Research, Society and Development, v. 10, n. 12, e444101220561, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v1012.20561

Mbl-2 gene polymorphisms in pediatric Burkitt lymphoma: an approach based on machine learning techniques

Polimorfismos do gene MBL-2 no linfoma de Burkitt pediátrico: uma abordagem baseada em técnicas de aprendizagem de máquina

Polimorfismos del gen MBL-2 en el linfoma de Burkitt pediátrico: un enfoque basado en técnicas de aprendizaje automático

Received: 09/13/2021 | Reviewed: 09/17/2021 | Accept: 09/24/2021| Published: 09/26/2021

Jonathan Wagner de Medeiros
ORCID: https://orcid.org/0000-0002-9679-273X
University of Pernambuco, Brazil
E-mail: jonathanmedeiros.biomed@gmail.com
Anthony José da Cunha Carneiro Lins
ORCID: https://orcid.org/0000-0002-7153-841X
Catholic University of Pernambuco, Brazil
E-mail: thonylins@gmail.com
Oluwarotimi Williams Samuel
ORCID: https://orcid.org/0000-0003-1945-1402
Chinese Academy of Sciences, China
E-mail: timitex92@gmail.com
Elker Lene Santos de Lima
ORCID: https://orcid.org/0000-0002-7171-7418
University of Pernambuco, Brazil
E-mail: elkerlene@yahoo.com.br
Maria Luiza Tabosa de Carvalho Galvão
ORCID: https://orcid.org/0000-0002-4593-7636
University of Pernambuco, Brazil
E-mail: mluizatcgalvao@gmail.com
Bárbara Oliveira Silva
ORCID: https://orcid.org/0000-0003-4158-5650
University of Pernambuco, Brazil
E-mail: barbaraoliveirabiomol@gmail.com
Giwellington Silva Albuquerque
ORCID: https://orcid.org/0000-0001-6610-963X
University of Pernambuco, Brazil
E-mail: giwellington@hotmail.com
Luísa Priscilla Oliveira de Lima
ORCID: https://orcid.org/0000-0002-4119-0491
University of Pernambuco, Brazil
E-mail: Luh.lima.ll6@gmail.com
Maria Tereza Cartaxo Muniz
ORCID: https://orcid.org/0000-0001-9498-5223
University of Pernambuco, Brazil
E-mail: tereza.cartaxo@upe.br

Abstract

Introduction: Burkitt lymphoma belongs to the group of non-Hodgkin lymphomas. Although curable in 80% of less advanced stages, it presents in advanced stages in about 75% of cases in Brazil's Northeast region, requiring urgent and intensive care in the early stages of treatment. *Objectives:* therefore, this study aimed to verify the participation of MBL-2 gene polymorphisms in the development of Burkitt lymphoma. *Methods:* In this article, computational approaches based on the Machine Learning technique were used, where we implemented the Random Forest and KMeans algorithms to classify patterns of individuals diagnosed with the disease and, therefore, differentiate them from healthy individuals. A group of 56 patients aged 0 to 18 years, with Burkitt lymphoma, from a reference hospital in the treatment of childhood cancer, was evaluated, together with a control group consisting of 150 samples, all of which were tested for exon 1 polymorphisms and the MBL2 gene -221 and -550 regions. *Results:* At first, an unsupervised classification was performed, which identified as two the number of groups that best represent the data present in our database, reaching 72.81% accuracy in the separation of patients and controls. Then, the supervised classification was performed, where the classifier obtained a 70.97% success rate, being possible to reach 75% accuracy in the best GridSearch configuration

when performing a cross validation. *Conclusion:* It was not yet possible to conclude about the participation of the evaluated polymorphisms in the development of the BL, however the computational techniques used proved to be very promising for carrying out studies of this nature.
Keywords: Machine learning; Burkitt lymphoma; MBL-2; Polymorphisms.

Resumo

Introdução: O linfoma de Burkitt pertence ao grupo dos linfomas não Hodgkin. Embora curável em 80% dos estágios menos avançados, se apresenta em estágios avançados em cerca de 75% dos casos no Nordeste brasileiro, necessitando de cuidados urgentes e intensivos nas primeiras fases do tratamento. *Objetivos:* afim de se obter mais informações sobre esta patologia, este trabalho teve como principal objetivo verificar a participação do gene MBL-2 no desenvolvimento do linfoma de Burkitt. *Métodos:* Neste artigo, foram implementadas abordagens computacionais baseadas na técnica de Aprendizado de Máquina, para a qual utilizamos os algoritmos Random Forest e KMeans para classificar padrões de indivíduos diagnosticados com a doença e, com estes, diferenciá-los de indivíduos saudáveis. Foi avaliado um grupo de 56 pacientes com linfoma de Burkitt, de 0 a 18 anos e um grupo controle composto por 150 amostras de indivíduos, todas testadas para polimorfismos do exon 1 e das regiões -221 e -550 do gene MBL2. *Resultados:* A classificação não supervisionada identificou como dois o número de grupos que melhor representam os dados presentes em nosso banco de dados, alcançando 72,81% de acerto na separação de pacientes e controles. Em seguida, foi realizada a classificação supervisionada, onde o classificador obteve uma taxa de sucesso de 70,97%, sendo possível atingir 75% de acerto na melhor configuração do GridSearch ao realizar uma validação cruzada. *Conclusão:* Neste estudo não foi possível concluir sobre a participação dos polimorfismos avaliados no desenvolvimento do LB, entretanto as técnicas computacionais utilizadas se mostraram bastante promissoras para realização de estudos desta natureza.
Palavras-chave: Aprendizagem de máquina; Linfoma de Burkitt; MBL-2; Polimorfismo.

Resumen

Antecedentes: El linfoma de Burkitt pertenece al grupo de los linfomas no Hodgkin. Aunque curable en el 80% de los estadios menos avanzados, se presenta en estadios avanzados en aproximadamente 75% de los casos en el noreste de Brasil, requiriendo atención urgente e intensiva en las primeras etapas del tratamiento. *Objetivos:* de esta manera, este estudio tuvo como objetivo verificar la participación de polimorfismos del gen MBL-2 en el desarrollo del linfoma de Burkitt. *Métodos:* En este artículo utilizamos enfoques computacionales basados en la técnica de Machine Learning, para lo cual se utilizaron los algoritmos Random Forest y KMeans para clasificar patrones de individuos diagnosticados con la enfermedad y, con ellos, diferenciarlos de individuos sanos. Se evaluó un grupo de 56 pacientes con linfoma de Burkitt, de 0 a 18 años, de un hospital de referencia para el tratamiento de cáncer infantil, y un grupo de control que constaba de 150 muestras de individuos, todos analizados para exón 1 y polimorfismos. 221 y -550 del gen MBL2. *Resultados:* Inicialmente se realizó una clasificación no supervisada, que identificó como dos el número de grupos que mejor representan los datos presentes en nuestra base de datos, alcanzando un 72,81% de precisión en la separación de pacientes y controles. Luego, se realizó la clasificación supervisada, donde el clasificador obtuvo una tasa de éxito del 70,97%, siendo posible alcanzar el 75% de acierto en la mejor configuración de GridSearch al realizar una validación cruzada. *Conclusión:* En este estudio aún no se pudo concluir sobre la participación de los polimorfismos evaluados en el desarrollo del BL, sin embargo las técnicas computacionales empleadas resultaron ser muy prometedoras para la realización de estudios de esta naturaleza.
Palabras clave: Aprendizaje automático; Linfoma de Burkitt; MBL-2; Polimorfismo.

1. Introduction

Burkitt lymphoma (BL) is a type of non-Hodgkin's lymphoma of mature B cells, malignant and extremely aggressive, presenting the highest cell proliferation rates among all neoplasms, with a doubling time between 24 to 48 hours. It represents almost half of all childhood lymphoma cases, with a higher incidence rate in caucasian and male children (Aydin et al., 2019; Derinkuyu et al., 2016; Swerdlow et al., 2016). Its main characteristic is the presence of mature B cells and monomorphic lymphocytes, presenting reciprocal translocation involving the *MYC* proto-oncogene (Hecht & Aster, 2000; Vardiman et al., 2008). Currently, three clinical forms of BL are considered: endemic, sporadic (non-endemic) and the form associated with immunodeficiency. Although these forms of BL are histologically identical and have similar clinical behavior, they have different epidemiological, clinical and genetic characteristics (Freedman et al., 2018).

Regarding the factors involved in the genesis of BL, genetic-based mechanisms (such as reciprocal translocation involving the *MYC* proto-oncogene) and the participation of infectious agents (Hsu & Glaser, 2000), mainly by the *Epstein-Barr Virus* (EBV), can be highlighted (Molyneux et al., 2012). In addition, other genetic factors have been associated with genetic

susceptibility, such as polymorphisms of the promoter genes of interleukin-10 and *Tumor Necrosis Factor* (TNF), especially in children without EBV infection (White, 2004). The primary deficiency of *Mannose-Binding Lectin* (MBL), usually caused by polymorphisms in the *MBL-2* gene (Kilpatrick, 2002a), can lead to immunodeficiencies and increased susceptibility to various infectious diseases (Da Cruz et al., 2013), cancers of lymphoid origin and autoimmune diseases (Martín-Mateos & Piquer Gibert, 2016).

In the promoter region of the *MBL-2* gene, among others, there are two well-studied polymorphic regions: -550 H / L (rs11003125) and -221 X / Y (rs7096206), both resulting from the exchange of guanine for cytosine (G□C) (Bouwman et al., 2006). For the structural region, there are three point mutations that have been described in exon-1 of the *MBL-2* gene: in the 52 codon (Arg → Cys, D allele), 54 codon (Gly → Asp, B allele), and 57 codon (Gly → Glu, C allele), resulting in the exchange of amino acids (Bouwman et al., 2006; Moslem et al., 2015). These variants are collectively referred to as "O" for the mutant allele and "A" for the wild allele (Martín-Mateos & Piquer Gibert, 2016). Carriers of the O allele show reduced MBL expression, both those with heterozygosis (A / O) and especially homozygosis (O / O) (Harrison et al., 2012).

The polymorphic sites of the promoter region are associated with different serum MBL levels independent of the variant alleles, being closely linked. Due to the imbalance of the link, these polymorphisms combine to form a limited number of only seven or eight haplotypes (Petersen et al., 2001). Among them, the HYA and LYA haplotypes are most often associated with high plasma concentrations of MBL (Soltani et al., 2014), and LXA, HYO and LYO are associated with low concentration (Bouwman et al., 2006). In this group, the X allele variant is the one that most negatively affects the MBL serum production (Hansen et al., 2004). Consequently, the determination of the Y / X polymorphism in the promoter region is important for a functional reading of MBL (Mendonça et al., 2010). In the promoter region of the *MBL-2* gene there is also another polymorphic site in the +4 region, also associated with a decrease in MBL levels, that represents the P/Q *loci* (Madsen et al., 1995). According to Boldt et al. (2006), haplotypes are associated with progressively lower concentrations of serum MBL in the following sequence: HYPA > LYQA > LYPA > LXPA >> HYPO = LYPO = LYQO (Boldt et al., 2006).

The search for new data analysis methodologies that are efficient and at the lowest possible cost has intensified, and it is precisely in this context that Machine Learning has been inserted. The term “Machine Learning” (ML) can be used to refer to algorithms that give computers the ability to learn without being explicitly programmed (learning from experience) (Van Der Aalst, 2016). To learn and adapt, a model is built from input data (instead of using fixed routines) and the constantly evolving model is used to make predictions or make decisions that are considered to be the most correct according to its experience (Van Der Aalst, 2016).

Therefore, the present study proposed another method, which is not as traditional as statistical methods used to analyze polymorphism data. The main objective is to adapt and use computational models based on ML with the purpose of verifying the participation of *MBL-2* gene polymorphisms in the susceptibility to the development of BL, and mainly to be able to classify patients and controls efficiently in the future.

2. Methodology

A total of 56 patients, with BL, aged 0 to 18 years between 2012 and 2017 were examined. Patients were recruited by spontaneous demand, both in the prospective and in the retrospective follow-up, at Hospital Universitário Oswaldo Cruz (HUOC). The clinical and biological data of the patients were collected by consulting the medical records of the Pediatric Oncohematology Center of HUOC (CEONHPE). Meanwhile, the control group is formed by a total of 150 randomly selected individuals in the same age group as the study group (patients), with no history of cancer, provided by the DNA Bank of the Human Molecular Genetics Laboratory of the Genetics Department of the Federal University of Pernambuco. The data collected

were: staging, recurrence, gender, age, treatment time, patient status (alive or dead).

It must be highlighted that the study design and experimental protocols were approved by the Human Research and Ethics Committee of the Hospital Complex, HUOC / PROCAPE, under the number CAAE: 02044612.3.0000.5198.

2.1 Genetic polymorphisms analysis

Somatic genetic material extracted from peripheral blood collected in an ethylenediamine tetraacetic acid - EDTA tube, followed by processing, DNA extraction using the Mini salting out technique, according to Miller (1988) (MWer et al., 1988), and stored at -20°c, was used to study the gene polymorphisms, in the prospective follow-up patients. DNA extracted from paraffinized material was used to evaluate the retrospective segment, using the QUIAGEN® Kit QIAamp® DNA FFPE Tissue (*QIAamp® DNA Mini and Blood Mini Handbook. Sample & Assay Technologies*, [s.d.]), according to the manufacturer's specifications. Subsequently the single nucleotide polymorphisms (SNPs) in the promoter (-550 G / C and -221 C / G) and exon-1 of *MBL-2* gene were amplified by the Rotor Gene 6000 real-time Polymerase Chain Reaction (PCR) methodology (Corbett Research, Sydney, Australia). The fluorophore SYBR Green® was used to test the exon-1 polymorphism (A/O) in accordance with Hladnik et al. (2002) (Hladnik et al., 2002), whereas the promoter region polymorphisms were detected by using Taqman® SNP (Single Nucleotide Polymorphism). PCR conditions for genotyping are described in Table 1.

Table 1: Data for the determination of MBL gene polymorphisms.

Polymorphisms	Primers and probes	PCR cycling	Amplicons	References
***MBL* Exon-1**	forward 5´-AGGCATCAACGGCTTCCCA-3´ Reverse 5´-CAGAACAGCCCAACACGTACCT-3´	95°C – 10 min 95°C – 15 sec 60°C – 1 min 40x Melt Range from 60° to 95°C increase of 0.2°C at each step 90sec pre-melt break 8 sec pause at each step	Allele A Allele O	(Hladnik et al., 2002) (adapted)
***MBL* -221**	forward 5´-GCACGGTCCCATTTGTTCTCA-3´ Reverse 5´-GCGTTGCTGCTGGAAGACTATAAA-3´ Allele Y, 5´-FAM-CATGCTTTCCGTGGCAGMGB- 3´ Allele X 5´-VIC-CATGCTTTCGGTGGCAG-MGB-3´	50°C - 2min 95°C – 10 min 92°C - 30 sec 50x 60°C - 60 sec	Allele Y Allele X	(*Division of Cancer Epidemiology and Genetics - National Cancer Institute*, 2018a). available in: http://snp500cancer.nci.nih.gov.] (adapted)

MBL -550	forward 5´-CCAACGTAGTAAGAAATTTCCAGAGA-3´ Reverse 5´-CAACCCAGCCCAGAATTAACTG-3´ Allele L VIC 5´-CCTGTCTAAAACACC-MGB-3´ Allele H 5´-FAM-AGCCTGTGTAAAAC-MGB-3´	50ºC - 2 min 95ºC – 10 min 92ºC - 30 sec 50 x 60ºC - 60 sec	Allele L Allele H	(*Division of Cancer Epidemiology and Genetics - National Cancer Institute*, 2018b). available in: http://snp500cancer.nci.nih.gov.] (adapted)

Source: Hladnik, U., Braida, L., Boniotto, M., Pirulli, D., Gerin, F., Amoroso, A., & Crovella, S. (2002). Single-tube genotyping of MBL-2 polymorphisms using melting temperature analysis. *Clinical and experimental medicine*, *2*(2), 105–108.
Division of Cancer Epidemiology and Genetics—National Cancer Institute (nciglobal,ncienterprise). (2018a, janeiro 1). [CgvHomeLanding]. https://dceg.cancer.gov/

2.2 Experimental setup

In this study, implementations of models based on Machine Learning techniques were carried out in order to analyze patterns related to groups of patients and controls, to find new insights about the development of BL. For this purpose, the code development environment available by Google Colab[1] was used with the Python[2] Scikit-learn library[3]. In order to verify whether there is a statistically significant difference between the expression groups, Fisher's exact test was used with the aid of the GraphPad Prism® V5.0 program. Then the first actions aimed to analyze the database, identifying the distribution of patient and control data. Then, a Principal Component Analysis (PCA) was performed (Niitsuma & Okada, 2007), aiming to reduce dimensionality and selecting relevant attributes. The main reasons why the dimensionality is as small as possible are: measurement cost and classifier precision. When the feature space contains only the most salient characteristics (which has better explanatory capacity), the classifier will be faster and will take up less memory. When the set of training examples is not very large, a small space of characteristics can tackle the curse of dimensionality and provide small error rates for the classifier (Watanabe, 1985). Following that, the analysis was divided into unsupervised and supervised analysis (Sathya & Abraham, 2013).

When the database has a large amount of data, the best way is to divide this set into three parts: training, validation and testing. When the data set is reduced, the most suitable is to use the resampling technique, which is used to approximate the validation set through the reuse of observations from the set used in training. This is the *k-fold* cross-validation technique, which consists of the random division of the training bench in k equal parts. Then, k-1 will compose the training data for the adjustment of models, and the other part will be reserved for the estimation of its performance, repeating this process until all parts have been used both in training and in model validation. This is expected to increase the accuracy of these estimates (Hastie et al., 2009; Kuhn & Johnson, 2013). For this reason, our database was divided as follows: 70% of the data was used for training and 30% for testing.

For the unsupervised analysis, the K-means algorithm was used (Xu & Wunsch, 2005). This algorithm uses the "*clustering*" technique, which is the process of dividing the data set by similarity, where individuals in one group are more similar

[1] https://colab.research.google.com
[2] https://www.python.org/
[3] https://scikit-learn.org/stable/

to each other than with individuals in other groups, that are called “clusters” (Jain & Dubes, 1988; Sharma, 2019). Starting with the clustering idea, the use of this algorithm is mainly based on the creation of a model, which when adding new data, this data will be automatically inserted in a given cluster. From this, it is possible to deduce its characteristics by the similarities with the components of that same cluster. In order to improve the dissimilarity between data points in the low-dimensional embedding space, we used the dimensional reduction technique *t-sne* (t-distributed stochastic neighbor embedding), in which preservation is non-linear, unlike Multidimensional Scaling (MDS) or PCA (Li et al., 2017; Van der Maaten & Hinton, 2008). This has many applications, being able to aid in the diagnosis or to determine a prognosis of a patient, for example. In order to interpret and validate an analysis of the obtained clusters, a silhouette technique was performed (Rousseeuw, 1987).

For the supervised technique, the Random Forest (RF) algorithm was used to analyze the data pattern, the same way it was done by Lins et al., (2017). RF is a supervised learning algorithm that creates multiple decision trees and combines them to obtain a more accurate and stable prediction (J. C. da Silva, 2018). Then, with the purpose of increasing the algorithm's predictive power and making the forecasting model more flexible (Kuhn & Johnson, 2013), two adjustment parameters (hyperparameters) were introduced. A parameter relates to the number of trees built by the algorithm before making a decision or making an average of predictions (Estimators) and the other parameter indicates the depth that trees should or could reach (Max Depth). These parameters were assigned different settings, as shown in table 2.

In this study, the cross-validation technique was also implemented (Refaeilzadeh et al., 2009). This technique uses the partitioning of the dataset into subsets (in our case, k = 10 folds) by randomly separating a subset for testing and the others for training. This process is repeated k times, until all parts are used for training and testing, thus more reliably evaluating the classifier predictive power.

Several configurations were tested with a *GridSearch* algorithm, to try to identify the best configuration to be applied to the supervised technique (Bergstra & Bengio, 2012), and the results expressed through a confusion matrix and accuracy, precision and recall metrics (Bland, 2015). To assess the efficiency of class separation, the Area Under the Curve - Receiver Operating Characteristic (AUC-ROC curve) was used, which is very useful to evaluate domains in which there is a large disproportion between classes (Khan & Rana, 2019; Prati et al., 2008), as shown in graphs 11 and 12.

Table 2: Hyperparameters and settings.

Parameters	Settings				
Estimators	5	10	15	20	100
Max Depth	2	5	7	9	10

Settings used for each hyperparameter used
Source: Authors

3. Results

3.1 Biological characteristics analysis

56 patients with BL were analyzed, 64% male and 36% female. The age group of the individuals who participated in the study was from 0 to 18 years old, (mean age of 6.7 years at diagnosis), with a standard deviation of 6.4295 ± 6.9705. However, 44% were under 5 years of age (Table 3). The main site of tumor involvement was the abdomen, in 80% of cases, and about 7% had mandibular/jaw involvement. 70% of patients were cured, with a period of more than 5 years without the disease.

Table 3: Biological characteristics of pediatric patients (n = 56) with BL treated at CEONHPE / HUOC.

Variables (Measurable unit)		Amount	%
Gender	Male	36	64
	Female	20	36
Age at diagnosis	< 5 years old	25	44
	5 – 9 years old	19	34
	≥ 10 years old	12	22

Source: Authors.

150 individuals formed the control group, randomly recruited, aged 0 to 18 years old and with no history of cancer. 77 of these were male (51.3%) and 73 were female (48,66%).

The proportional difference between genders in this study was approximately 2 males to 1 female in the patients group, and approximately 1 male to 1 female in the control group (p = 0.1161).

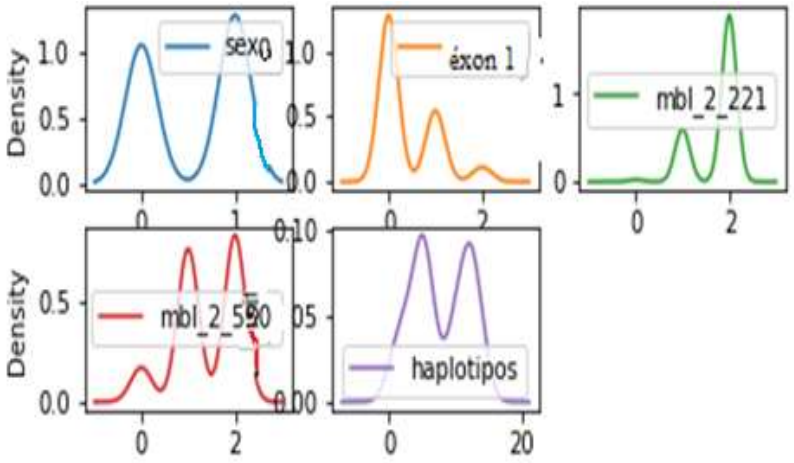
Analyses of the *MBL-2* gene polymorphisms were performed and the results are shown in table 4. We created graph 1 that summarizes the distribution of components with greater variability within the base. Among the polymorphisms present in the general study population (patients and controls), a total of 15 combinations were obtained, as described in Figure 2. Meanwhile, Figure 3 shows the comparison between the patient and control groups in terms of haplotypes.

Table 4: Result of the analysis of *MBL-2* gene polymorphisms. Differences in the amount of polymorphisms observed in the MBL-2 gene of patients and controls in total numbers (n) and percentage (%).

Genotypes ***MBL-2***	**Patients** **n= 56 (%)**	**Controls** **n= 150(%)**
Éxon 1		
AA	33 (58,92)	103 (68.66)
AO	21 (37.5)	37 (24.66)
OO	2 (3.57)	10 (6.66)
-221		
YY	43 (76.78)	112 (74.67)
YX	13 (23.22)	36 (24)
XX	0 (0.0)	2 (1.33)
-550		
LL	29 (51.79)	68 (45.33)
LH	23 (41.07)	66 (44)
HH	4 (7.14)	16 (10.67)

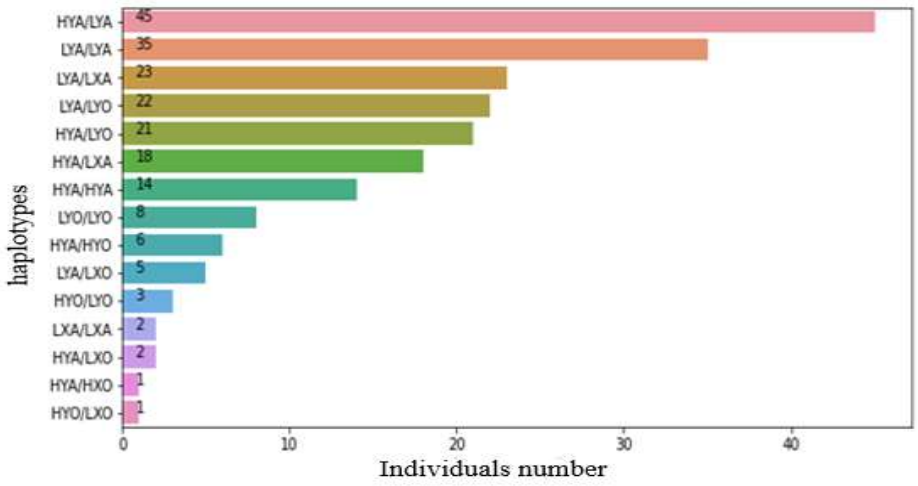
Source: Authors.

Figure 1: variation between some components of the base in a summarized form.



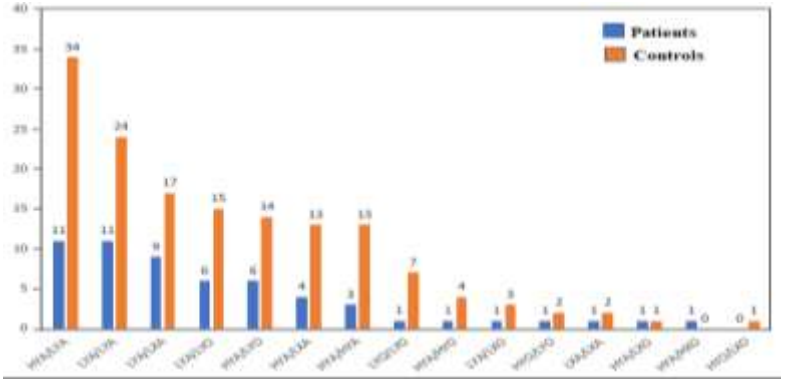
About the gender, there is a greater number of males than females; in the polymorphisms of the *MBL-2* gene curve, it is shown a large majority of individuals homozygous for the wild allele followed by heterozygous individuals and a small number of individuals homozygous for the mutant allele.
Source: Authors

Figure 2: Total frequency of haplotypes found in patients and controls



Source: Authors.

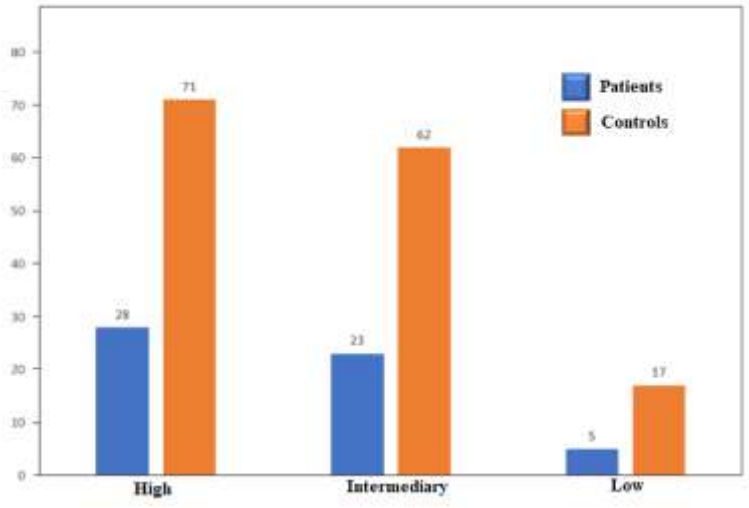
Figure 3: Comparisons between haplotypes in patients and controls.



Source: Authors.

In order to verify if there is a significant difference between the haplotypes of the patients and the control group, we grouped the study population into three: high expression, intermediate expression, and low expression of MBL, as shown in Figure 4. However, no significant difference was found between the patients and controls (p = 0.9142).

Figure 4: Comparison between the frequencies of high, intermediate and low haplotypes expression in the patients and control group. p = 0.9142.

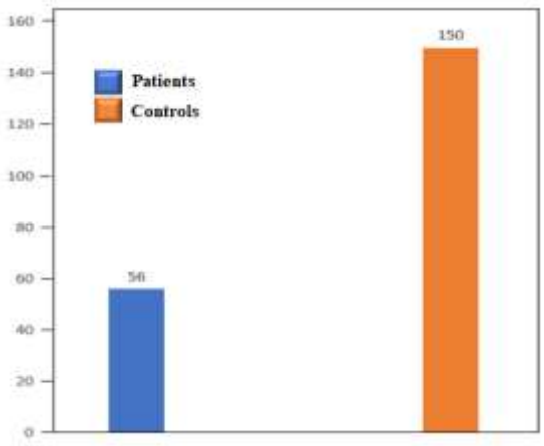


Graph shows no statistical difference between patients and controls when comparing haplotypes by expression groups
Source: Authors

3.2 Machine learning models experiments

Initially, the database had been analysed, with a focus on balancing, thus verifying the possibility of "*overfitting*". This phenomenon happens when the model has adapted very well to the data it is being trained with, however, it does not specialize for new data. In other words, the classifier "learns" much more about one group of data than another due to the large disparity of components between them. In this study, there is an imbalance in the number of patient and control groups, as can be seen in Figure 5, in addition to the low number of individuals in the patient group.

Figure 5: The imbalance regarding the number of individuals present in each group (patients and controls), a fact that contributed to the overfitting.

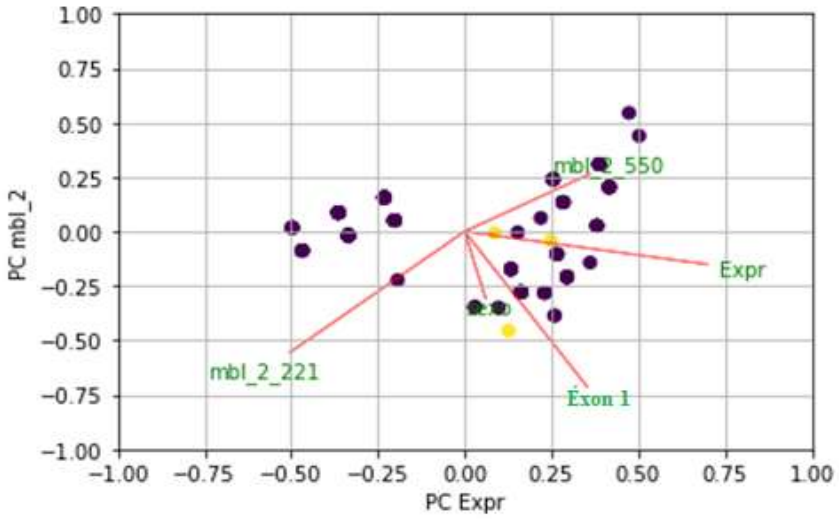


Source: Authors.

From the PCA technique, analyses were performed to indicate the two main components that best explain the variability of the data.

According to the analyzed parameters, the PCA technique was able to identify the MBL expression groups (high, intermediate or low) and the polymorphism present in *MBL-2* gene exon 1 as the two main variables that explain the variability of our data, followed by the *MBL2* -221 polymorphism. Figure 6 shows the most salient variables on the x-axis and on the y-axis.

Figure 6: A PCA plot showing the two most salient variables on the x and y axes.

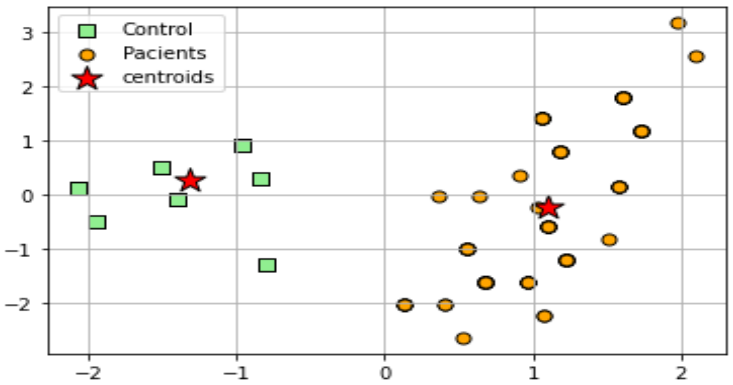


The most important variables are those that are furthest from the center of the graph. “Expr” refers to the MBL protein group expression; “Éxon 1” refers to the MBL-2 exon 1 polymorphirsm.
Source: Authors

According to the data, these variables had the highest factor loadings, thus, it was verified whether the use of these variables alone could be sufficient for the development of the classifier, along with the inherent benefits of dimensionality reduction, as discussed above.

After this phase, unsupervised clustering techniques were performed in order to verify the best number of clusters for these data. Through association analysis, it was identified that two clusters better represent our dataset showing that, for the training set, the algorithm managed to satisfactorily separate patients and controls, as can be seen in Figure 7.

Figure 7: classification using k means



The satisfactory separation of patients and controls performed by the algorithm, where it takes into account the distance from the point to the centroid to perform the classification.
Source: Authors

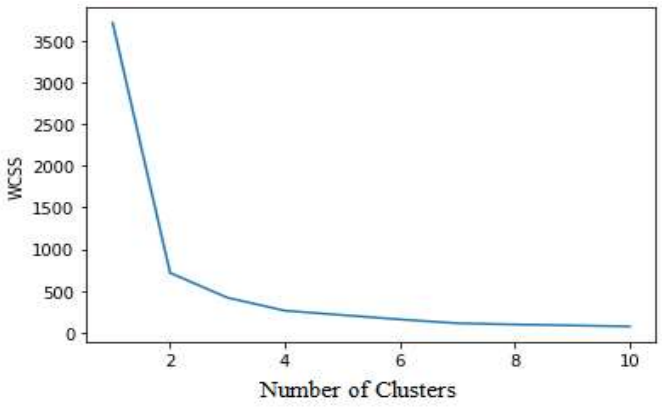
Aiming to obtain a better interpretation and validation for the clusters analysis, the Silhouette Technique was used. Table 5 shows the silhouette valuesobtained and Figure 8 shows the so-called “elbow curve”, which indicates the point from which there are no gains in relation to the increase in clusters.

Table 5: Unsupervised analysis results, which shows the best number of groups, according to the similarity of individuals in the same group and disparities with individuals from other groups.

Silhouette Analysis	
Number of Clusters	Result (Score)
2	0.706
3	0.632
4	0.571
5	0.572

Source: Authors.

Graph 8: An “elbow curve” showing the best number of clusters, in this case, two clusters.



Source: Authors.

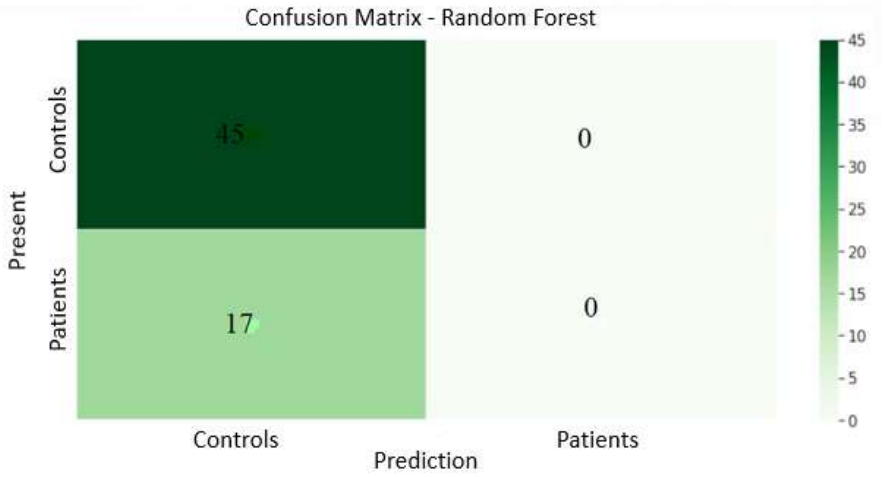
After this first analysis, part of our data was used to test the model's hit rate when a new dataset was introduced. Thus, it was possible to verify that the classifier’s hit rate was 49.51%, however when the analysis was performed using the *t-sine*, the model accuracy rose to 72.81%.

In the face of overfitting, the supervised analysis was performed using the Random Forest (RF) algorithm because, as is known, this avoids overfitting more efficiently than decision trees, in addition to obtaining greater accuracy and being more stable (J. C. da Silva, 2018).

The RF was first used for the complete database, and with this, we obtained the following confusion matrix, shown in Figure 9. It is possible to observe that, from the portion randomly selected for testing, 45 individuals classified as controls were actually controls, 17 classified as patients were actually controls, no control was classified as patient and there was no occurrence

of truly patient in this portion.

Figure 9: Matrix of confusion resulting from the use of Random Forest.



Of the portion randomly selected for testing, 45 individuals classified as controls were actually controls, 17 classified as controls were actually patients, no controls were classified as patients, and there was no classification of true patients.
Source: Authors.

According to this matrix, it was possible to calculate that the classifier obtained an accuracy of 72.58%, with a precision of 72.58% and a recall of 100%. However, as the confusion matrix shows, due to the strong imbalance in the data, the model learned much more about controls than about patients.

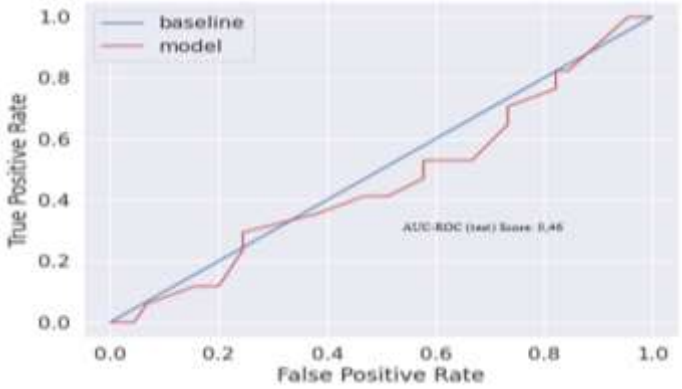
Following this analysis, we used the RF again, in its standard configuration, but this time we used only the most salient variables according to the PCA, and in this phase, we obtained an accuracy value of 68.77%.

After this, the hyperparameters to be applied in the RF algorithm were implemented along with a cross validation. These multiple configurations were assembled using a *GridSearch*-type technique, and with this, it was possible to achieve an accuracy value of 75% in the best configuration found by the algorithm where the average accuracy of the configurations was 72.93%, with a precision of 72.58 and a recall of 100%.

To verify the model efficiency in separating classes (patients and controls) two AUC-ROC curves were plotted (Khan & Rana, 2019; Prati et al., 2008). The first plot was made using the complete database with all variables and using the RF at default settings; and the second plot was made with the complete database, but after the introduction of the hyperparameters.

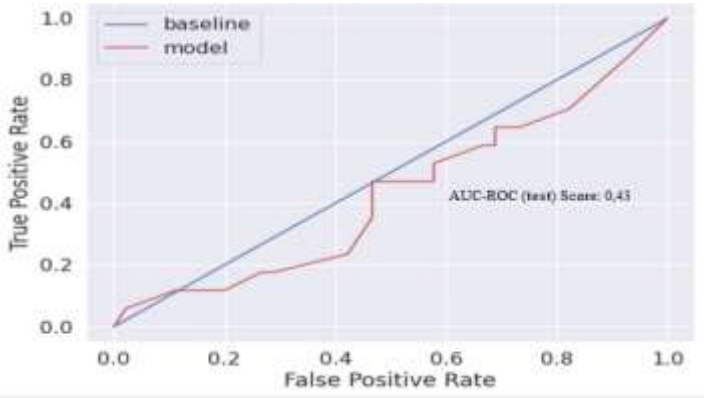
The first curve, shown in Figure 10, presented a rate of 46%, a value considered low. The second, Figure 11, showed a rate of 43%, demonstrating that overfitting caused an unsatisfactory data separation.

Figure 10: AUC-ROC curve for complete database and using standard RF settings.



The curve demonstrates a low efficiency of the model in separating patients and controls.
Source: Authors

Figure 11: AUC-ROC curve after using GridSearch.



AUC-ROC curve demonstrating the model's low efficiency in separating patients and controls, even when used in GridSearch.
Source: Authors

4. Discussion

This study was intended to analyze a possible risk factor for BL through using the Machine Learning approach. Then, it is focused on bringing a new and promising analysis tool that can assist in several future analyzes, besides being able to help in the diagnosis and prognosis of a great number of diseases

Like most types of lymphoma, BL is more prevalent in men with a 3-4: 1 ratio, between men and women, according to world statistics (Dozzo et al., 2017). In our study, a higher incidence in males was also observed, however in a slightly lower proportion (about 2: 1). This ratio is lower than that found by Silva et al. (2020), which was approximately 4: 1 when studying patients with BL and carriers of the human immunodeficiency virus in Brazil (W. F. da Silva et al., 2020) and Rodrigues-Fernandes et al. (2020), which was 3: 1, when in a systematic review compiled data on pediatric patients from several countries (Rodrigues-Fernandes et al., 2020).

Regarding the average age, Hassan et al. (2008), observed an average of 7.8 ± 3.7 years in Brazil (Hassan et al., 2008) very similar to that observed by Rodrigues-Fernandes et al. (2020), which was 7.4 years (Rodrigues-Fernandes et al., 2020). Our data, although showing a slightly lower average age (6.7 years at diagnosis with a standard deviation of 6.4295 ± 6.9705) corroborate with Hassan et al. (2008) and with Rodrigues-Fernandes et al. (2020), showing that there is a little variation at the age most affected by this pathology.

When performing a simple comparison, it is possible to observe the haplotype differences that exist between the

population in our study and the population studied by Kilpatrick, (2002b). The majority of our population presented the HYA / LYA haplotypes (21.84%), followed by LYA / LYA (16.99%) and LYA / LXA (11.16%) versus HYPA / HYPA (17.64%), HYPA / LXPA (11.76%) and HYPA / LYQA (11.76%). In both populations, the vast majority of individuals are in the genotype categories that are related to the expression of high concentration (48.06% versus 52.86%) and intermediate concentration (41.26% versus 32.09).

However, in this study, the haplotypes frequencies remained similar when comparing the patient groups and the control group, demonstrating that there were no statistically significant differences between them. This suggests that there is no participation of *MBL-2* polymorphisms in the development of Burkitt Lymphoma, unlike what happens in susceptibilities to bacterial and viral infections (Eisen & Minchinton, 2003), atherosclerosis (Rugonfalvi-Kiss et al., 2002), autoimmune disease (for example, type 1 diabetes) (Tsutsumi et al., 2003) and rheumatoid arthritis (Graudal et al., 2000), which have a well-described participation in the literature.

In view of the results obtained, we implemented some algorithms in our database. First, the PCA test was performed and it was possible to identify the variables that best explained the variability of the data, in other words, those that had the best explanatory capacity for the problem. It must be considered that the variables referring to the *MBL-2* mutations were highlighted in this test, which leads us to think that there is a participation of these polymorphisms in BL. Possibly this is due to a greater infection rate and maintenance of the EBV in patients with low expression genotypes, leading to a low concentration of serum MBL.

Another hypothesis is that these variables can facilitate the creation of a classifier capable of separating patients and controls. The K-means algorithm is a very popular approach to finding clusters due to its simplicity of implementation and quick execution (Davidson, 2002). Some of these applications are already being documented in the literature, as was done by Salma, (2016) that used a variation of K-means (fast K-means) to select the most relevant resources from a high-dimension breast cancer data set, reaching an accuracy 99.39%. In this context, it is also worth mentioning the study of Kakushadze and Yu, (2017), in which they used 1389 published samples of 14 types of cancer and found that 3 types of cancer (liver cancer, lung cancer and renal cell carcinoma) stand out from the others and had no similar structures to the cluster. In our study, using this same algorithm, we identified that two groups have the best explanatory capacity for our data, dividing them between patients and controls with a hit rate that reached 72.81% when analyzed using the *t-sne*.

In the context of supervised analysis, some authors have been using the cross-validation strategy to classify some types of cancer, for example Lee et al., (2019), who used the RF with 10-fold cross-validation to classify 31 types of cancer and managed to reach 84% accuracy, reaching up to 94% for the 6 most common types of cancer. As mentioned, RF was used in our study because it avoids overfitting more efficiently than decision trees, in addition to obtaining greater accuracy and being more stable (J. C. da Silva, 2018). Thus, by using this algorithm, it was possible to reach a reasonable accuracy level, even for a small number of individuals, especially when we use cross-validation as well as Lee et al., (2019). Therefore, we can conclude that the option of cross-validation allowed a variability of the training data and a better test than the standard RF approach based on percentages. It is notable that when using the variables highlighted by the PCA, slightly less accuracy was obtained. However, this strategy can be considered when dealing with a large amount of data, due to the lower computational costs required and the reduction in execution time.

It is also worth noting that, as the distribution analysis showed, our data is extremely unbalanced. We have a low number of patients, contributing to a strong bias for the control data represented in the results. Therefore, we understand that we could evaluate a higher number of patients and include other polymorphisms related to the innate immunity. We confirmed this fact through the AUC-ROC curve, which evidenced a low separation efficiency in this model. This low efficiency can be explained by two ways: the database is unbalanced and the algorithm found no differences between patients and controls and/or the studied

polymorphisms of the *MBL-2* gene do not exert significant force on the pathogenesis of BL, being the variables present in this study insufficient for the creation of a classifier.

5. Conclusion

Machine Learning techniques have brought new expectations to the medicine field, mainly for the diagnosis and prognosis of diseases. These are extremely promising techniques to assist in the analysis of biomedical data, as they make it possible to extract new insights from data sets that are often previously analyzed or too large to be analyzed by methods that are more conventional.

In this article, we used a Machine Learning model to verify the participation of *MBL-2* polymorphisms in the development of BL, as well as verifying whether with only the aforementioned parameters, it would be possible to, satisfactorily, classify patients and controls.

Thus, it was possible to observe that the dimensionality reduction techniques should be considered, especially when dealing with large databases, because the losses in hit rates in our study were low, compared to the benefits that these techniques provide.

The two algorithms proved to be quite efficient in classifying individuals, especially when using their variations (72.81% for KMeans and 75% for Random Forest), showing that the sophistications implemented actually bring improvements to their performance and these proposals return even better rates.

Therefore, it was not possible to conclude about the participation of the *MBL-2* polymorphisms in the development of Burkitt lymphoma. We believe that this is due to the low number of individuals present in our database (56 patients and 150 controls) and the imbalance of groups, facts that led to overfitting.

Even though it is a disease with a relatively low incidence, the results encourage us, as we believe that with the cooperation of several reference centers in the treatment of childhood BL and the creation of a unified digital medical record, approved in Brazil by Bill 3814/ 2020, it will be possible to significantly increase the robustness of our database. Soon, it will be possible to create a classifier containing the main components related to the disease that will serve as a decision support tool, based on a computational intelligence and Machine Learning algorithm.

For this reason new larger and more robust studies, especially regarding the number of patients, are needed to obtain this answer. Another suggestion would be to use other algorithms and variation, to compare the efficiency with those we used in this study.

Acknowledgments

The authors thank the Oswaldo Cruz University Hospital, in Recife, State of Pernambuco, Brazil, for allowing patients to be enrolled in this study.

References

Aydin, B., Akyuz, C., Kalkan, N., Kurucu, N., Varan, A., Yalcin, B., & Kutluk, T. (2019). FAB LMB 96 Regimen for Newly Diagnosed Burkitt Lymphoma in Children: Single-center Experience. *Journal of Pediatric Hematology/Oncology*, *41*(1), e7–e11. https://doi.org/10.1097/MPH.0000000000001270

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Bland, M. (2015). *An introduction to medical statistics*. Oxford University Press (UK).

Boldt, A. B. W., Culpi, L., Tsuneto, L. T., De Souza, I. R., Kun, J. F. J., & Petzl-Erler, M. L. (2006). Diversity of the MBL2 gene in various Brazilian populations and the case of selection at the mannose-binding lectin locus. *Human immunology*, *67*(9), 722–734.

Bouwman, L. H., Roep, B. O., & Roos, A. (2006). Mannose-binding lectin: Clinical implications for infection, transplantation, and autoimmunity. *Human immunology*, *67*(4–5), 247–256.

Da Cruz, H. L. A., Da Silva, R. C., Segat, L., de Mendonça Gomes, M. S. Z., Brandão, L. A. C., Guimarães, R. L., Santos, F. C. F., de Lira, L. A. S., Montenegro, L. M. L., & Schindler, H. C. (2013). MBL2 gene polymorphisms and susceptibility to tuberculosis in a northeastern Brazilian population. *Infection, Genetics and Evolution*, *19*, 323–329.

Davidson, I. (2002). Understanding K-means non-hierarchical clustering. *SUNY Albany Technical Report*, *2*, 2–14.

Derinkuyu, B. E., Boyunağa, Ö., Öztunalı, Ç., Tekkeşin, F., Damar, Ç., Alımlı, A. G., & Okur, A. (2016). Imaging features of Burkitt lymphoma in pediatric patients. *Diagnostic and Interventional Radiology*, *22*(1), 95.

Division of Cancer Epidemiology and Genetics—National Cancer Institute (nciglobal,ncienterprise). (2018a, janeiro 1). [CgvHomeLanding]. https://dceg.cancer.gov/

Dozzo, M., Carobolante, F., Donisi, P. M., Scattolin, A., Maino, E., Sancetta, R., Viero, P., & Bassan, R. (2017). Burkitt lymphoma in adolescents and young adults: Management challenges. *Adolescent health, medicine and therapeutics*, *8*, 11.

Eisen, D. P., & Minchinton, R. M. (2003). Impact of mannose-binding lectin on susceptibility to infectious diseases. *Clinical Infectious Diseases*, *37*(11), 1496–1505.

Freedman, A. S., Aster, J. C., & Rosmarin, A. G. (2018). *Epidemiology, clinical manifestations, pathologic features, and diagnosis of Burkitt lymphoma*.

Graudal, N. A., Madsen, H. O., Tarp, U., Svejgaard, A., Jurik, A. G., Graudal, H. K., & Garred, P. (2000). The association of variant mannose-binding lectin genotypes with radiographic outcome in rheumatoid arthritis. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, *43*(3), 515–521.

Hansen, T. K., Tarnow, L., Thiel, S., Steffensen, R., Parving, H.-H., & Flyvbjerg, A. (2004). Association between mannose-binding lectin and vascular complications in type 1 diabetes. *Scandinavian Journal of Immunology*, *59*(6), 613–613.

Harrison, E., Singh, A., Morris, J., Smith, N. L., Fraczek, M. G., Moore, C. B., & Denning, D. W. (2012). Mannose-binding lectin genotype and serum levels in patients with chronic and allergic pulmonary aspergillosis. *International journal of immunogenetics*, *39*(3), 224–232.

Hassan, R., Klumb, C. E., Felisbino, F. E., Guiretti, D. M., White, L. R., Stefanoff, C. G., Barros, M. H. M., Seuánez, H. N., &

Zalcberg, I. R. (2008). Clinical and demographic characteristics of Epstein-Barr virus-associated childhood Burkitt's lymphoma in Southeastern Brazil: Epidemiological insights from an intermediate risk region. *haematologica*, *93*(5), 780–783.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Hecht, J. L., & Aster, J. C. (2000). Molecular biology of Burkitt's lymphoma. *Journal of Clinical Oncology*, *18*(21), 3707–3721.

Hladnik, U., Braida, L., Boniotto, M., Pirulli, D., Gerin, F., Amoroso, A., & Crovella, S. (2002). Single-tube genotyping of MBL-2 polymorphisms using melting temperature analysis. *Clinical and experimental medicine*, *2*(2), 105–108.

Hsu, J. L., & Glaser, S. L. (2000). Epstein–Barr virus-associated malignancies: Epidemiologic patterns and etiologic implications. *Critical reviews in oncology/hematology*, *34*(1), 27–53.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, New Jersey.*

Kakushadze, Z., & Yu, W. (2017). *K-means and cluster models for cancer signatures. *Biomolecular Detection and Quantification*, *13*, 7–31. https://doi.org/10.1016/j.bdq.2017.07.001

Khan, S. A., & Rana, Z. A. (2019). Evaluating performance of software defect prediction models using area under precision-Recall curve (AUC-PR). *2019 2nd International Conference on Advancements in Computational Sciences (ICACS)*, 1–6.

Kilpatrick, D. C. (2002a). Mannan-binding lectin and its role in innate immunity. *Transfusion Medicine*, *12*(6), 335–352.

Kilpatrick, D. C. (2002b). Mannan-binding lectin: Clinical significance and applications. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1572*(2–3), 401–413.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Lee, K., Jeong, H., Lee, S., & Jeong, W.-K. (2019). CPEM: Accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network. *Scientific Reports*, *9*(1), 16927. https://doi.org/10.1038/s41598-019-53034-3

Li, W., Cerise, J. E., Yang, Y., & Han, H. (2017). Application of t-SNE to human genetic data. *Journal of Bioinformatics and Computational Biology*. https://doi.org/10.1142/S0219720017500172

Lins, A. J. C. C., Muniz, M. T. C., Garcia, A. N. M., Gomes, A. V., Cabral, R. M., & Bastos-Filho, C. J. A. (2017). Using artificial neural networks to select the parameters for the prognostic of mild cognitive impairment and dementia in elderly individuals. *Computer Methods and Programs in Biomedicine*, *152*, 93–104. https://doi.org/10.1016/j.cmpb.2017.09.013

Madsen, H. O., Garred, P., Thiel, S., Kurtzhals, J. A., Lamm, L. U., Ryder, L. P., & Svejgaard, A. (1995). Interplay between promoter and structural gene variants control basal serum level of mannan-binding protein. *The Journal of Immunology*, *155*(6), 3013–3020.

Martín-Mateos, M. A., & Piquer Gibert, M. (2016). Primary immunodeficiencies and B-cell lymphomas. *Boletín Médico Del Hospital Infantil de México*, *73*(1),

18–25. https://doi.org/10.1016/j.bmhimx.2015.11.009

Mendonça, T. F., Oliveira, M., Vasconcelos, L. R. S., Pereira, L., Moura, P., Bezerra, M. A. C., Santos, M. N. N., Araújo, A. S., & Cavalcanti, M. S. M. (2010). Association of variant alleles of MBL2 gene with vasoocclusive crisis in children with sickle cell anemia. *Blood Cells, Molecules, and Diseases*, *44*(4), 224–228.

Molyneux, E. M., Rochford, R., Griffin, B., Newton, R., Jackson, G., Menon, G., Harrison, C. J., Israels, T., & Bailey, S. (2012). Burkitt's lymphoma. *The Lancet*, *379*(9822), 1234–1244.

Moslem, M., Mahmoudabadi, A. Z., Fatahinia, M., & Kheradmand, A. (2015). Mannose-binding lectin serum levels in patients with candiduria. *Jundishapur Journal of Microbiology*, *8*(12).

MWer, S., Dykes, D., & Polesky, H. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic acids res*, *16*(3), 1215.

Niitsuma, H., & Okada, T. (2007). Covariance and PCA for Categorical Variables. *arXiv:0711.4452 [cs]*. http://arxiv.org/abs/0711.4452

Petersen, S. V., Thiel, S., & Jensenius, J. C. (2001). The mannan-binding lectin pathway of complement activation: Biology and disease association. *Molecular immunology*, *38*(2–3), 133–149.

Prati, R. C., Batista, G., & Monard, M. C. (2008). Curvas ROC para avaliação de classificadores. *Revista IEEE América Latina*, *6*(2), 215–222.

QIAamp® DNA Mini and Blood Mini Handbook. Sample & Assay Technologies. ([s.d.]).

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, *5*, 532–538.

Rodrigues-Fernandes, C. I., Pérez-de-Oliveira, M. E., Aristizabal Arboleda, L. P., Fonseca, F. P., Lopes, M. A., Vargas, P. A., & Santos-Silva, A. R. (2020). Clinicopathological analysis of oral Burkitt's lymphoma in pediatric patients: A systematic review. *International Journal of Pediatric Otorhinolaryngology*, *134*, 110033. https://doi.org/10.1016/j.ijporl.2020.110033

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rugonfalvi-Kiss, S., Endrész, V., Madsen, H. O., Burián, K., Duba, J., Prohászka, Z., Karádi, I., Romics, L., Gönczöl, É., &

Füst, G. (2002). Association of Chlamydia pneumoniae with coronary artery disease and its progression is dependent on the modifying effect of mannose-binding lectin. *Circulation*, *106*(9), 1071–1076.

Salma, M. U. (2016). Pso based fast k-means algorithm for feature selection from high dimensional medical data set. *2016 10th International Conference on Intelligent Systems and Control (ISCO)*, 1–6.

Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, *2*(2), 34–38.

Sharma, P. (2019). The Most Comprehensive Guide to K-Means Clustering You'll Ever Need. *URL: https://www. analyticsvidhya. com/blog/2019/08/comprehensiveguide-k-means-clustering*.

Silva, J. C. da. (2018, março 13). *O Algoritmo da Floresta Aleatória*. Medium. https://medium.com/machina-sapiens/o-algoritmo-da-floresta-aleat%C3%B3ria-3545f6babdf8

Silva, W. F. da., Garibaldi, P. M. M., Rosa, L. I. da., Bellesso, M., Clé, D. V., Delamain, M. T., Rego, E. M., Pereira, J., & Rocha, V. (2020). Outcomes of HIV-associated Burkitt Lymphoma in Brazil: High treatment toxicity and refractoriness rates – A multicenter cohort study. *Leukemia Research*, *89*, 106287. https://doi.org/10.1016/j.leukres.2019.106287

Soltani, A., RahmatiRad, S., Pourpak, Z., Alizadeh, Z., Saghafi, S., HajiBeigi, B., Zeidi, M., & Farazmand, A. (2014). Polymorphisms and serum level of mannose-binding lectin: An Iranian survey. *Iranian Journal of Allergy, Asthma and Immunology*, 428–432.

Swerdlow, S. H., Campo, E., Pileri, S. A., Harris, N. L., Stein, H., Siebert, R., Advani, R., Ghielmini, M., Salles, G. A., Zelenetz, A. D., & Jaffe, E. S. (2016). The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*, *127*(20), 2375–2390. https://doi.org/10.1182/blood-2016-01-643569

Tsutsumi, A., Ikegami, H., Takahashi, R., Murata, H., Goto, D., Matsumoto, I., Fujisawa, T., & Sumida, T. (2003). Mannose binding lectin gene polymorphism in patients with type I diabetes. *Human Immunology*, *64*(6), 621–624. https://doi.org/10.1016/S0198-8859(03)00054-5

Van Der Aalst, W. (2016). Data science in action. In *Process mining* (p. 3–23). Springer.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

Vardiman, J. W., Arber, D. A., Brunning, R. D., Larson, R. A., Matutes, E., Baumann, I., Swerdlow, S. H., Campo, E., Harris, N. L., & Jaffe, E. S. (2008). WHO classification of tumours of haematopoietic and lymphoid tissues. *Lyon: International Agency for Research on Cancer*.

Watanabe, S. (1985). *Pattern recognition: Human and mechanical*. John Wiley & Sons, Inc.

White, L. R. (2004). *Análise de polimorfismo do promotor dos genes da interleuccina 10 e do fator de necrose tumoral como fator de suscetibilidade genética em linfomas de Burkitt de crianças*. 126–126.

Xu, R., & Wunsch, D. C. (2005). *Survey of clustering algorithms*.