XIII Argentine Congress of Bioinformatics and Computational Biology

XIII International Conference of the Iberoamerican Society of Bioinformatics

III Annual Meeting of the Ibero-American Artificial Intelligence Network for Big BioData

Machine Learning to Medical imaging study for cancer prognosis

Marcio Aparecido Bulla Junior[a] , Fabiano Yokaichiya[a] , Israel Tojal[b]
Universidade Federal do Paran´a- Brazil [a], AC Camargo - Brazil[b]

INTRODUCTION

The use and role of medical imaging technologies in clinical oncology has expanded greatly from being primarily a diagnostic tool to a more central role in the context of individualized medicine in the last decade. In this context, the objective of this research consists of the following parts:

1. Understanding and development of machine learning tools for analyzing tomography images for cancer diagnosis, based on the Martin Carrier-Vallieres master's thesis [1]. The objective is to produce the software, developed in the thesis in Mathlab, using Python language;
2. Use of machine learning tools to study mutated and non-mutated epidermal growth factor receptor (EGFR) associated with magnetic resonance tomography images of brains with cancer development;
3. Study the relationship between magnetic resonance tomography images of breasts with cancer, associated with germline mutations in the BRCA1 and BRCA2 genes that predispose to common human malignancies, mainly breast tumors and ovaries.

MATERIALS AND METHODS

In this study, we used quantitative data and MRI images obtained from The Cancer Imaging Archive (TCIA) [2]. The TCIA is a widely used oncology research database that provides free access to a wide range of information about cancer patients, including medical images such as x-rays, CT scans and MRIs, and quantitative data such as clinical and treatment. The use of quantitative data from TCIA magnetic resonance images allowed us to apply machine learning techniques to analyze quantitative information from these images, which enabled the development of our study. Additionally, we also used data obtained from Polak's article [3]. This article is a complete and detailed review of breast cancer, including information on epidemiology, risk factors, diagnosis and treatment. Incorporating data from this article highlights our ability to seek out and utilize relevant and reliable data sources. Furthermore, it demonstrates our concern with providing accurate and up-to-date information on the subject in question. The tools used in this research are composed of Python, Jupyter Lab, virtualenv and Git. These tools are widely used in Data Science projects and offer great flexibility and resources for performing data analysis. In addition, several Python libraries are used, such as numpy, pandas, scipy, seaborn, matplotlib, yellowbrick, scikit-learn and dtreeviz. These libraries offer a wide variety of functions and resources for manipulating, analyzing and visualizing data, making the research development process much more efficient and productive.

Programming Softwares

Python is a high-level programming language, very versatile and with a large, active community, which makes it easier to find solutions and integrate new libraries. Jupyter Lab is a web interface that allows us to write, run and document code interactively. Virtualenv is a tool that allows the creation of virtual environments for project development, ensuring compatibility between the libraries used. Git is a version control system that allows team collaboration and versioning control in the project.

Hypothesis Test

The Mann Whitney U function from the SciPy package was used to calculate the P-value between quantitative data, which enabled accurate and efficient statistical analysis. The P-value is a crucial measure in hypothesis testing, as it indicates the probability that the observed result occurs or is even more extreme, if the null hypothesis is true. The smaller the P-value, the stronger the evidence against the null hypothesis, indicating that the samples come from different distributions. The SciPy package is a Python library widely used in diverse fields, including computer science, engineering, physics, mathematics, and statistics, and provides a wide variety of scientific computing algorithms and techniques. The library is known for being easy to use, with clear and comprehensive documentation, which made using the Mann Whitney U function even more viable and accessible.

Correlation

Spearman correlation is a statistical measure that evaluates the monotonic relationship between two variables. It differs from linear correlation because it does not assume that the relationship between variables is linear, but only that there is a monotonic relationship. By using the Pandas correlation method to calculate the Spearman correlation, we were able to evaluate the relationship between the pre-selected features in the hypothesis test. This allowed us to identify whether there is any relationship between the variables, which may be relevant for the research and analysis of the results. The Spearman correlation is especially useful in cases where the variables are ordinal or when there are outliers, as these situations can affect the validity of the linear correlation, but not that of the Spearman correlation. By evaluating the correlation between features, we can use the results to understand the relationship between variables, which is important for making informed decisions about which features to include or exclude in research.

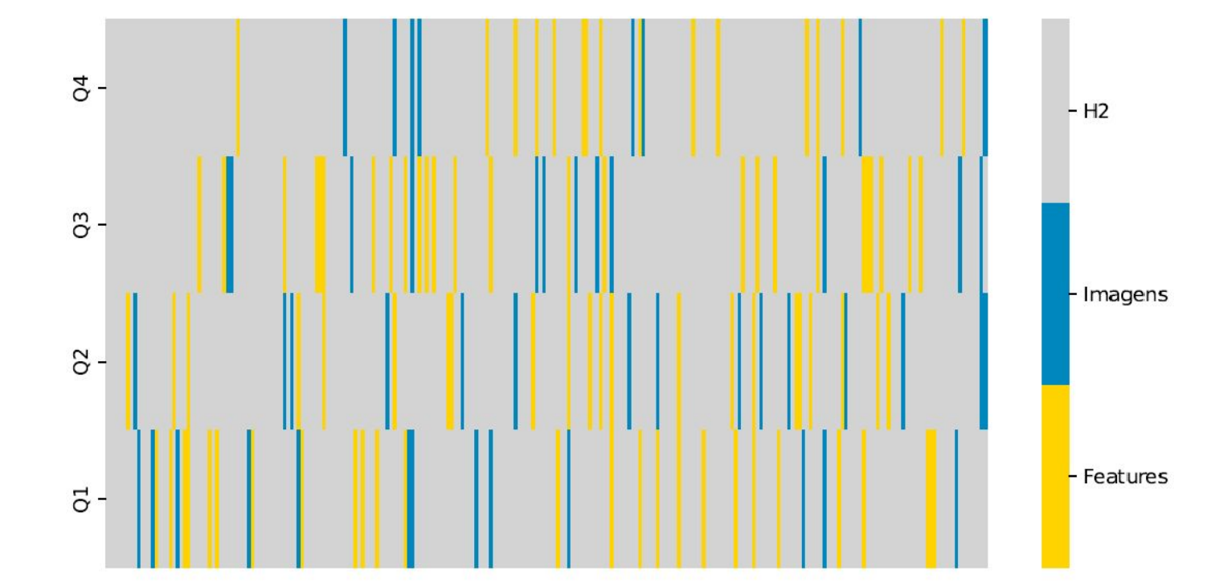
Machine Learning Models

We used two machine learning models, Decision Tree and Random Forest, in our study. Both are classification models that allow us to predict the category a given entry belongs to. The Decision Tree is based on creating a series of questions that are asked about the data, allowing categories to be inferred from the answers. It is a simple and intuitive technique, but it may have difficulties in generalization if the questions are not appropriate. Random Forest is a more robust model that uses multiple decision trees to predict the category of an input. By using multiple trees, the Random Forest is able to handle cases of under-adjustment and variations in data, which results in better overall performance. Both models are available in the scikit-learn library, which is widely used in machine learning research. We chose to use these models because of their efficiency and ease of use, and because they are widely known and used in the machine learning community.

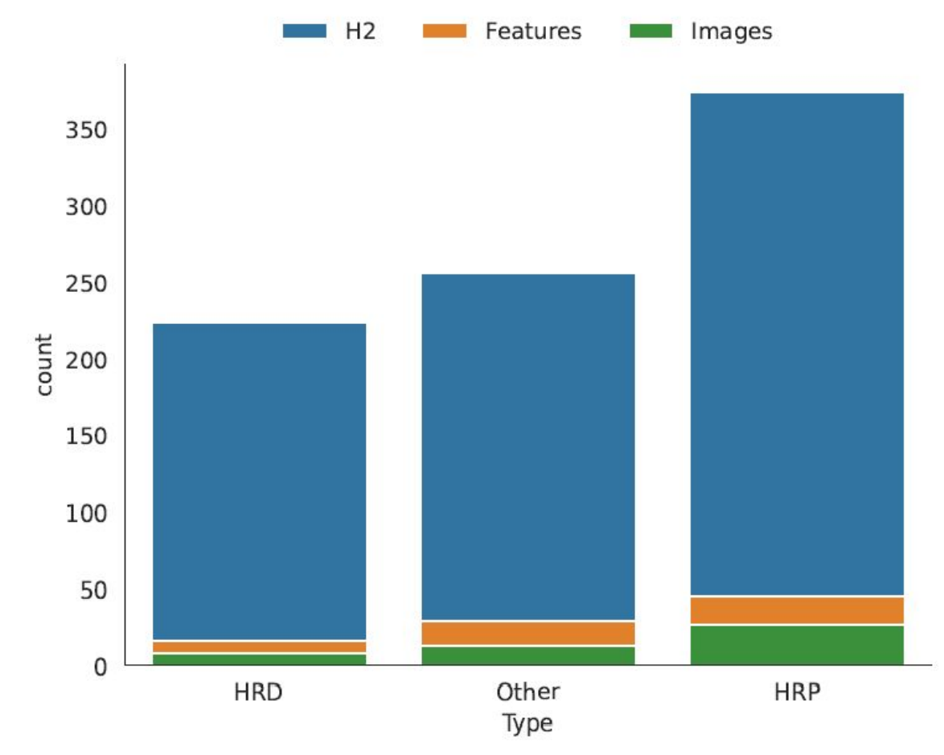
Data visualization

The Matplotlib, Seaborn, Yellowbrick, and Dtreeviz libraries are powerful tools for creating graphs and visualizations in Python. Matplotlib is a 2D plotting library that allows you to create a wide variety of graphs, including lines, bars, histograms, scatter plots, and more. It is a versatile tool that allows you to customize various aspects of graphs, including choosing axes, scales, legends and much more. Seaborn is a library based on Matplotlib that provides a more intuitive and visually appealing interface for creating graphs. Additionally, Seaborn includes a number of functions for creating more specialized charts, such as distribution plots, violin plots, regression plots, and more. Yellowbrick is a visualization library for machine learning that allows the visualization of performance metrics, model validation, among other aspects related to machine learning. It is a useful tool for evaluating the efficiency of models and for identifying problems such as overfitting or underfitting.

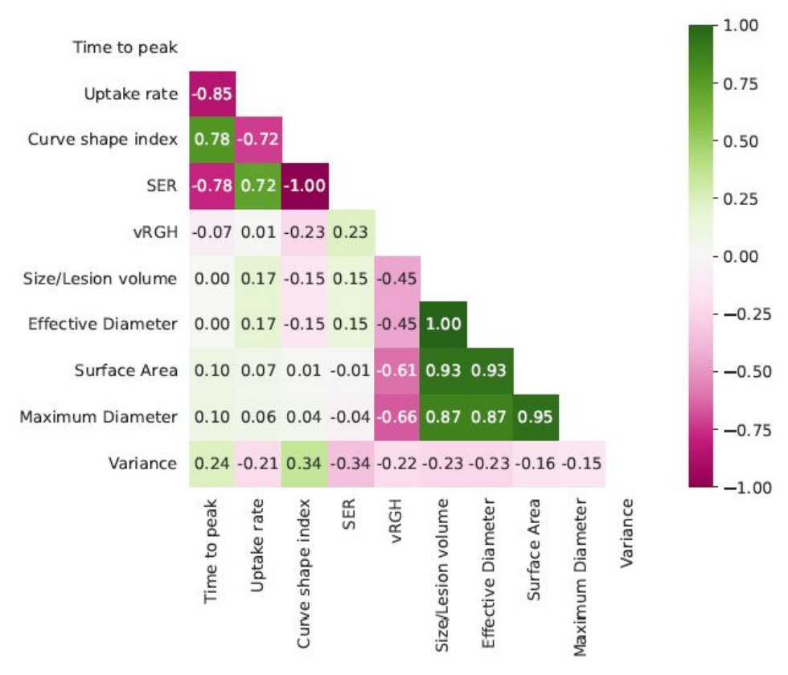
RESULTS



Samples separated into quantiles where in the yellow samples there is availability of features (quantitative data) and images and in the blue there are only images and gray there is only H2.



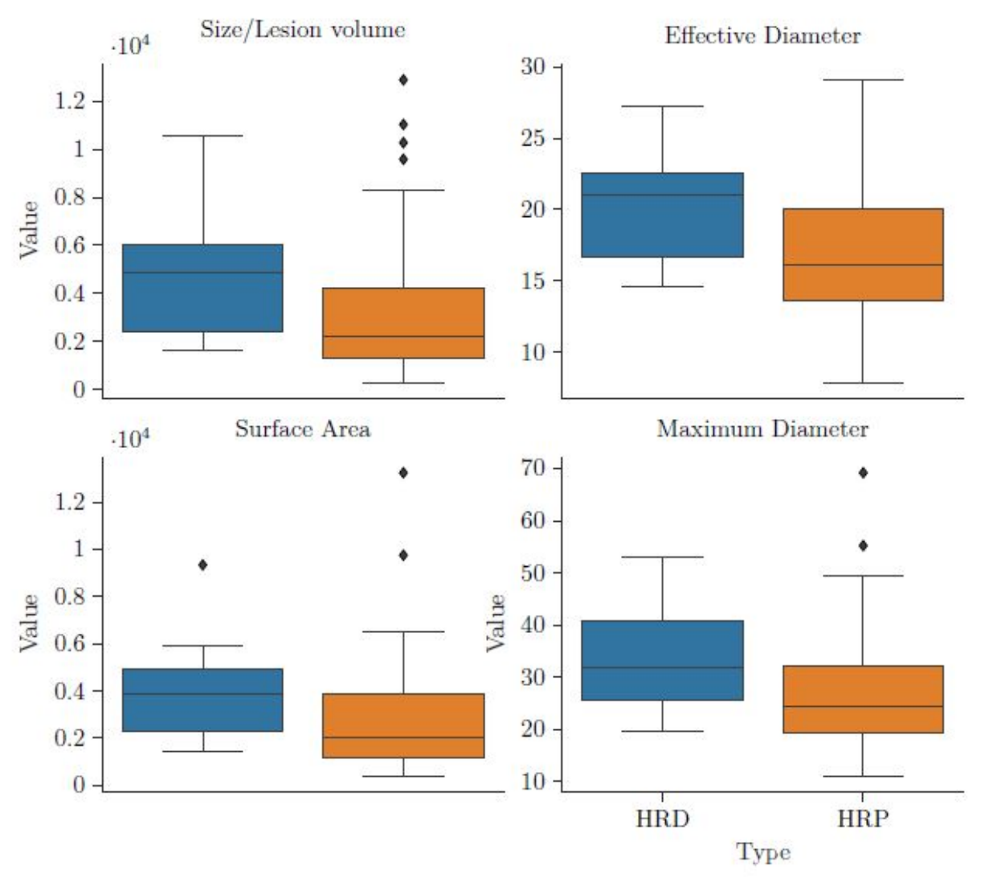
Proportion of classified samples



Heatmap of features with P(value) < 0, 05

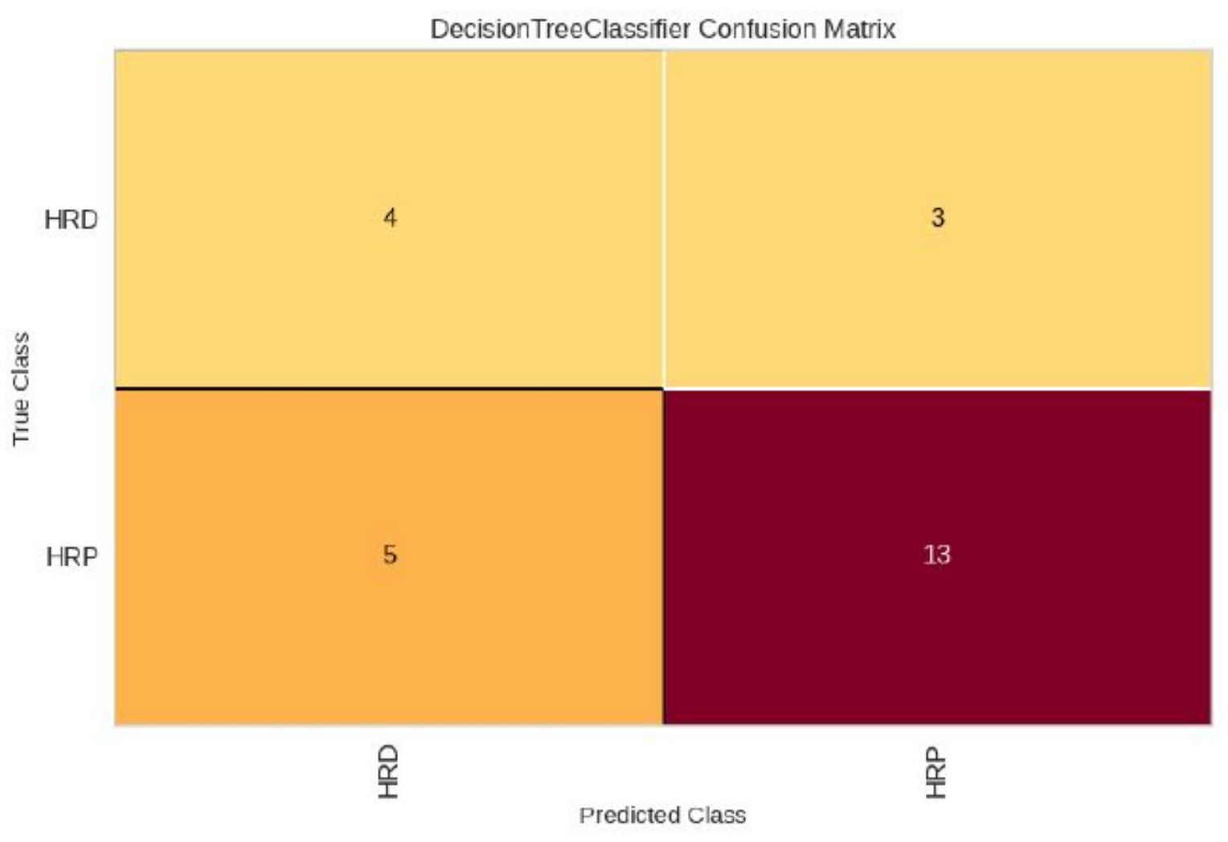
Family	*Feature*	P_{value}
Size	Size/Lesion volume	0,009814
	Effective Diameter	0,009814
	Surface Area	0,024171
	Maximum Diameter	0,021246
Kinetic Curve Assessmen	Time to peak	0,008359
	Uptake rate	0,018634
	Curve shape index	0,048192
	SER	0,048192
Enhancement Texture	Variance	0,004926
Morphology	vRGH	0,080787

P(value) of selected features



Box plot of Size features

Model Results



Decision tree

CONCLUSIONS

In summary, this study explored the application of machine learning methods for prognosis of HRD and HRP based on features extracted from MRI images. A relevant discussion in this context is the possibility of expanding the sampling universe, considering the extraction of features from other images using tools such as PyRadiomics [4]. However, it must be emphasized that although this approach could increase the amount of data, it would not necessarily solve the fundamental problem of data amount. To obtain substantial results, it would be necessary to rely on more comprehensive databases, such as those from hospitals, which offer more rigorous control over the origin of images and detailed clinical information. Furthermore, it is important to highlight that many researches face significant challenges due to the scarcity of samples. This is reflected in the inadequate application of bootstrap methods and the lack of adequate separation between training and validation samples, which can result in data leakage and, consequently, distorted results. A common limitation is the availability of an insufficient number of cases, making it unfeasible to apply machine learning methods for cancer prognosis with acceptable performance. Ultimately, this study highlights the need for comprehensive data collection and rigorous trait selection to support future research in this field. Only with these advances will we be able to explore the full potential of machine learning techniques in the medical field and, eventually, contribute to significant improvements in the diagnosis and treatment of cancer. To promote the transparency and replicability of this study, all source code created, along with the necessary data and resources, was made available in the public GitHub repository [5]

REFERENCES

[1] Carrier-Vallieres, M. FDG-PET/MR *imaging for prediction of lung metastases in soft-tissue sarcomas of the extremities by texture analysis and wavelet image fusion. Master thesis. [2] SHARMA, A. TCGA Breast Phenotype Research Group Data sets (TCGA-Breast-Radiogenomics). 2022. [3]* POLAK, P. et al. A mutational signature reveals alterations underlying deficiente homologous recombination repair in breast cancer. *Nature Genetics*, Springer Science and Business Media LLC, v. 49, n. 10, p. 1476–1486, ago. 2017. [4] GRIETHUYSEN, J. J. van et al. Computational radiomics system to
decode the radiographic phenotype. *Cancer Research*, American Association for Cancer Research (AACR), v. 77, n. 21, p. e104–e107, out. 2017. [5] BULLA, M. *EGFR*. 2022. Available: <https://github.com/AC4UFPR/TCGA-BRCA>.