UNIVERSITÀ
DI TRENTO

Department of Information Engineering and Computer Science

Bachelor's Degree in
Computer Science

Final Dissertation

A machine learning approach for tumor staging using multi-omics data

Supervisor

Alberto Montresor

Student

Giulia Grotto

Academic year 2022/2023

Acknowledgements

...thanks to...

Contents

Abstract

In 2020 Fatima and Rueda created a method, called iSOM-GSN, where they transform ‘multi-omic’ data with higher dimensions onto a 2D grid, and afterward, they apply a CNN to predict disease states of various types [4].

This thesis shows how machine learning methods can be used for the diagnosis of serious diseases like tumors, but also that the same good result obtained by Fatima and Rueda can be achieved with an easier approach like random forest.

The data taken into account are contained in two datasets: TCGA Prostate Adenocarcinoma (PRCA) and TCGA Breast Invasive Carcinoma (BRCA). These repositories of multi-omics data encompassed crucial biological information, including mRNA-seq, DNA methylation, and CNA (copy number alteration) data.

The final goal of the project is to classify in three different classes the two types of tumors considered.

The project was carried out in collaboration with FBK (Fondazione Bruno Kessler), in particular with Data Science for Health (DSH) office. Their primary objective is to develop machine learning-based IT solutions that can be seamlessly integrated into the healthcare industry.

The work is divided into four main parts:

- Data Quality Check and Cleaning: this foundational phase is aimed at comprehending the intricacies of the provided data, with the goal of shaping it into a coherent and robust dataset for subsequent analysis.
- Dimensionality Reduction: deviating from conventional machine learning practices for dimensionality reduction, the choice is to incorporate the MutSigCV technique for feature selection, the same as [4]. Nonetheless, this decision was not without its challenges, as certain significant obstacles emerged during its implementation.
- Data Augmentation: confronting the complexities posed by imbalanced datasets, it was strategically employed the Synthetic Minority Over-sampling Technique (SMOTE). This strategy proved instrumental in rectifying data imbalances and enhancing overall predictive performance.
- Modeling and Visualization: the concluding stage involves the implementation of the random forest algorithm on the pre-processed datasets, now enriched through the preceding adjustments.

The results achieved are nothing short of remarkable, showcasing accuracy levels ranging from 95% to 99%. This collective effort underscored the assertion that even simpler methodologies could yield very good outcomes. Fatima and Rueda’s work demonstrated that while the iSOM-GSN approach was highly effective, the application of the random forest algorithm, bolstered by meticulous data preprocessing and augmentation, could achieve comparably impressive results.

1 Introduction

This thesis is based on biological data, so in the next section, it is provided a brief introduction to this argument and on project input data and the final goal. This project is coded with R. This choice have been made because it can offer several distinct advantages that make it a compelling choice. For example:

- Diverse Package Collection: R presents an extensive range of specialized packages designed for various machine learning and data analysis tasks. Some of these are randomForest, caret, performanceEstimation, that are the ones used here.

- Statistical Proficiency: R have a strong statistical foundation, so it excels in tasks requiring statistical analysis and modeling. This proficiency supports thorough exploratory data analysis and effective interpretation of results.

- Powerful Visualization: R includes robust visualization libraries enabling the creation of customizable and informative visualizations that reveal insights from data and model outcomes.

- Interdisciplinary Applicability: R has widespread use across various domains like biology, economics, and social sciences promotes interdisciplinary collaboration and knowledge exchange.

1.1 Brief introduction to DNA

DNA, or deoxyribonucleic acid, serves as the fundamental genetic blueprint of living organisms, encoding the essential instructions required for their growth, development, and functioning. DNA, in figure 1.1, is composed of subunits called nucleotides, which are the building blocks of its structure. Each nucleotide consists of three primary components: a phosphate group, a deoxyribose sugar molecule, and a nitrogenous base. The nitrogenous bases, namely adenine (A), thymine (T), cytosine (C), and guanine (G), establish the genetic code through their specific pairing interactions – adenine with thymine and cytosine with guanine - in order to create the typical DNA double-helix form.
RNA, or ribonucleic acid, is a vital molecule in the cellular realm, serving as a versatile intermediary between DNA and protein synthesis. Unlike DNA, RNA is usually single-stranded and contains uracil (U) as a base instead of thymine. It comes in several forms, including messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA), each playing a distinct role in conveying genetic information, translating it, and facilitating protein production within cells.

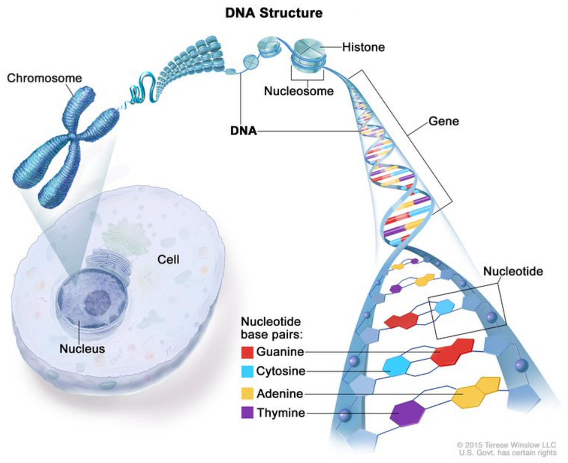


Figure 1.1: Structure of DNA [14]

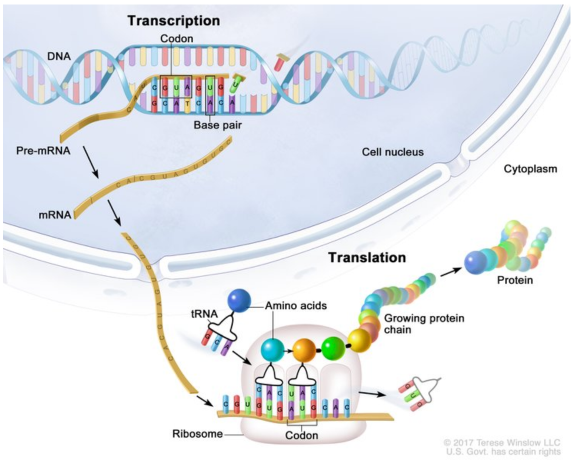


Figure 1.2: Central dogma [14]

The genetic information carried by DNA is organized into discrete units called genes. Genes are segments of DNA that encode instructions for synthesizing proteins, the molecules responsible for driving numerous cellular processes. The central dogma, in figure 1.2, of molecular biology outlines the flow of genetic information: DNA is transcribed into mRNA in a process catalyzed by the enzyme RNA polymerase. Subsequently, the mRNA carries the genetic code from the nucleus to the ribosomes in the cytoplasm, where it serves as a template for protein synthesis during translation.

Within the cell nucleus, DNA is organized into structures known as chromosomes. Chromosomes are thread-like structures composed of DNA and associated proteins. They contain multiple genes arranged linearly along their length. Humans, for instance, possess 46 chromosomes, including 23 pairs, where one chromosome of each pair is inherited from each parent.

DNA replication errors, mutagens like UV light, and imperfect DNA repair cause mutations. Genetic mutations drive differences between individuals and fuel evolution by providing genetic variation. Beneficial mutations spread through natural selection, adapting populations over time. Mutations vary in size and impact, including base substitutions, deletions, insertions, inversions, and translocations. They affect gene function, leading to gain or loss of molecular functions.[1] This mutation can also lead to cancer.

1.2 Input data

As mentioned before, the input data for this project is based in two different datasets:

- TCGA Prostate Adenocarcinoma (PRCA) [1]
- TCGA Breast Invasive Carcinoma (BRCA) [2]

The total number of samples for PRCA and BRCA is 500 and 817, respectively.

These repositories of multi-omics data encompassed crucial biological information, including mRNA-seq, DNA methylation, and CNA (copy number alteration) data, but also a medical record for each patient (sample). The next subsections contain a description of the three layers of data that are used in this thesis and how they are derived.

1.2.1 RNA-seq

RNA-seq (RNA-sequencing) is a technique that can examine the quantity and sequences of RNA in a sample. It analyzes the transcriptome, indicating which of the genes encoded in our DNA are turned on or off and to what extent.

After acquiring the RNA sample, the initial technique step involves transforming the targeted RNA population into complementary DNA (cDNA) fragments, creating what is known as a cDNA library. This conversion is achieved through reverse transcription, allowing the RNA to be integrated into a next-generation sequencing (NGS) workflow. Subsequently, the cDNA is fragmented, and adapters containing functional elements, such as amplification sequences to enable clonal amplification and the primary sequencing priming site, are appended to the fragment. After undergoing amplification, size selection, cleaning, and quality assessment steps, the cDNA library is subjected to NGS analysis, generating concise sequences representing all or portions of the original fragments. After library preparation, it is possible to employ the selected sequencing platform to sequence the cDNA library to the desired specifications. Once transcript data is generated, it is aligned with a reference genome, or if no reference exists, a de novo assembly is performed. After the alignment stage, it is time to focus on analyzing data [10].

1.2.2 DNA methylation

DNA methylation is an epigenetic mechanism involving the transfer of a methyl group onto the C5 position of the cytosine [11], in fact, is commonly found on CpG dinucleotides[3]. DNA methylation is a crucial component of epigenetic gene expression control. It involves specific modifications in different cell types. While it is reversible, it usually maintains stability during cell division.

[1] http://www.cbioportal.org/study/summary?id=prad_tcga

[2] http://www.cbioportal.org/study/summary?id=brca_tcga_pub2015

[3] Cytosine-Guanine base pairs.

High-throughput bisulfite sequencing is a widely trusted approach for measuring DNA methylation. This and similar methods offer the ability to measure methylation at the level of individual nucleotides. By converting unmethylated cytosines to thymines while keeping methylated cytosines unaffected, the subsequent step involves aligning reads with these C-to-T changes and counting C-to-T mutation to calculate the fraction of methylated bases. This process allows for a precise, comprehensive assessment of DNA methylation across the genome [1].

1.2.3 CNA

Copy number alterations (CNAs) occur due to changes to DNA structure that lead to the gain/amplification or loss/deletion of copies of DNA sections from a normal genome.[13] CNAs comprise deletions or amplifications of fragments of genomic material that are particularly common in cancer and play a major contribution to its development and progression [3].
Copy number alterations (CNAs) are explored by analyzing tumor cell genomes to identify regions with abnormal copy numbers. This method could offer insights into diverse cancer types through CNA analysis, but it demands handling substantial data volumes [5].

1.3 Final goal of the project

The final goal of the project is to classify into three different classes the two types of tumors considered. In the case of Prostate Adenocarcinoma, tumor grading is called Gleason Score. The Gleason Score is a number from 2 to 10 expressed like a sum of two grades in the range from 1 to 5. These two grades are numbers that the doctor assigns respectively to the two most predominant patterns in your biopsy. A Gleason score equal to or lower than 6 corresponds to a low-grade cancer, a Gleason Score of 7 is a medium-grade cancer, and a Gleason Score equal to or higher than 8 is a higher-grade cancer. However, in the case of Breast Invasive Carcinoma, tumor grading is called staging, and it goes from 0 to 4. These stages depend on the tumor size, the number of nearby lymph nodes with cancer, and whether the cancer has metastasized or not. Based on these parameters, there are sub-stages like 2A and 2B. In particular:

- on PRCA dataset: the model makes predictions, especially on patients with a Gleason Score 7 or 9, but for 7 ones it makes a distinction between 3+4 and 4+3.
- on BRCA dataset: the model makes predictions, particularly on patients with tumor stage II or III, but for stage II it makes a distinction between IIA and IIB.

2 Data pre-processing

Data pre-processing is required because all the files containing data tables need to be cleaned and well-formed. This process can be made in parallel for both datasets cause the original data tables for both types of tumor have the same structure. Both datasets require the creation of four data frames:

- patient info: all the information for each patient, like a clinical record that contains the present diagnosis and the clinical history. From that data frame we take as features only the patient ID and the tumor stage, that is the label used for the training of our model.
- rna-seq: each patient's sequenced mRNA gene expression, that considers about 20000 different genes.
- DNA methylation: methylation (HM450) beta-values for genes (about 16000).
- linear CNA: relative linear copy-number values for each gene (about 23000).

From the patient info dataset, are obtained only the patient ID and the tumor stage. The patient ID is useful to merge the two features obtained from this data frame with the other features, i.e. the best features selected from the other three data frames. Both for the RNA-seq data frame, and for the DNA methylation one, and for the CNA one, there are too many features that have to be analyzed, and a lot of these are useless. That's why a feature selection is performed in the same way for all of these three.

But before doing this, all four data blocks require to be structured according to certain specifications, i.e. features must be placed in columns, and the samples arranged in rows. In addition to these adjustments, data need to be "clean":

- NA values: they are an indicator of missingness, so the data in that position may have been lost. The chosen solution in this case is to replace that value with 0.
- 0 columns: there could be 0 columns among the features, and they are obviously useless, so the solution is to drop them.

2.1 Dimensionality Reduction

As said above, in all three data frames, containing the three levels of data used (mRNA-seq, DNA methylation, CNA), there are thousands of features, most of them useless. So the best solution is dimensionality reduction.

There are two ways to perform dimensionality reduction: feature extraction and feature selection.

Feature selection involves choosing a subset of features from an original set based on specific criteria to retain them relevant or not. This aids in data processing by eliminating redundant and irrelevant features. Effective feature selection enhances learning accuracy, speeds up learning, and simplifies results.

In contrast, feature extraction transforms original data into high-pattern recognition features, differing from the weak recognition abilities of the original data [6].

Taking a cue from the original paper [4], the used method is feature selection, with the help of the MutSigCV algorithm explained below.

Feature selection can be categorized into three different methods: filter methods, wrapper methods, and embedded methods.

The chosen method is a filter method, i.e. ranking features based on a statistical measure and selecting a subset of the top-ranked features for the model [6].

2.1.1 MutSigCV

The main objective of this algorithm is to identify genes that are significantly mutated in cancer genomes, using a model with mutational covariates (so it makes a ranking of the genes, which in this case are the features).

MutSigCV analyzes whole genome or exome[1] sequencing data from multiple samples, detecting genes mutated more frequently than expected by chance. While growing sample sizes were expected to enhance sensitivity and specificity in identifying cancer-driver genes, larger sample sizes introduce false positives by detecting overly mutable genes as implausibly cancer-related. This issue arises due to using an average mutation rate for a cancer type across the genome.

MutSigCV addresses this by considering patient-specific mutation rates, spectra, gene-specific mutation rates, gene expression, and replication times. This correction reduces false positives, especially in high mutation rate tumors [8].

In the abstract, the presence of a problem in this part of the project was introduced. The arose problem during the code drafting is that the MutSigCV algorithm is no longer available for use, neither in the online version nor on MATLAB as a function. The solution is to trust the MutSig made on the data directly from TCGA in both datasets. This choice has been made because the algorithm was executed specifically on the same data used for the research.

Despite it all, the first thirty genes for each level of data are selected from the output ranking of the algorithm to be used as features for the model.

2.2 Data Augmentation

After the dimensionality reduction, the three layers of data and the patient information are merged to form a unique data table with all the features that need to be considered.

At this point, an attempt at model training was made, but there was an overfitting problem caused by the data, which is too few and imbalanced, so there is a need for data augmentation and over/undersampling.

Imbalanced data is data in which observed frequencies are very different across the different possible values of a categorical variable. Basically, there are many observations of some type and very few of another type [7] (majority and minority class). So, the problem is that a machine learning algorithm could ignore or have poor performance in the minority class.

Data augmentation involves boosting data volume by creating fresh data points from existing ones. This encompasses making slight modifications to the data or employing machine learning models to craft new data points within the latent space of the original data, thereby expanding the dataset.

Data augmentation improves the performance and results of machine learning models by generating new and diverse instances for training datasets. Moreover, it reduces operation costs related to data collection avoiding the time spent on data collection and data labeling [12].

The technique used to balance and augment data is SMOTE.

2.2.1 SMOTE

SMOTE stands for Synthetic Minority Over-sampling Technique. It is a technique that enhances datasets by generating data points from existing ones. Essentially, SMOTE can be considered an upgraded form of oversampling or a dedicated method for augmenting data. [7] So, this technique can solve both our problems at the same time.

The pro of SMOTE is that it doesn't only duplicate samples to augment data volume, but creates new samples from the older ones to provide also new information for the model.

The SMOTE works by selecting nearby instances in the feature space and creating synthetic samples along a line connecting them. A random minority class example is chosen, and its k nearest neighbors (often k=5) are identified. From these neighbors, a random one is picked, and a synthetic sample is generated on a random point between the two in the feature space. This process can be repeated as needed. The technique recommends combining random undersampling of the majority class with SMOTE oversampling of the minority class to balance class distribution. However, the method's drawback is that it can lead to ambiguous samples if significant class overlap exists [2].

[1] The collection of exons, the protein-coding part of DNA.

An issue is that the SMOTE works with 2 classes: the majority class and the minority class, but there are three classes both in PRCA (Gleason Score 7:3+4, 7:4+3, 9) and in BRCA (Stage IIA, IIB, III). The solution is to build two data tables from the one that was merged before based on different classes. The result of this division is:

- for PRCA: one data table which contains all the samples, but instead of classifying differently the two types of Gleason Score 7, classifying all of them with 7 and the others with 9; and the other data table with only Gleason Score 7 patient, classified with 3 for the 3+4 class, and with 4 for the 4+3 class.
- for BRCA: one data table which contains all the samples, but instead of classifying differently the two types of Stage 2, classifying all of them with 2 and the others with 3; and the other data table with only Stage 2 patients, classified with 1 for the 2A class, and with 2 for the 2B class.

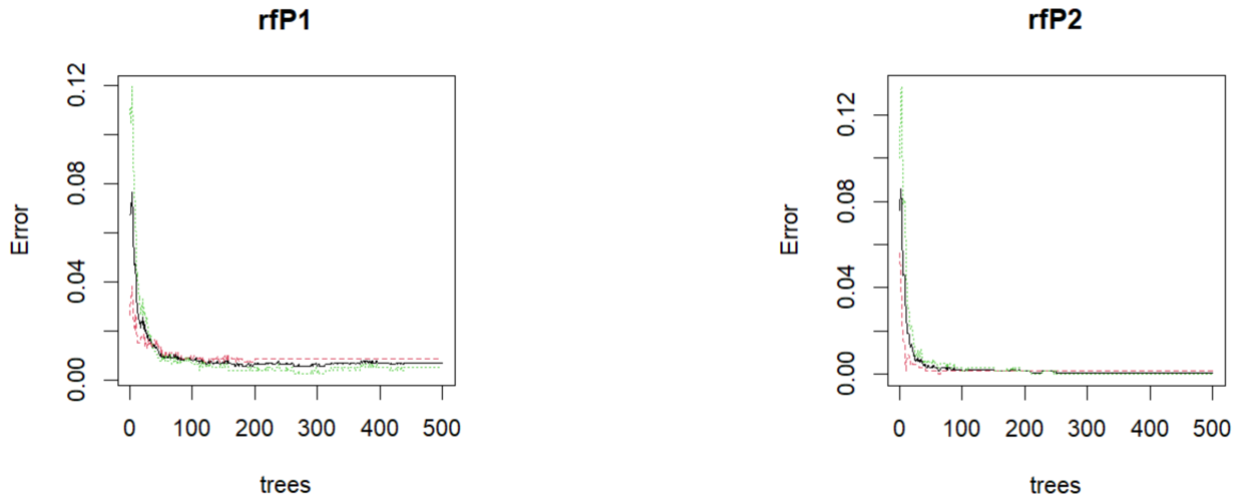
At the end of this operation and the SMOTE, the data tables for training and testing the model are ready.

3 The model

Now that data are well-formed, the model can be trained, so all data tables are split into 2 sets of data: training set and test set. The division between the training and test set follows the rules of holdout cross-validation, i.e. the samples are divided randomly for each split in a fixed division (in this case 70-30).

3.1 Random Forest

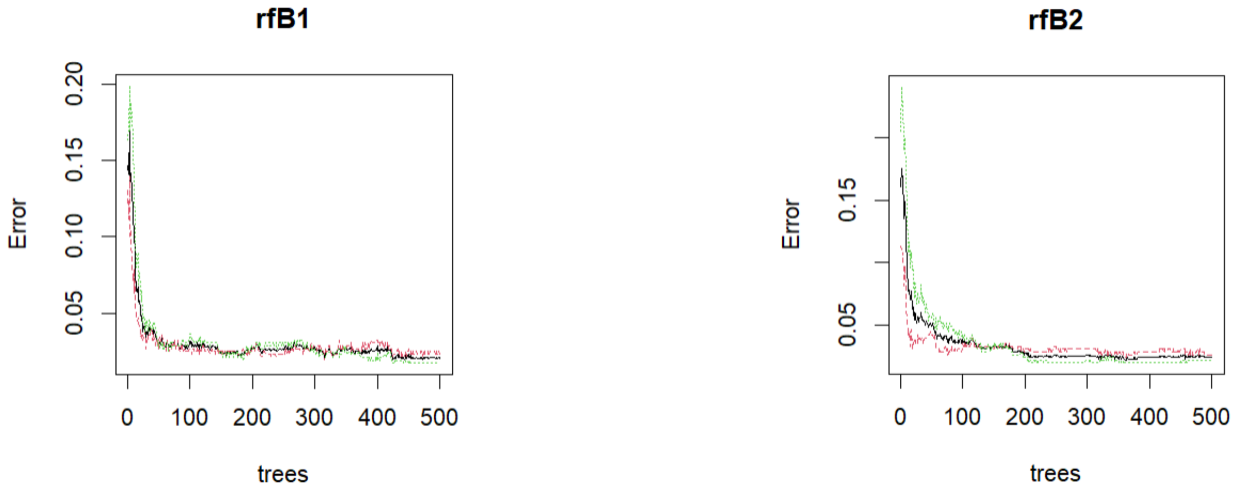
In 2001, Leo Breiman defined random forests as follows:
”A random forest is a classifier consisting of a collection of the tree–structured classifiers $\{\mathrm{h}(\mathrm{x},\Theta_k),\mathrm{k}=1,\ldots\}$, where the $\{\Theta_k\}$ are independent identically distributed random vectors, and each tree casts a unit vote for the most popular class at input x” [9].
In general, a Random Forest is an ensemble learning method used in machine learning. It combines multiple decision trees to enhance predictive accuracy and control overfitting. Each tree is trained on a different subset of the data and makes its prediction. The final output is determined by averaging the predictions of all trees. This approach enhances robustness and applies to both classification and regression tasks.
One key advantage of Random Forests is their ability to handle high-dimensional datasets with various types of features. By randomly selecting subsets of features for each tree, the algorithm maintains diversity and prevents over-reliance on any single feature. This mitigates the risk of overfitting and improves generalization to unseen data.
Despite their effectiveness, Random Forests can become computationally intensive for a large number of trees or features. Additionally, they may struggle with capturing intricate relationships in the data, which more complex models like gradient boosting methods could address.
In conclusion, Random Forests stand as a versatile and powerful tool in the machine learning arsenal. Their ability to combine simplicity with strong predictive performance makes them the best choice for this project.
The randomForest function, to be called, requires the number of trees needed to build the model. The choice fell on 500 trees, because, as you can see in figure 3.1 and 3.2, that is the best number of trees in order to minimize the error and maximize the accuracy of the model.



(a) Plot for the Gleason Score 7/9 classification model

(b) Plot for the 3+4/4+3 types of Gleason Score 7

Figure 3.1: These are the 2 plots for PRCA tumor



(a) Plot for the Stage 2/3 classification model

(b) Plot for the A/B types of Stage 2

Figure 3.2: These are the 2 plots for BRCA tumor

Like the split for cross-validation, the model training is also done on each data table. Therefore, four models have been designed:

1. classification between Gleason Score 7 and Gleason Score 9 (PRCA);
2. classification between 3+4 and 4+3 types of Gleason Score 7 (PRCA);
3. classification between Stage II and Stage III (BRCA);
4. classification between A and B types of Stage II (BRCA).

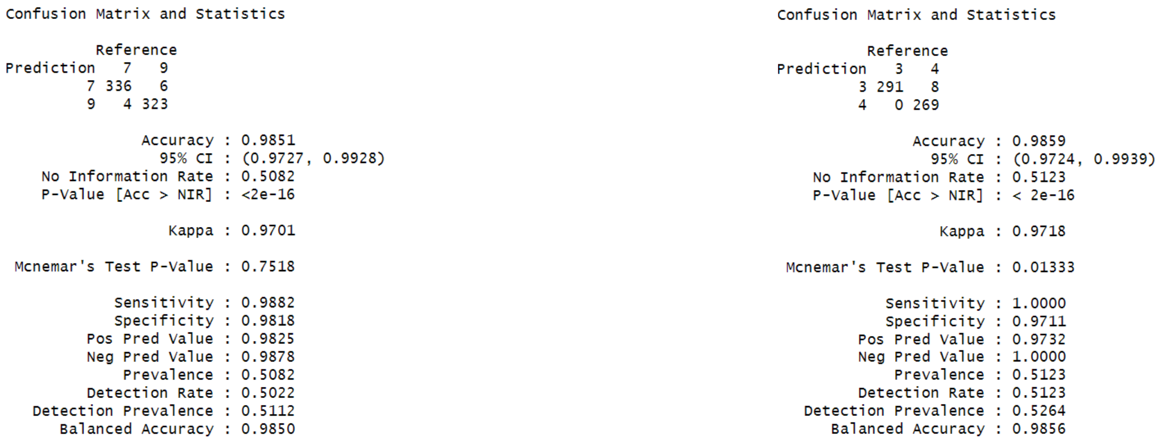
The idea is to take a sample and classify it first with the model trained on Tumor Grades (1, 3), and then, if the result is the grade with the 2 sub-grades, the sample is classified also by the second model (2,4).

4 Results

After model training, there is the test phase. In the test phase, a prediction on the test set with a visualization of the results through a confusion matrix.
Evaluating a trained machine learning model involves testing its predictive capabilities on a separate test dataset. This dataset contains new, unseen examples that the model hasn't encountered during training. By applying the model to this data, it generates predictions or classifications. These predictions are compared against the actual outcomes in the test dataset to measure the model's accuracy and effectiveness. It is crucial to avoid using the test data for any adjustments or tuning, as it ensures the model's true generalization ability. The process ensures the model's effectiveness in making accurate predictions on new, unseen data.

4.1 Results obtained

The final results are displayed in a confusion matrix.
A confusion matrix is a visual representation used in machine learning to evaluate how well a model's predictions match actual outcomes. It is organized as a table with rows for actual classes and columns for predicted classes. The matrix highlights four types of outcomes: true positives (accurate positive predictions), true negatives (accurate negative predictions), false positives (incorrect positive predictions), and false negatives (incorrect negative predictions). By analyzing these outcomes, the confusion matrix provides insights into a model's performance and helps in calculating metrics like accuracy, precision, recall, and F1-score. It assists in understanding where a model excels and where it needs improvement. In figures 4.1, 4.2, there are 4 confusion matrices, one for each model. On top of that, there are prediction statistics, including accuracy and balanced accuracy.
Accuracy is used to evaluate a classification model's performance. In classification, the model assigns data to specific classes. Accuracy measures the proportion of correct predictions made by the model, calculated by dividing the number of correct predictions by the total predictions. While useful, accuracy may not reflect a model's effectiveness with imbalanced data. Complementing accuracy with metrics like precision, recall, and F1-score offers a more comprehensive understanding of the model's performance, especially in scenarios where class distribution varies, and that is why there is also balanced accuracy. While accuracy measures overall correctness, balanced accuracy adjusts for imbalanced datasets. It averages class-specific accuracies, making it a better choice when handling varying class sizes and ensuring a fairer evaluation.



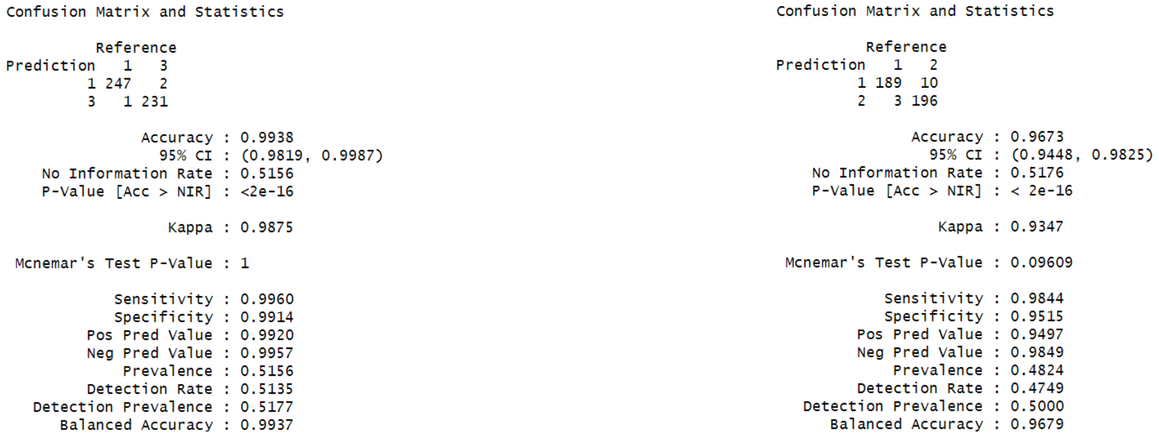
```
Confusion Matrix and Statistics

          Reference
Prediction   7   9
         7 336   6
         9   4 323

               Accuracy : 0.9851
                 95% CI : (0.9727, 0.9928)
    No Information Rate : 0.5082
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9701

 Mcnemar's Test P-Value : 0.7518

            Sensitivity : 0.9882
            Specificity : 0.9818
         Pos Pred Value : 0.9825
         Neg Pred Value : 0.9878
             Prevalence : 0.5082
         Detection Rate : 0.5022
   Detection Prevalence : 0.5112
      Balanced Accuracy : 0.9850
```

(a) Confusion matrix for the Gleason Score 7/9 classification model

```
Confusion Matrix and Statistics

          Reference
Prediction   3   4
         3 291   8
         4   0 269

               Accuracy : 0.9859
                 95% CI : (0.9724, 0.9939)
    No Information Rate : 0.5123
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9718

 Mcnemar's Test P-Value : 0.01333

            Sensitivity : 1.0000
            Specificity : 0.9711
         Pos Pred Value : 0.9732
         Neg Pred Value : 1.0000
             Prevalence : 0.5123
         Detection Rate : 0.5123
   Detection Prevalence : 0.5264
      Balanced Accuracy : 0.9856
```

(b) Confusion matrix for the 3+4/4+3 types of Gleason Score 7

Figure 4.1: These are the 2 confusion matrices for PRCA tumor



```
Confusion Matrix and Statistics

          Reference
Prediction   1   3
         1 247   2
         3   1 231

               Accuracy : 0.9938
                 95% CI : (0.9819, 0.9987)
    No Information Rate : 0.5156
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9875

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9960
            Specificity : 0.9914
         Pos Pred Value : 0.9920
         Neg Pred Value : 0.9957
             Prevalence : 0.5156
         Detection Rate : 0.5135
   Detection Prevalence : 0.5177
      Balanced Accuracy : 0.9937
```

(a) BRCA: plot for the Stage 2/3 classification model

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 189  10
         2   3 196

               Accuracy : 0.9673
                 95% CI : (0.9448, 0.9825)
    No Information Rate : 0.5176
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9347

 Mcnemar's Test P-Value : 0.09609

            Sensitivity : 0.9844
            Specificity : 0.9515
         Pos Pred Value : 0.9497
         Neg Pred Value : 0.9849
             Prevalence : 0.4824
         Detection Rate : 0.4749
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.9679
```

(b) BRCA: plot for the A/B types of Stage 2

Figure 4.2: These are the 2 confusion matrices for BRCA tumor

5 Conclusion

As can be seen, the results and outcomes clearly demonstrate that the model is capable of ensuring precision within the range of 95% to 99%. This indicates that the random forest-based model's performance is on par with the achievement of iSOM-GSN, a creation of Fatima and Rued. It's evident that the model's predictive prowess falls within an impressive spectrum, showcasing accuracy levels akin to those witnessed in the work done by Fatima and Rued on their iSOM-GSN project. This parity in outcomes emphasizes the potency of the random forest approach, underlining its ability to yield results that are in line with the innovative techniques employed by the aforementioned individuals. This substantiates the effectiveness of the random forest methodology, mirroring the outcomes realized through the inventive techniques harnessed by the individuals in their notable iSOM-GSN work. In essence, the model's performance echoes the impressive achievements witnessed in the iSOM-GSN project, showcasing the credibility of the random forest framework in delivering comparable and noteworthy results, close to the approaches adopted by Fatima and Rued.

Bibliography

[1] A. Akalin. *Computational Genomics with R.* Chapman and Hall/CRC, 2020.

[2] J. Brownlee. Smote for imbalanced classification with python. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification.

[3] L. Esteves, F. Caramelo, I.P. Ribeiro, et al. Probability distribution of copy number alterations along the genome: an algorithm to distinguish different tumour profiles. *Scientific Reports*, 10, 09 2020.

[4] N. Fatima and L. Rueda. iSOM-GSN: an integrative approach for transforming multi-omic data into gene similarity networks via self-organizing maps. *Bioinformatics*, 36(15):4248–4254, 05 2020.

[5] S. Franch-Expósito et al. Cnapp, a tool for the quantification of copy number alterations and integrative analysis revealing clinical implications. *eLife*, 9, jan 2020.

[6] C. Jie, L. Jiawei, et al. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018.

[7] J. Korstanje. Smote. *Towards Data Science*, 2021.

[8] M. Lawrence et al. Mutsigcv (v1). https://www.genepattern.org/modules/docs/MutSigCV, 2013.

[9] Y. Liu, Y. Wang, and J. Zhang. *Information Computing and Applications.* Springer, Berlin, Heidelberg, 2012.

[10] R.J. Mackenzie. Rna-seq: Basics, applications and protocol. *Technology Networks*, 2018.

[11] L.D. Moore, T. Le, and G. Fan. Dna methylation and its basic function. *Neuropsychopharmacology*, 07 2012.

[12] D. Shah. The essential guide to data augmentation in deep learning. https://www.v7labs.com/blog/data-augmentation-guide, 2022.

[13] E.S. Tan et al. Copy number alterations as novel biomarkers and therapeutic targets in colorectal cancer. *Cancers (Basel)*, 04 2022.

[14] Terese Winslow. https://visualsonline.cancer.gov/.