sigma^j (CNN_{(s_i sj)} + 1)}),$$

where d_{ij} denotes the Euclidean distance between points s_i and s_j, σ is a local scaling parameter, CNN denotes the number of points in the intersection between the two sets of nearest neighbors of points s_i and s_j. The kernel increases A_{ij} when s_i and s_j share more neighbors and therefore adapts to the data's local density. Two additional parameters, P and S, are passed on to the algorithm to define the size of the kernel: P is the number of nearest neighbors to use when calculating local sigma, and S is the number of nearest neighbors to use when calculating common nearest neighbors. (John et al., 2020). We ran the algorithm with P = 4 and S = 6. These numbers were chosen empirically judging by runtime and similarity matrix values (higher values will prefer global structures, while lower values local structures).

2) xGeneModel: In this method, the functional similarities of the cancer driver genes, which are genes in which acquired mutations are causally linked to cancer progression, are integrated with the mutational to calculate the genetic distance between tumors. Two precalculated matrices, FSM and WM, are used to perform clustering via integrating biological information about driver genes. FSM is the functional similarity matrix for all the putative cancer driver genes, which are taken into account in the analysis. Functional similarity is defined as the correlation between two genes in terms of the correlation of the biological process and molecular function terms (from Gene Ontology database)(Harris et al., 2004)) that those genes are involved in. WM is the weighted between-gene similarity matrix, where weights are the confidence scores of the putative cancer driver genes being the 'true' cancer driver genes. Given these two matrices, the mutation similarity matrix is calculated element by element for the tumors in a cohort. This is then passed on to Ward's hierarchical clustering analysis, which provides cluster assignments.

3) Data Enrichment: Keeping in mind sparsity and dimensions of somatic mutation and cnv data matrices, we move on to perform biological enrichment of data. The stratification algorithm we propose implies functional scoring of the mutational profile (set of mutated genes per patient) as follows: a set of mutated genes is matched to different genesets of various biological context, such as T cell co-stimulation, regulation of cell cycle, etc. and geneset Z-score (GSZ) is estimated for each patient and functional gene set. The GSZ-score calculates the fraction of mutated genes of a patient that match to a given functional set minus the fraction of mutated genes in the set averaged over all patients and divided by the overall standard deviation. The mutational GSZ-score of a functional set is consequently high/positive for patients with mutations in the functional set exceeding its average. The output of this analysis returns for each patient a vector of GSZ-scores where each element refers to one functional set. In our application, we used functional sets from the gene ontology category biological process (GO BP). The method was implemented using its application in the R-package oposSOM (Loeffler-Wirth, 2015).

This step helps do reduce data sparsity as the resulting functionally enriched matrix contains correlation scores for every sample-geneset pair, as well as reduces the feature space from 459 genes to 84 genesets.

The above-mentioned clustering methods were taking as input patient-gene boolean matrix indicating mutation existence for a given patient in a given gene. The xGeneModel was integrating functional information within its algorithm. The following two methods use prior biologically stratified input.

4) KMeans Clustering on enriched data: This is a simple integration of KMeans Clustering (in Python's Scikit-Learn library (Pedregosa et al, 2011)) on the enriched data.

5) Deep Embedded Clustering on enriched data: DEC is a method that simultaneously learns feature representations and cluster assignments using deep neural networks. It learns a mapping from the data space to a lower-dimensional feature space in which it iteratively optimizes a clustering objective (Xie et al., 2016). Given n tumors with the feature vectors in space X with m dimension that should be grouped to k clusters with centers μ_i, instead of clustering the data in the initial space X, the data are mapped to the latent feature space Z. This is done by a nonlinear function

$$f_\theta : X \to Z,$$

where θ is a set of trainable parameters. A deep neural network can be used to implement f , because of its theoretical function approximation characteristics and the capabilities in learning features (Hornik, 1991). As said, DEC is an iterative method. In each iteration, the cluster centers μ_i, as well as parameters θ, are updated. The algorithm consists of two parts:

- Parameter initialization using a stacked autoencoder (SAE) (for θ) (Suk et al., 2015) and k-means algorithm (for centroids).
- Parameter optimization that calculates the auxiliary target distribution function and updates the parameters using minimization of the Kullback–Leibler divergence (KLD) (Rohani et al., 2020).

These two steps are iterated until the convergence. The convergence criterion is satisfied when the assigned clusters to samples in two subsequent iterations are changed in < 0.001 portion of data.

D. *Classification*

After obtaining clusters, each patient was assigned a label with its cluster assignment and passed on to the classification step. Here, data is split to train and test dataset, and applicable classifiers from a set of 10 (9 traditional ML classifiers from Python's Scikit-learn library and 1 custom DL classifier) are fit to the train data. These are KNN (K Nearest Neighbors), LDA (Linear Discriminant Analysis), GNB (Gaussian Naive Bayes), LR (Logistic Regression), SVC (Support Vector Classifier), DTC (Decision Tree Classifier), RF (Random Forest), BG (Bagging Classifier), AB (AdaBoost), MLP (Multi-Layered Perception). Each of the classifiers, except MLP, was trained with Grid Search Cross-Validation to pick the best parameters. MLP model had a fixed architecture:

$$FC \to ReLu \to FC \to ReLu \to FC \to Softmax$$

The chosen optimizer was Adam, with a learning rate of 0.0001, and the loss was calculated with Cross-Entropy Loss. In some cases, we also had to account for class imbalance by doing random oversampling. After training the fitted models were tested on the test dataset, and evaluated using performance measures, such as Precision, Recall, Accuracy, per-class F1, Balanced Accuracy, Cohen Kappa Score(Cohen, 1960), were reported.

RESULTS

E. *Clustering and Classification*

Clustering quality was measured using silhouette score, which measures the difference between the similarity of a tumor to its own cluster (cohesion) compared to its similarity to other clusters (separation). The similarity is measured using Euclidean distance. The value of this criterion ranges from 1 to +1. The below table (Tab. I) summarizes silhouette scores for each of the four clustering methods:

TABLE I
SILHOUETTE SCORES FOR CLUSTERING METHODS

Method	**Score**
Spectrum	-0.672
xGene	0.008
KMeans (on s.d.*)	0.042
DEC (on s.d.)	0.065

*s.d. refers to stratified data

In general, all of our methods showed scores far less than the best possible score, which is 1. However, in the case of mutational profile subtyping, the value of this criterion is not very descriptive of the biological quality of the clusters even for state-of-the-art methods. However it can be used to compare and contrast two different methods. As we can see, Spectrum method performs the worst in terms of clustering quality. Spectrum and xGene perform relatively worse than KMeans and DEC methods. This indicates that clustering quality is improved when stratified data is being clustered. The best performing is the DEC model of stratified data, the score of which is pretty close to the score of a slightly different architecture and a different application of the Deep Embedded Clustering method on breast cancer data (0.07)(Rohani et al., 2020). Considering the fact that colorectal cancer is much more heterogenic and highly variant from patient to patient compared to breast cancer, reaching a similar performance level validates the goodness and improvement of the model. For DEC assigned clusters, PCA plots were obtained. 2 component PCA plots showed visually identifiable 3 clusters, where

3 component PCA plot showed horizontal separation of the other clusters.

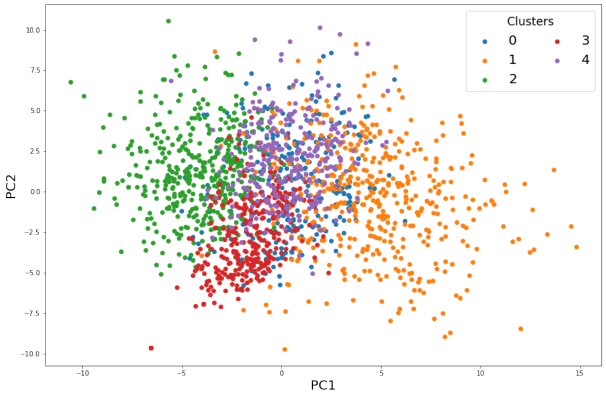

Fig. 7. 2 component PCA plot for DEC clusters

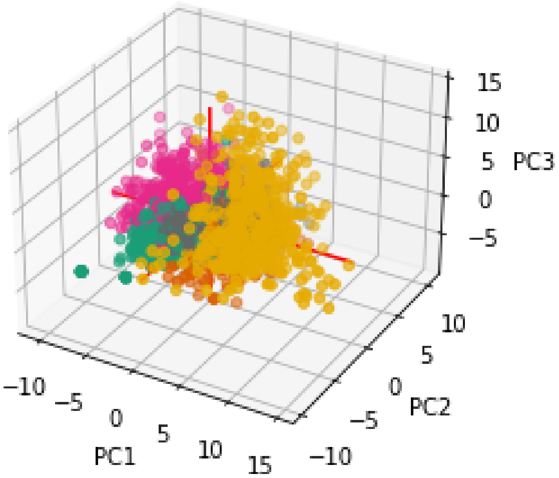

Fig. 8. 3 component PCA plot for DEC clusters

Classification was performed with two purposes. First, was to find the model which can serve as a predictor of a cluster assignment for an incoming record. This is needed in the clinical application of subtyping. The second purpose was, to have the best performing classifier in mind, to compare its performance given the different cluster assignments. This could also help for clustering quality validation. Logistic Regression and custom MLP models proved to perform relatively better than other models at classifying all cluster assignments. However, the performance measures for these classifiers varied depending on the input data and clustering method used for obtaining labels. because we had class imbalance, we picked F1 score, Balanced Accuracy, and Cohen Kappa score for evaluation, as these are the most descriptive metrics to look at in case of unbalanced data. xGene model labels along with non stratified data were the most poorly classified with best models. The best values are obtained using LDA with SVD (Singular Value Decomposition) as solver, and were 0.47 for Balanced Accuracy and 0.23 for Cohen Kappa score. Class imbalance here affects the performance as per-class F1 scores become close to 0 for some classes.

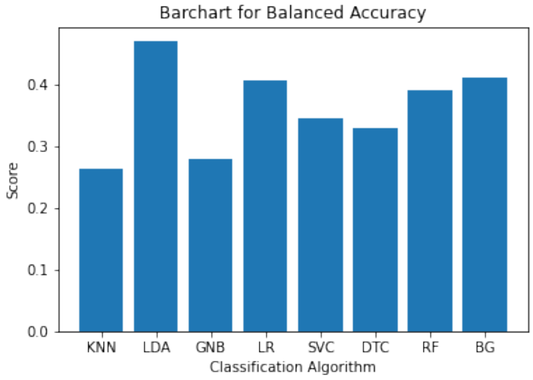

Fig. 9. Balanced Accuracy on xGene assigned labels

Spectrum assigned labels along with non stratified data provided better results. Best performing were Logistic Regression and Bagging Classifiers, with LR giving Balanced Accuracy of 0.89 and Cohen Kappa score of 0.86.

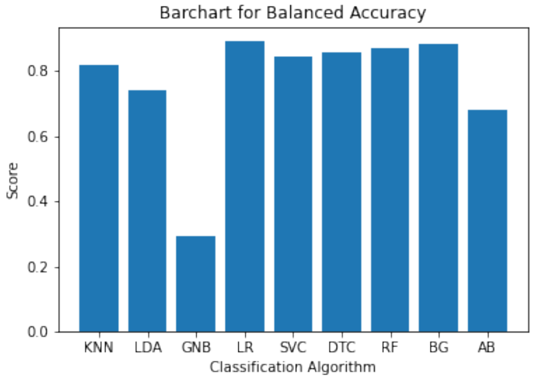

Fig. 10. Balanced Accuracy on Spectrum assigned labels

As anticipated, the clustering as well as classification were more accurate for the stratified data and clusters assigned to that data as labels. Here LR and MLP models performed the best. LR giving 0.92 balanced accuracy and 0.90 Cohen Kappa Score.

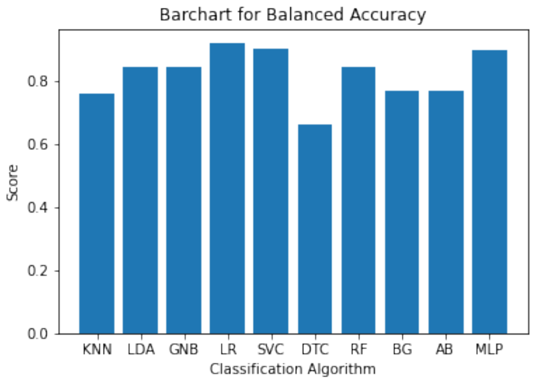

Fig. 11. Balanced Accuracy on KMeans assigned labels

Even better performance was reported by DEC assigned clusters on stratified data, with Balanced Accuracy of 0.93 and Cohen Kappa of 0.94.

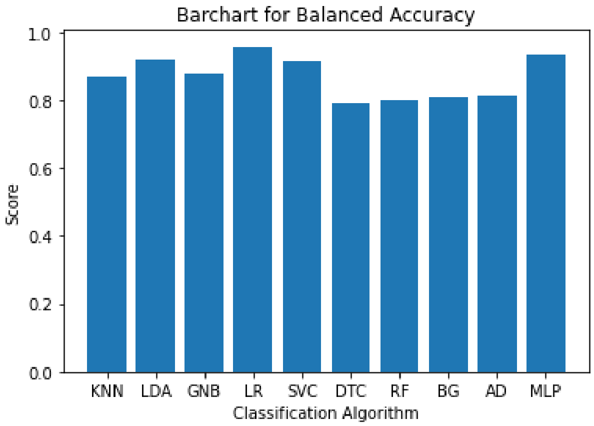

Fig. 12. Balanced Accuracy on DEC assigned labels

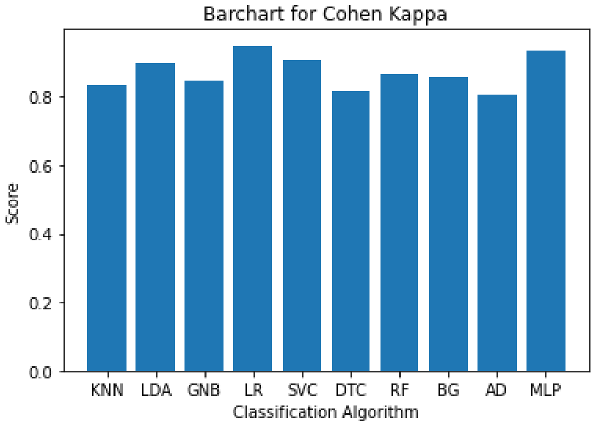

Fig. 13. Cohen Kappa Score on DEC assigned labels

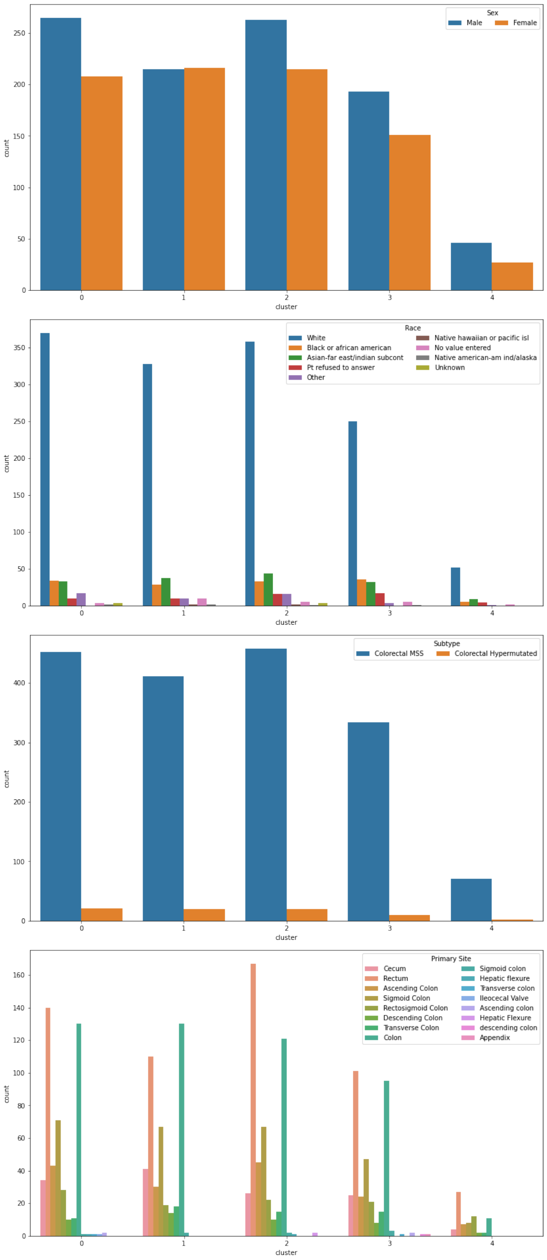

Fig. 14. Distribution of patients by covariates per cluster

Additionally, Matthews Correlation Coefficient was calculated for this method, as it is another method for evaluating multi-class classification, and, as some sources claim, can sometimes be competing with Cohen Kappa score (Delgado & Tibau, 2019). However, it still showed a high value - 0.93.

In all cases Logistic Regression was trained using L2 regularization as the penalty norm.

F. *Biological Evaluation*

1) Batch Effect Analysis: After we were able to achieve high predictive power with classification, we moved on to analyze each cluster separately according to various biological criteria. Firstly, it is possible in biological data clustering that sometimes clusters can form based on a variable that is not used as a feature during clustering. Such covariates can be the demographic variables, biological characteristics of the tumor, stage of the cancer development or any other confounding variables. To be sure our clusters were not we checked for batch effects accounting for Sex, Race, Subtype, Primary Site of Tumor, and Tumor Mutation Burden. None of the variables contributed to batch effect in any cluster assignment, so the given clusters were not affected by any covariates. The below figure shows batch effect analysis results for DEC clusters (Fig. 14), graphs for other covariates is provided in the Supplementary Material.

2) Kaplan Meier Survival Analysis: To check how survival differs in these clusters, we performed Kaplan Meier Survival Analysis. The curves were fitted per cluster, and the log-rank test was performed for each pair from the 5 clusters. Over (Fig. 15). As we also can see from the figure, the survival curves are pretty close to each other, and the log-rank test proved this by showing significant (p-value = 0.08 with $\alpha = 0.1$) difference only between clusters 1 and 4.

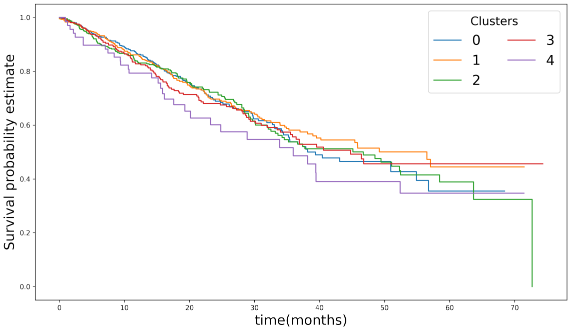

Fig. 15. Kaplan Meier curves by cluster (k = 5)

When the same analysis was done using k=10 clusters (Fig. 16), batch effects again were not reported and the survival curves were now showing more significantly divergent pairs, namely 0/1, 0/2, 0/5, 0/7, 2/6, 2/8, 2/9, 3/5, 5/6, 5/8, 5/9, however not all clusters were showing distinct survival patterns. This means the molecular subdivision of tumors does not necessarily infer clinical outcome differences.

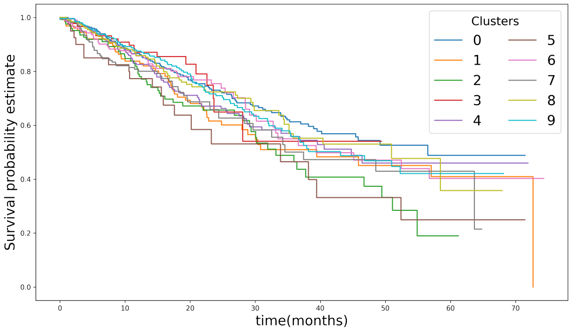

Fig. 16. Kaplan Meier curves by cluster (k = 10)

Hidden information under the curve of each cluster can be the treatment data that has been prescribed to each patient. These curves may be a result of the average of survival given a specific treatment in each cluster subgroup. So to be able to better evaluate the goodness of clusters this information need to be taken into account. However, treatment data was not available for this study,so this idea will be moved to further investigation using other datasets.

Conclusion:Findings and Future Work

Cancer, and specifically colorectal cancer, is a highly complex disease for which research to find personalized treatment options is crucial. With a growing pool of sequencing data, it is made available for researchers to explore the predictive potential of somatic mutations and copy number variations in the process of molecular subtyping of cancer. Our research firstly aimed to test and compare previously proposed promising methods that have not been applied to colorectal cancer data because of its biological complexity. We have tailored these methods to fit to our data to get appropriate results. Moreover, we proposed methods that combined theoretically and practically tested state-of-the-art concepts, such as biological enrichment and clustering using deep neural networks. These newly proposed combination methods were also compared with the other methods we tested, as well as the state-of-the-art results previously obtained on different datasets.

Our findings showed that considering the complexity of colorectal cancer mutational spectrum, biological stratification of data is an essential step to perform before subtyping analysis. This conclusion was a result of the evaluation of the four different clustering methods, as well as classification methods applied to each of the clustering labels. The models ran on non-stratified data had poorer silhouette score and smaller classification power. The best clustering algorithm was found to be the Deep Embedded Clustering method ran on GSZ-score stratified data, with 0.065 silhouette score for cluster quality and 93% balanced accuracy with Logistic Regression and MLP classifiers. Compared to the result from another DEC method developed and tested by Rohani et al. for breast cancer subtyping, our suggested model, which inherited many concepts from the latter but added and adjusted some of the algorithmic solutions and parameter choices, provided very similar results. Considering the fact that colorectal cancer is much more heterogenic and highly variant from patient to patient, getting this close to the performance of the algorithm on breast cancer data is a good milestone that was reached.

Another finding was that these optimized clusters did not show significantly different survival patterns. The reason for this can be the hidden specific treatment effect averaging. However, this hypothesis could not be checked because the used dataset did not include treatment data for the patients. This poses an ”to-do” item for the future work in this area when these algorithms are tested on other datasets. If this data is made available, performing separate survival analysis for each cluster may help identify treatments that significantly change survival pattern in a specific cluster and not the other. In clinical application, this analysis can serve as an efficient treatment prediction for a predefined subgroup of patients.

The drawback of this research has been the absence of justification for the choice of the cluster number. The analysis was run with k=5, which was chosen simply based on biological intuition. Algorithms to predict the most efficient number of clusters were suggesting to use 2 clusters, which was not going to be efficient for the analysis. So future work will include designing domain-specific algorithms that may be able to predict the number of clusters that will be biologically meaningful.

Future work may also include integrating more biological meaning into the clusters, such as functional analysis of clusters based on the most over-represented genes in them, and obtaining descriptive gene signature (set of mutated genes specific to each cluster). These signatures can be later used to model organisms representing each cluster and experimentally test the efficiency of a predicted treatment.

All of the methods may later be applied to other types of cancer as well. Pancreatic cancer is of special interest for us, as it is a common cancer type of colorectal metastasis.

The benefits of having this type of subtyping and classification are reduced costs and more targeted, more efficient and less toxic treatment for incoming cancer patients. This research, along with many others which have and are still trying to optimize solutions to this complex task, are of fundamental importance to modern precision medicine research and promise great potential when reached a level of confidence to be applied in the clinic.

Supplementary Information

All supplementary data and figures can be found here.

Acknowledgment

I would like to express my gratitude to my primary supervisor, Hans Binder, who guided me throughout this project.

References

[1] Vural, S., Wang, X., and Guda, C. (2016). Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. BMC Syst. Biol. 10:62. doi: 10.1186/s12918-016-0306-z

[2] Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W., and Quackenbush, J. (2018). Cancer subtype identification using somatic mutation data. Br. J. Cancer 118, 1492–1501. DOI: 10.1038/s41416-018-0109-7

[3] Zhang, W., Flemington, E. K., Zhang, K. (2018). Driver gene mutations based clustering of tumors: methods and applications. Bioinformatics (Oxford, England), 34(13), i404–i411. https://doi.org/10.1093/bioinformatics/bty232

[4] Rohani, N., Eslahchi, C. (2020). Classifying Breast Cancer Molecular Subtypes by Using Deep Clustering Approach. Frontiers in genetics, 11, 553587. https://doi.org/10.3389/fgene.2020.553587

[5] Nguyen, B., Fong, C., Luthra, A., Smith, S. A., DiNatale, R. G., Nandakumar, S., Walch, H., Chatila, W. K., Madupuri, R., Kundra, R., Bielski, C. M., Mastrogiacomo, B., Donoghue, M., Boire, A., Chandarlapaty, S., Ganesh, K., Harding, J. J., Iacobuzio-Donahue, C. A., Razavi, P., Reznik, E., Schultz, N. (2022). Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. Cell, 185(3), 563–575.e11. https://doi.org/10.1016/j.cell.2022.01.003

[6] Molinari, C., Marisi, G., Passardi, A., Matteucci, L., De Maio, G., Ulivi, P. (2018). Heterogeneity in Colorectal Cancer: A Challenge for Personalized Medicine?. International journal of molecular sciences, 19(12), 3733. https://doi.org/10.3390/ijms19123733

[7] Christopher R John, David Watson, Michael R Barnes, Costantino Pitzalis, Myles J Lewis, Spectrum: fast density-aware spectral clustering for single and multi-omic data, Bioinformatics, Volume 36, Issue 4, 15 February 2020, Pages 1159–1166, https://doi.org/10.1093/bioinformatics/btz704

[8] Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., … Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. Nucleic acids research, 32(Database issue), D258–D261. https://doi.org/10.1093/nar/gkh036

[9] Loeffler-Wirth H, Kalcher M, Binder H (2015). ”oposSOM: R-package for high-dimensional portraying of genome-wide expression landscapes on Bioconductor.” Bioinformatics.

[10] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011

[11] Xie, J., Girshick, R., and Farhadi, A. (2016). ”Unsupervised deep embedding for clustering analysis,” in International Conference on Machine Learning (Vienna), 478–487.

[12] Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. Neural Netw. 4, 251–257. DOI: 10.1016/0893-6080(91)90009-T

[13] Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A. D. N. (2015). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct. Funct. 220, 841–859. doi: 10.1007/s00429-013-0687-3

[14] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

[15] Delgado, R., Tibau, X. A. (2019). Why Cohen’s Kappa should be avoided as a performance measure in classification. PloS one, 14(9), e0222916. https://doi.org/10.1371/journal.pone.0222916