Machine Learning Applied to the Early Diagnosis of Leukemia Using Biomarkers

Fernanda T. Ferry [*] **Giulia Zanon Castro** [**]
Ramon Gonçalves Pereira [***]

[*] *Computer Science - UniBH | Centro Universitário de Belo Horizonte Av. Cristiano Machado, 4000 - União, Belo Horizonte - MG, 31160-900, (e-mail: nandateixferry@gmail.com).*
[**] *Graduate Program in Electrical Engineering - Universidade Federal de Minas Gerais - Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil (e-mail: giuliaz@ufmg.br)*
[***] *Computer Science - UniBH | Centro Universitário de Belo Horizonte Av. Cristiano Machado, 4000 - União, Belo Horizonte - MG, 31160-900, (e-mail: ramon.pereira@prof.unibh.br)*

Abstract: Leukemia is a rare and lethal blood cancer. One of the factors that increase the patient's chances of better treatment results is early diagnosis. The best attempt to discover leukemia usually is the image analysis exams, but this is costly, and sometimes it is late. Thus, this paper uses attributes of a complete blood count as input to Machine Learning algorithms to predict earlier and cheaper leukemia diagnoses. In this paper, we collected actual exam results. We developed a synthetic dataset with 1000 examples based on the distribution and limits of each attribute to classify a patient in positive or negative for leukemia. We tested different classifiers (Logistic Regression, Random Forest, XGBoost, and SVM) to predict sample classes. We show that it is possible with an accuracy of 96% to predict if a patient is likely to have leukemia based on its blood count.

Keywords: Machine Learning; Leukemia; Earlier Diagnoses; Explainability; Decision Making; Complete Blood Count

1. INTRODUCTION

Currently, many diseases already have a known cause, easy diagnosis or treatments, without causing severe damage to the patient. However, since cancer consists of more than 100 diseases with uncontrolled cell growth in common (Ministério da Saúde, 2020), each diagnosis is still tricky. The varieties of cancer are the most common causes of death in the world, representing 9.6 million deaths in 2018 (Wild, 2020). In Brazil, there is an estimate that 8460 Brazilians with 1 to 19 years old (4310 men, 4150 women) are diagnosed with cancer annually (Instituto Nacional do Câncer, 2019).

However, leukemia has a high cure rate when the appropriate treatment started (World Health Organization, 2020). Like any other cancer, early diagnosis is one factor that increases the patient's probability of better treatment results. The earlier treatment begins, the greater the probability of successful. On the other hand, some patients have symptoms not linked to leukemia, treated like other pathologies and delaying the diagnosis. Symptoms commonly reported are tiredness, pallor, sweating, shortness of breath, signs of bleeding, fever, and frequent infections (Hamerschlak, 2008).

A complete blood count (CBC) may also be used to diagnose leukemia. In this disease, the malignant blood cells multiply and hinder the production of healthy blood cells from the bone marrow (Abrale, 2020a). This barrier causes a weak immune system in the patient. The myelogram helps confirm the disease – which is bone marrow collection (Abrale, 2020b). The blood count can reveal anemia with a high level of leucocytes. In the literature, are also reported changes in the number of hemoglobin and platelets (Abrale, 2020b).

In order to aid in early diagnosis, there is the development of computer systems to perform tasks that humans usually perform. Most of the research in the construction of systems that aim to diagnose diseases took place in Artificial Intelligence (AI) (Nilsson, 2014). These systems are trained by machine learning algorithms and have a big computational capacity. In order to carry out data training, it is necessary to access large amounts of data, a good data model, and a powerful machine for fast and accessible processing. These systems can learn their patterns through an input dataset and be used, for example, in a classification task. They are used in problems of a different nature, such as "yes" or "no", or in categories. The classification can be used to determine whether or not a

patient is at risk of being diagnosed with leukemia and aid clinical decision–making.

However, data acquisition in medicine is a high–cost task due to its rarity. Thus, we develop a study to test if the CBC can generate good results on the task of an early leukemia diagnosis. This study can be used as a baseline to compare with other approaches using real–world data or even synthetic data in other scenarios. We also provide an explanation to answers questions such as "How does the model decides?". Thus, this work presents the main contributions:

- evaluation/ standardization of blood count attributes to be used as input to machine learning (ML) models;
- evaluation of ML models for the classification of leukemia using blood tests;
- explanation of the factors that lead the model to its decision in the classification.

2. RELATED WORKS

AI and medicine are used worldwide (Lobo, 2017). The computational algorithms propose accurate diagnoses, and new ways of preventing and recovering from illnesses are created. Currently, many systems and works already focus on automatic diagnosis. For example, exists the Portal Telemedicina platform, which is an innovative solution for teleconsultation. This system uses the TensorFlow software, a ML library created by Google.

The study carried out by Mesquita (2017) discusses how AI and ML can be used in complex problems in science. With a focus on Cardiology, the author demonstrates that methods are now incorporated to find abnormalities accurately. Fatima and Pasha (2017) discusses that AI algorithms promise greater precision and perception in diagnosing diseases in the field of biomedicine. The authors used ML algorithms to diagnose diseases such as diabetes, dengue, hepatitis, and cardiovascular disease.

Regarding leukemia, Cheng (2019) created a diagnosis system with AI to identify white blood cells, reducing the diagnosis time. This diagnosis required two steps, first get images of blood cell samples, and after using a ResNet classification method. Following this line, Maria and T. Devi (2020) discuss some ML algorithms used to treat, classify or detect leukemia to assist hematologists. In this work, the input data of the studied algorithms are blood images. Then, Support Vector Machines (SVM), k–Nearest Neighbors (KNN), Neural Networks, Naive Bayes, and Deep Learning were used for the classification task. This paper used SVM to classify leukemia as acute lymphoblastic or acute myeloid through blood image samples from lymphoid and myeloid stem cells. SVM has reached an accuracy of 92%. The KNN algorithm was used to classify leukemic cells. KNN achieved an accuracy of 80% when classifying blasts in leukemic cells. Neural Networks are used to classify blood smear images in normal and leukemic blood cells, assisting in the clinical decision to diagnose leukemia. Furthermore, the Naive Bayes algorithm was used to classify leukocytes. Finally, a deep learning approach was used to classify subtypes of acute lymphocytic leukemia through blood images with an accuracy of up to 97,78%.

This diagnosis is challenging by different factors, including lack of access to health care and incorrect classification due to the shortage of experts in some regions of the world (Salah et al., 2019). Salah et al. (2019) carried out a bibliographic study on the applications of ML in the diagnosis of leukemia, and the algorithms were being applied prospectively in real–world scenarios. The work of Sossela (2017) presents the main clinical aspects of the disease and methods for diagnosis, focusing on the features changes found in the blood count. In this work, the idea that the blood count still has great relevance because it is an easily accessible methodology is reinforced.

Given the above, it is seen that AI positively assists the diagnosis and classification of leukemia. However, most of these algorithms use images as input data. This work evaluates data on CBC instead of images for training machine learning algorithms, bringing a new approach to classifying this disease. Therefore, the work assists in the early diagnosis of leukemia, but it can also increase the importance of the results of the CBC in clinical analyzes.

3. METHODOLOGY

This section presents the methodology of the three stages of our work (Fig. 1): (i) generation of a dataset from actual leukemia diagnosis data, collected and anonymized for this research; (ii) creation of ML models for the prediction of leukemia; (iii) explainability of the factors that lead the models to its decision–making in the classification.

3.1 Acquisition and generation of the leukemia dataset

In this work, we carried out a complete blood count (CBC) survey of the target population (people who have or had leukemia) for the study and creation of the dataset. For this, we made an explanatory video on the theme of the present work and its objectives. Then, the target population, who openly contributed to this research, sent the CBC data. The videos reached approximately 10,000 views on the internet, with 55 answers containing the results of the exams. All data collected was wholly anonymized. We transcript it from PDF Image format to a text file with the desired format and pattern, e.g., data format and decimal points, to standardize this data. We got data from 26 actual exams; we excluded the other 29 by inconsistency. Thus we create a synthetic dataset.

(1) In order to create data in which leukemia was present, we use tests from our dataset of people who had or have leukemia. Those exams were used to estimate the maximum and minimum values of each attribute. Thus, for each feature, we created a normal distribution $N(\bar{x}, S)$ based on the average $\bar{x}$ and the standard deviation S. We generated random values on the range of the features of each attribute, following the normal distribution $N(\bar{x}, S)$, generating synthetic samples.
(2) Likewise, for the creation of data referring to healthy people with absent leukemia, we used the typical reference values for each variable as a basis. We perform the generation of new data through random values following a normal distribution within the specified threshold. Their respective clinical significance are shown in Table 1.

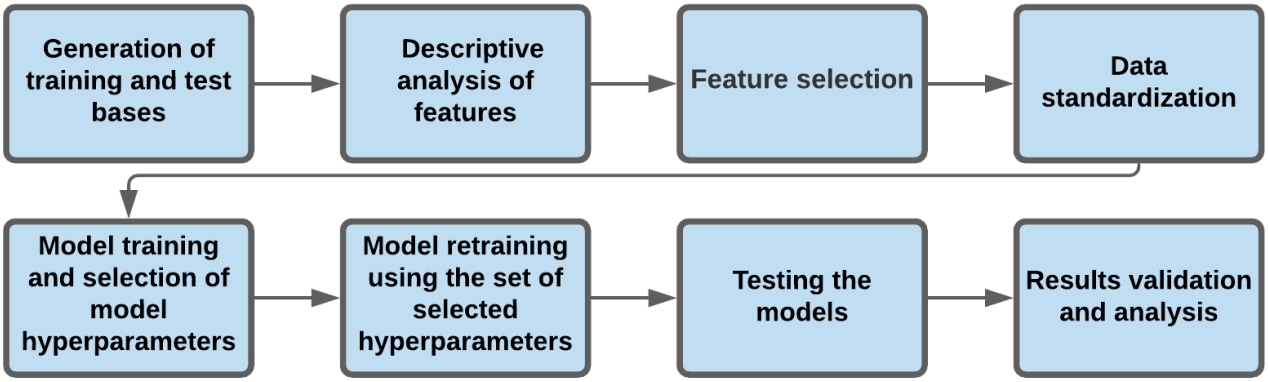

Figure 1. Flowchart of research steps. The training dataset is made up of a synthetic dataset, and the test dataset is made up of the actual dataset

Table 1. Hemogram attributes used for the dataset generation for the non sick people

Features	Women	Men	Meaning
Red Cells (million/µL)	3.9 – 5.4	4.2 – 5.9	Also known as red blood cells. Its function is to bring oxygen to the cells
Hemoglobin (g/dL)	12.0 – 16.0	13.0 – 18.0	Protein responsible for transporting oxygen from the lungs to tissues, by red blood cells
Hematocrit (%)	35 – 47	38 – 52	Percentage of red cells in the total blood volume
Vcm (fL)	80.0 – 100.0	80.0 – 100.0	Index that represents the average of the sizes of the red blood cells
Hcm (pg)	27.0 – 32.0	27.0 – 32.0	Parameter that measures the size and color of hemoglobin within the blood cell
Chcm (g/dL)	31.0 – 36.0	31.0 – 36.0	Hemoglobin concentration within a red cell
Rdw (%)	10.0 – 16.0	10.0 – 16.0	Determines the size variation within a red cell
Leukocytes (µL)	4000 – 11000	4000 – 11000	White blood cells. Responsible for defending the body from diseases, infections, allergies
Lymphocytes (µL)	900 – 4000	900 – 4000	Type of defense cell of the organism, produced more intensively when there is an infection
Monocytes (µL)	100 – 1000	100 – 1000	Group of cells of the immune system.Defending the organism from foreign bodies, such as virus
Eosinophils (µL)	0 – 500	0 – 500	Defends against foreign microorganisms
Platelets (µL)	140 000 – 450 000	140 000 – 450 000	Blood cells produced by the bone marrow. Are responsible for the blood clotting process

(3) In the generation of the dataset, we divided the data into 60% without leukemia and 40% with leukemia, aiming for an ideal world for machine learning but considering the real-world proportion seeking for leukemia diagnosis. We need many positive tests for leukemia to train the model without escaping the reality that the number of people with the disease is smaller than the number of healthy people.

3.2 Generation of the machine learning models

Fig. 1 shows the process diagram. In the first step, we evaluate the attributes presents in the blood tests and their patterns. In the second step, we perform a descriptive analysis of these attributes. We used the generated dataset to train and validate our models and the actual data to test our models. The data division considers that most patients who have a complete blood test do not have leukemia, which is an imbalance of classes, a common problem regarding machine learning techniques.

Fig. 2 shows the correlation coefficients between the features, which refers to how close two attributes are to having a linear relationship with each other. The closer to 1, the more correlated the attributes are. For example, in Fig. 2 the leukocytes and Rdw have a high correlation. In order to validate these data, we perform tests excluding one or more attributes from the dataset to understand the model's behavior concerning this data.

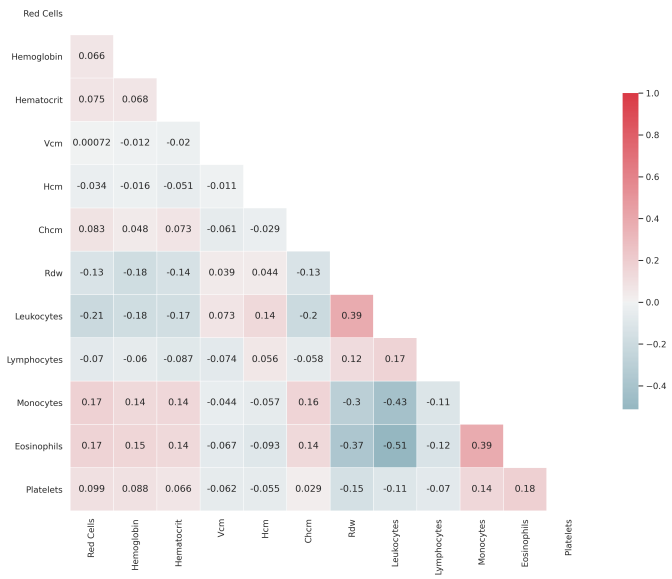

Figure 2. Correlation between attributes

The ML algorithms chosen for training the model were Logistic Regression (LR), Random Forest (RF), XGBoost, and Support Vector Machines (SVM). The RF (Breiman, 2001) is a supervised machine learning method. Through this algorithm, it is possible to train several decision trees obtained from samples of the dataset. XGBoost (Chen and Guestrin, 2016) is an evolution of Random Forest, also being an algorithm based on the decision tree. It can be used to solve, for example, regression, classification, and ranking problems. SVM (Cortes and Vapnik, 1995) is a supervised algorithm that can be used for classification or regression. These algorithms presented good results for several problems and challenges and were chosen to check if a specific type was better than others. For this type of problem, trees are a great choice as they quickly learn and make predictions. They are often also accurate for a wide range of questions. Deep learning was not used because the size of the dataset.

We perform the experiments using the Python language and the Scikit–Learn library (Scikit Learn, 2020), an open–source library for creating ML models. To achieve our results, we tested the Random Forest and XGBoost algorithms with the depth of the tree varying from 1 to 20. On the SVM, we used the kernel RBF and tested it with 7 variations in hyperparameter C {0.0001, 0.01, 0.05, 0.1, 0.5, 1, 10}. On the LR we tested the set of C {0.001, 0.01, 0.05, 0.1, 0.5, 1}.

To select the best hyperparameters, we divide the set into 75% of training and 25% of validation. After this selection, the models have trained again with the training and validation set. To test our models, we used the actual data as a test set to obtain the results reported in Section 3.3.

3.3 Evaluation and explainability of the machine learning models

One of the biggest problems with ML is overfitting. It occurs when a model can memorize the data and overestimate the performance due to that memorization. Thus, an overfitting model does not have good generalizability. That is, when submitted to new data, it may not perform well. The metrics presented in this work are used only on actual data to minimize this effect, while the model training is done from the synthetic set.

We evaluate our models in accuracy and the f1–score. We choose the f1–score measure since it is more appropriate when the output classes are not in the same proportion, i.e., presenting a data imbalance. Accuracy, which corresponds to the number of hits of the model concerning the total, can give a false impression of how good a model is. For example, in the medical field, if the proportion of a disease is one person affected in 100, and the model always predicts that the person is not affected by the disease, the accuracy is 99%. At the same time, the prediction for the only person with the disease would be wrong. The f1–score, in turn, which is a harmonic mean between precision and recall, is a more consistent measure for this problem.

For binary classification problems, f1–score is given by

$$fscore = \frac{(1+\beta^2) \times (precision \times recall)}{(\beta^2 \times precision + recall)}, \quad (1)$$

where the value of $\beta = 0.5$ is commonly used. Thus, Equation 1 becomes

$$fscore = 2 \times \frac{precision \times recall}{precision + recall} \quad (2)$$

Also, we did an exchange between the features used to train the algorithm to see if some feature was providing overfitting in the model (Table 2). Finally, we analyze the impact of input features on the model decision using the SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017). These values are inspired by game theory and are used to explain the output of a ML model.

4. RESULTS

This section presents the results obtained by the application of the models on the dataset to predict leukemia. Table 2 shows the f1–score performance and accuracy measures for each trained model. The results consist of the evaluation of the models using the test set of actual data. We did the training of the models, in turn, using the synthetic dataset. As explained above, this training data was formed by a set of 60% (600) negative for leukemia and 40% (400) positive.

As can be seen in Table 2, all algorithms were able to identify patterns in the leukemia data when the input characteristics were at least the number of red cells, hemoglobin, hematocrit, vcm, hcm, chcm, rdw, lymphocytes, monocytes, and platelets. Other essential features are the leukocyte and eosinophil numbers since the cancer attacks mainly the leukocyte cells. However, we remove these features over most tests due to their high correlation with other features and the feature of interest. In this case, the best results were found with the SVM and XGBoost models, with f1–score of 96%. When using all characteristics, the best result was obtained with the RF model.

Fig. 3 shows the distribution of predictions for each algorithm used, in which we show the decision boundary. Predictions with a probability greater than 0.5 are classified as positive leukemia (concentrated on the right side of the graphics). In contrast, predictions with probabilities less than 0.5 are classified as negative for leukemia. The actual positive results are shown in the hatched bars.

Regarding the miss classification of a patient, the results of the Random Forest algorithm show that false positive and false negative are closer to the decision surface, i.e., incorrect predictions have a probability close to 0.5. In other algorithms, incorrect predictions have probability farther from the decision threshold and may be considered more incorrect in terms of probabilities showing a lower performance.

XGBoost achieved the best results in terms of accuracy. However, Random Forest achieves a f1–score above 83% due to all experiments performed, considering all the variations of the input features. Thus, we will proceed with the analysis using Random Forest and the model trained with the Red Cells, hemoglobin, hematocrit, vcm,

Table 2. Comparison of accuracy and f1–score between Logistic Regression (LR), Random Forest, XGBoost and Support Vector Machines (SVM) for leukemia classification

Features/Algorithm	**LR**		**Random Forest**		**XGBoost**		**SVM**	
	Acc	Fscore	Acc	Fscore	Acc	Fscore	Acc	Fscore
Red Cells, hemoglobin, hematocrit, vcm, hcm and chcm	56 %	56 %	84 %	83 %	76 %	74 %	76 %	76 %
Red Cells, hemoglobin, hematocrit, vcm, hcm, chcm and monocytes	72 %	72 %	88 %	87 %	80 %	78 %	88 %	87 %
Red Cells, hemoglobin, hematocrit, vcm, hcm, chcm and lymphocytes	68 %	68 %	92 %	91 %	84 %	80 %	96 %	96 %
Red Cells, hemoglobin, hematocrit, vcm, hcm, chcm, rdw, lymphocytes, monocytes and platelets	84 %	83 %	88 %	86 %	88 %	86 %	96 %	96 %
Red Cells, hemoglobin, hematocrit, vcm, hcm, chcm, rdw, lymphocytes, monocytes, eosinophils and platelets	88%	87 %	88%	86 %	96 %	96 %	92 %	91 %
Red Cells, hemoglobin, hematocrit, vcm, hcm, chcm, rdw, leukocytes, lymphocytes, monocytes, eosinophils and platelets	84 %	83 %	92 %	91 %	84 %	83 %	92 %	91 %

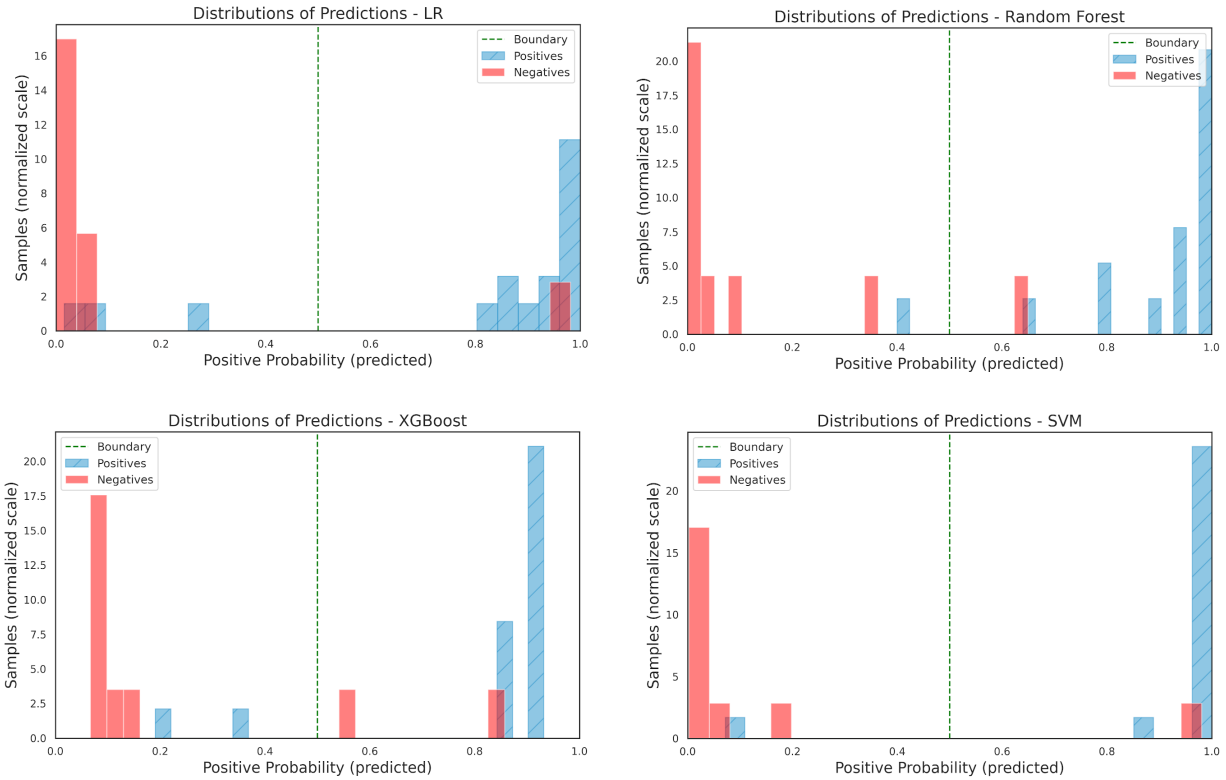

Figure 3. Decision surface. Predictions with probability greater than 0.5 are classified as positive leukemia. Actual leukemia cases are shown in the hatched bar

hcm, chcm, rdw, leukocytes, lymphocytes, monocytes, and eosinophils to understand how much each feature influences the model's decision.

Fig. 4 shows this impact of the attributes on the model's output, using SHAP values. The attributes that most influenced the model in decision making were leukocytes, lymphocytes, and monocytes. Their importance is very high, and this is consistent. There is a change in the number of leukocytes in a person with chronic myeloid leukemia concerning a healthy person. Furthermore, chronic lymphoid leukemia is characterized by an increase in the number of lymphocytes. In chronic myelomonocytic leukemia, there is an increase in the number of monocytes in the blood.

Finally, Fig. 5 shows an explanation of the algorithm diagnosing a patient as positive for leukemia. In this case, the probability of the patient having leukemia is 90% for the model. The values of monocytes, eosinophils, and

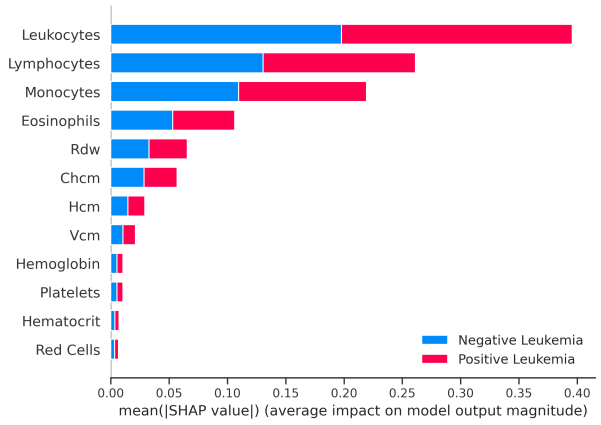

Figure 4. Impact of each attribute on the model's classification of leukemia

Figure 5. Explainability for a patient with a positive diagnosis of leukemia

lymphocytes presented increased the probability of the output until it reaches that value. These local explanations could be used, in this way, to understand the patient's context and work as a diagnostic aid tool by the health professional.

5. CONCLUSIONS

There is a long path to achieve good results on the task of predicting diagnoses. In Machine Learning, we have two choices for cancer predictions: blood count exams and image analysis. The first of one is cheaper than the second one and less assertive, but it can help alarm the physicians in the early stage of the disease, while cancer has not been considered a diagnosis yet.

In this paper, we showed that is possible to have a model to work on blood count exams with a f1–score higher as 96% using SVM and XGBoost, and 91% with Random Forest. This means that we can have a potent tool to work. One of the essential features to support the decisions was hemoglobin and platelets. It has been cited as already have been cited on Abrale (2020b). Changes in the level of those attributes can lead us to diagnose leukemia. Meanwhile, modifications on the leukocytes and the lymphocytes can indicate the presence of chronic leukemia. Leukocytes and Lymphocytes were also reported as essential attributes by our model. The results seem to be pretty good, but the training set was a synthetic one. Our approach to testing real-world data seems to be valid, and the way we selected the features for each model tested makes it away from the overfitting as the data will have more varied values than the synthetic ones. However, we know that in a scenario with a large amount of actual data, the accuracy and f1 score of the models may be lower than the reported in this paper.

Since it, as future work, we suggest applying the models created in real-world data and creating a new scenario with the actual data in all those steps to train, validate, and test a new model to compare both. We also expect to develop a tool to be side by side with the physicians on the clinical medical to support their decisions.

REFERENCES

Abrale (2020a). Leucemias: Saiba tudo sobre todos os tipos de leucemias. `https://www.abrale.org.br/doencas/leucemia/`. (accessed: 24.11.2020).

Abrale, R. (2020b). Como é o hemograma de uma pessoa com leucemia? `https://revista.abrale.org.br/hemograma-e-diagnostico-de-leucemia/`. (accessed: 07.04.2021).

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Cheng, S. (2019). Videl: A vision-based ai diagnoser for early leukemia. In *Proceedings Of The Hpc Asia 2019 Workshops*.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273–297.

Fatima, M. and Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. In *Journal Of Intelligent Learning Systems And Applications*, 1–16.

Hamerschlak, N. (2008). Leucemia: fatores prognósticos e genética. *Jornal de Pediatria*, 84(4). doi:https://doi.org/10.1590/S0021-75572008000500008.

Instituto Nacional do Câncer (2019). *Estimativa Incidência de Câncer no Brasil*.

Lobo, L.C. (2017). Inteligência artificial e medicina. In *Revista Brasileira de Educação Médica*, 185–193.

Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, 4765–4774. Curran Associates, Inc. URL `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

Maria, I.J. and T. Devi, D.R. (2020). Machine learning algorithms for diagnosis of leukemia. *International Journal of Scientific & Technology Research*, 09.

Mesquita, C.T. (2017). Artificial intelligence and machine learning in cardiology - a change of paradigm. In *International Journal Of Cardiovascular Sciences*, 187–188.

Ministério da Saúde (2020). Câncer: sintomas, causas, tipos e tratamentos. `https://saude.gov.br/saude-de-a-z/cancer`. (accessed: 29.04.2020).

Nilsson, N.J. (2014). *Principles of Artificial Intelligence*. Morgan Kaufmann.

Salah, H.T., Muhsen, I.N., Salama, M.E., and Hashmi, T.O.S.K. (2019). Machine learning applications in the diagnosis of leukemia: Current trends and future directions. *International Journal of Laboratory Hematology*, 41(6). doi:https://doi.org/10.1111/ijlh.13089.

Scikit Learn (2020). Machine learning in python. `https://scikit-learn.org/stable/`. (accessed: 24.11.2020).

Sossela, F.R. (2017). Leucemia mieloide crônica: aspectos clínicos, diagnóstico e principais alterações observadas no hemograma. In *Revista Bras. de Análises Clínicas*, 127–130.

Wild, C.P. (2020). *World Cancer Report: Cancer research for cancer prevention*. Lyon França: International Agency for Research on Cancer. URL `http://publications.iarc.fr/586`.

World Health Organization (2020). Cancer. `https://www.who.int/news-room/fact-sheets/detail/cancer`. (accessed: 29.04.2020).