INTERNSHIP REPORT

Cancer Research Center of Toulouse

April 1st 2021 - August 31st 2021

Ibrahim Didi

ÉCOLE
POLYTECHNIQUE

IP PARIS

ABSTRACT

In recent years, the medical field has witnessed an accumulation of patients' data thanks to the digitization of processes in hospitals and clinics. With the emergence of machine learning, new opportunities have risen in terms of building tools to explore, visualize and model this data in order to facilitate data-driven decision making for the medical profession.

It is with this purpose in mind that the hematology department of the University Cancer Institute of Toulouse (IUCT) has compiled a database of patients with acute myeloid leukemia (AML) spanning over 20 years and keeping track of the evolution of their disease, the treatments they received, and their response to these treatments. Collaborating with the hematology department of CHU Bordeaux, the database contains information about more than five thousand patients, making it one of the most extensive databases in the world regarding AML.

The purpose of this report is to present the work that has been done in exploring and modeling this extensive database. In particular, we will present the performance and results of a neural network we used to estimate probabilities of different outcomes such as early death, relapse, or cure when given a patient's diagnostic data as input. We will compare the performance of this model both with the state-of-the-art studies in AML as well as with the performance of other architectures such as Gradient Boosting algorithms on our dataset, and discuss the limitations encountered. At the time of this writing, we are also designing a multi-modal model that takes as input both tabular data as well as images of microscopy slides of the patients' bone marrow, which we will also present.

ACKNOWLEDGEMENTS

I would like to thank the staff of the Cancer Research Center of Toulouse (CRCT), of the University Cancer Institute of Toulouse (IUCT) and of the Computer Science Research Institute of Toulouse (IRIT). Without them, my integration in the research center would have been much more difficult. By their kindness and their listening, they allowed me to efficiently take my place in the functioning of CRCT and to give the best of myself in the roles that were assigned to me.

In particular, I would like to thank the REVA team at CRCT/IRIT who kindly welcomed me in their fold and have always been available and helpful. I would also like to thank Ambre Santin and Muriel Serthelon for having accompanied me and helped me with all the administrative steps I had to take from the edition of the internship agreement onward.

Special thanks go to my tutors Dr. Sarah Bertoli and David Simoncini, who have been available, attentive and involved throughout this internship. They showed great benevolence and a constant trust that allowed me to demand more of myself and to surpass myself, while granting me an emancipating freedom as to the methods and tracks I chose to explore.

I would like to warmly thank Jean-Marc Alliot (X1983) for putting me in touch with my tutors and allowing me to find this internship. He showed kindness, generosity and availability that make one proud to belong to the Polytechnician family.

I would like to thank all the people at the École polytechnique involved in the smooth running of the third year internships, and in particular Mr. Samy Kara-Mosli and my referent teacher Pr. Patrick Loiseau, who have been available, benevolent and very reactive and whose great involvement has enabled many of my fellow students and myself to obtain an internship agreement in a very short time and to be able to start the internship in the best possible conditions.

Finally, I thank you, who are reading me.

CONTENTS

1 INTRODUCTION

Acute myeloid leukemia (AML) is a cancer of the myeloid line of bone marrow cells, characterized by the rapid growth of abnormal cells that build up in the bone marrow and blood and interfere with normal blood cell production. The underlying mechanism involves replacement of normal bone marrow with leukemia cells, which results in a drop in red blood cells, platelets, and normal white blood cells. Acute myeloid leukemia starts in the bone marrow (the soft inner part of certain bones, where new blood cells are made), but most often it quickly moves into the blood, as well. It can sometimes spread to other parts of the body including the lymph nodes, liver, spleen, and the central nervous system (brain and spinal cord).

As an acute leukemia, AML progresses rapidly and is typically fatal within weeks or months if left untreated, thus emphasizing the importance of a quick diagnosis and treatment. Risk factors can include previous chemotherapy or radiation therapy or myelodysplastic syndrome. Diagnosis is generally based on bone marrow aspiration and specific cytogenetic, molecular and phenotypic tests.

It is important to note that AML has several subtypes for which treatments and outcomes may vary. Regarding treatment, AML is usually treated first with chemotherapy, with the aim of inducing remission. Patients may then go on to receive additional chemotherapy or allogeneic cell transplant. Patients that are deemed too fragile to receive intensive chemotherapy may receive other treatments, such as Azacitidine (AZA). The specific genetic mutations present within the cancer cells may guide therapy, as well as determine how long that person is likely to survive. In particular, when considering whether to give a graft to a patient, it is of vital importance to weight the risks (approx. 25% chance of dying from the graft) against the expected benefits (increased probability of survival). Thus, having models that can predict the probability of survival is of great interest for the field.

Since 2007, the hematology department at the University Cancer Institute of Toulouse (IUCT) has collected data about their current patients with AML, as well as about former patients. The department, working hand in hand with its homologous department at Bordeaux, has built an extensive database of more than five thousand patients with AML. This database contains various information about the patients, ranging from diagnostic data (weight, sex, age or platelet counts in the blood) to the information of the treatments given to the patients and their response to these treatments.

It is in this context that my internship took place. I was tasked with studying the data, and trying to build machine learning models of interest for the medical profession. In particular, a **model to predict and compare the probabilities of the different outcomes** (remission, relapse, death...) given the diagnostic data of a patient and their treatment would for example allow the doctors to weight the different treatments at the time of decision.

2

CONTEXT OF THE INTERNSHIP AND PRESENTATION OF THE FRAMEWORK

2.1 The host entities of the internship

My internship took place in the context of a partnership between the Cancer Research Center of Toulouse (CRCT) and the Toulouse Institute for Research in Computer Science (IRIT).

The Cancer Research Center of Toulouse is a brand-new research center that has been created in 2011 and moved in 2014 on the site of the Oncopole of Toulouse. The CRCT is home to 21 research teams, and more than 300 researchers, clinicians and supports. It also welcomes more than 100 PhD students and post-doctoral fellows. The project has been financed by the Région Occitanie Pyrénées-Méditerranée, Toulouse-Métropole, the French Government, the Inserm, and the CNRS among others.

For its part, IRIT is a joint research unit (UMR) with nearly 600 permanent and non-permanent members, as well as about 100 external collaborators. The research themes of this structure are grouped around five major scientific topics:

- System design and construction
- Real-world digital modelling
- Concept formalization for cognition and interaction
- Environment-aware autonomous adaptative systems
- The transition from raw data to intelligible information

These topics are materialized in six application areas:

- Health, Autonomy, Living, Well-being
- The Smart City
- Aerospace and Transportation
- Social Media, Digital Social Ecosystems
- e-Education for learning and teaching
- Heritage and People Safety

At IRIT, the REVA team addresses issues related to the use of mathematical and algorithmic methods for the analysis of multimodal, visual, and biological data. Its research focuses on techniques for generating, optimizing, and analyzing multidimensional data to produce, calibrate and even control simulated environments, and proposes a wide variety of methods in the context of artificial life paradigms. Like many other IRIT teams, REVA works in partnership with several other institutes and research centers in the context of collaborations on specific projects. Among these partners is the CRCT which hosts a detachment of researchers and computer scientists from REVA, with whom I have been working for the last few months.

2.2 Frameworks and tools used

2.2.1 • Tools used

Programming language and libraries used
Concerning the programming languages and libraries, all of the code for both the exploratory data analysis and building the models was written in Python. The libraries and frameworks used were:

- **pandas**: for handling and operating on the tabular data
- **seaborn** : for data visualization and plotting
- **scikit-learn**: for data pre-processing and for managing k-fold cross-validation
- **xgboost**: to implement a gradient boosting algorithm for classification
- **PyTorch**: for all purposes related to neural networks, from creating a custom dataset class and defining and training the neural networks
- **SHAP**: to plot feature importance using the Shapley Additive Explanation method

Software and specific libraries for slide analysis
For the specific task of slide analysis, we used:

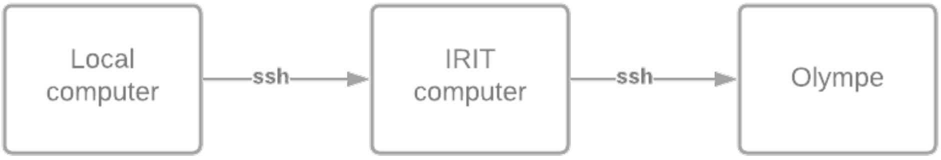
- **OpenSlide** : a library developed to view and manipulate digitized microscopy slides [1]
- **Pathomation**: a viewer to open and read *.mrxs* slides, used to extract the etiquette with the patient's name from the *.mrxs* file;
- **PathAIA**: a library developed at CRCT to process *.mrxs* files and select zones of interest on the images

2.2.2 • Organizing work with Git

To manage the code, as well as facilitate collaboration and code review, the teams at CRCT use **Git**. As such, all the code of the project is currently hosted on a Github repository, which allows for everyone to review code, push commits and pull the latest changes from the server.

2.2.3 • Working with the Olympe supercomputer

Ultimately, to gain time on the training of the neural networks and leverage the power of GPU, all the while ensuring that the data did not leave the CRCT computer I was lent, we connected via **ssh** to Olympe, the supercomputer for the Occitanie region. The GPUs made available through Olympe are NVIDIA Tesla V100. This solution both ensured the best performance as well as confidentiality, as opposed to other solutions such as using Google Colab or similar services.



2.3 Context of our work

2.3.1 • The different stages of AML

AML is a multistage disease, like most cancers (Figure 1). From diagnosis and the choice of an induction treatment to remission and the choice of performing a stem cell transplantation, the questions that the doctors must decide on do evolve according to each stage. The question of transplant, for example, emerges for a patient in remission for whom it is feared that they will relapse. The choice of salvage treatment, or second line of treatment can emerge when a patient has relapsed. As such, it was important to understand the different stages of the disease and how the transitions work between each stage to be able to design meaningful models.

2.3.2 • Objectives of the internship

In this context, I was tasked with the following missions:

- designing models to predict overall survival for a patient, when given their diagnostic data. The idea was to be both able to question the model about the most important variables for prediction and confront them with medical knowledge, as well as compare the expectancy of survival for a given patient in function of treatment ;
- designing models to predict the treatment given to a patient. Here again, the idea was to question the model about the most important variables. A long-term idea would be to deploy the model and allow practitioners in small centers around the world to get an indication of which treatment the experts at the center of Toulouse and Bordeaux would choose.

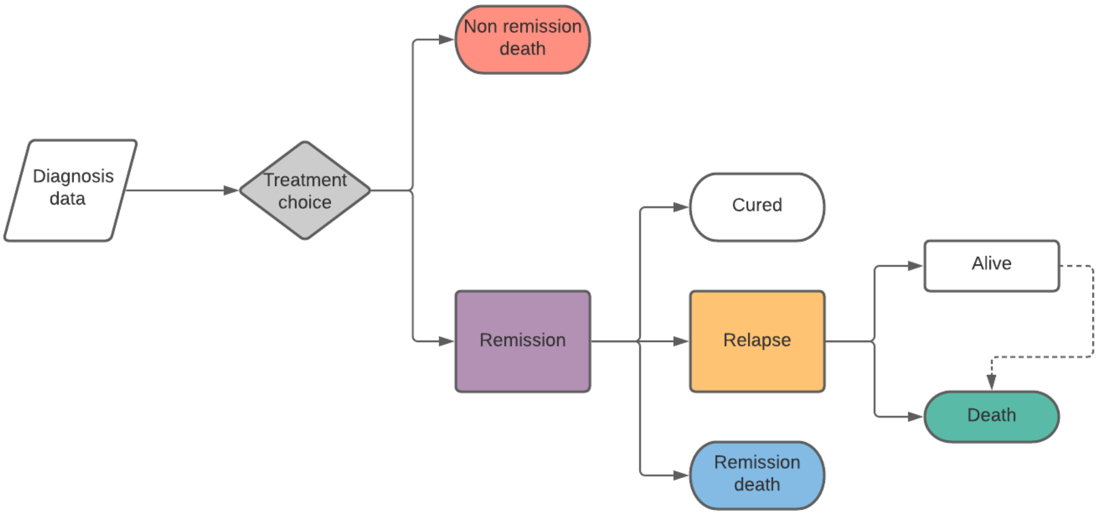


Figure 1: The different stages of AML

We restricted ourselves to diagnostic data for these tasks as the use-case envisioned was that of a patient newly diagnosed.

I was later on during the internship asked to develop a pipeline to use the information from images of microscopy slides in addition to the tabular data, in a multimodal model.

2.3.3 • State-of-the-art

In the field of AML, the use of machine learning methods is relatively new. A recent review of the use of machine learning for AML [2] highlighted some of the most relevant new studies incorporating machine learning in the context of AML diagnosis, prognosis and treatment.

Regarding diagnosis:

- Analyzing >12 000 samples from >100 different studies, Warnat-Herresthal et al [3] combined transcriptomic and genomic data with ML to develop classifiers that accurately detect AML in a near-automated and low-cost method

Regarding treatment and prognosis:

- Morita et al [4] analyzed bone marrow samples of 868 patients with myeloid leukemias (AML, myelodysplastic syndrome [MDS], chronic myelomonocytic leukemia, and myeloproliferative neoplasm) and generated an ML-based model that accurately predicts clinical phenotype based on somatic mutation data
- Siddiqui et al [5] proposed an ML model based on clinical parameters known before treatment that predicts mortality rates for patients undergoing chemotherapy, thereby enabling clinicians to identify patients suitable for intensive induction regimens

- Gerstung et al [6] have used a large dataset combining clinical and genomic data to predict relapse, remission, and overall survival. They have concluded that such datasets in the form of knowledge banks can be used to guide clinicians to precisely tailor a treatment approach for the individual patient

This last study from Gertsung et al was one of the inspiration for our work, as we focused on —without limiting ourselves to— the same objective of providing an accurate prediction of relapse, remission, and overall survival.

3
MATERIALS AND METHODS

3.1 The DATAML database

3.1.1 • Format and particularities of the data

The dataset was provided in the form of a *.sas7bdat* file, which is a binary database storage file. Each row contained information about a specific patient, and the columns corresponded to diagnostic, genomic and treatment information about this patient.

The dataset is comprised of 5116 rows (patients) and 456 columns. The columns can be split in three main categories.

- **diagnostic features**, e.g. *gender, age, platelet counts* or *white blood cells counts*
- **treatment and response to treatment**, e.g. *chemotherapy, transplant, relapse.* These variables describe the evolution of the patient's disease (remission, relapse and the associated time for each event, as well as blood measurements at the time of the event) as well as the treatment they received
- **latest news and updates about the patient**, e.g. *date of latest news, status (alive or dead), date of death (if applicable)*

Among the 456 variables, 122 are diagnostic variables.

The dataset contains features of four types.

- **numeric features**, such as age, weight, hemoglobin levels, or white blood cells counts to mention a few
- **categorical features**, such as gender, information about the existence of previous blood anomalies or the administration of previous treatment, or genomic data, in the form of the presence of certain known driver mutations for acute myeloid leukemia
- **dates**, such as *date of diagnosis, date of relapse...*
- **free text**, such as the doctor's commentaries and observations on the treatment etc.

A few relevant observations should be made about the dataset for a better overview of the problem:

- Medical data is inherently **acquired with time**, both regarding rows and columns. Regarding the rows, the patients are naturally ordered in the data by their order of diagnosis at the CHU. The data spans more than 20 years, with the earliest patients present in the data having been diagnosed in the year 2000 and the latest ones having been diagnosed in 2021. Regarding the columns, for each patient, as their state and response to treatment evolves, columns are completed on the go when this data becomes available. Ultimately, the last impact of time is that with the evolution of treatments and diagnosis technology, the features used to diagnose a patient as well as the treatments used do evolve. As such, some of the variables present among both the 456 columns in general and the 122 diagnostic variables in particular have started being used only recently and as such are not filled for the earliest patients. Conversely, some variables have stopped being used (being replaced by other ones) and are not filled for the newest patients.

- We can also note that **some of the columns are conditional** to others. Indeed, the column *history of therapeutic treatment 2* is conditioned by the value of the *history of therapeutic treatment 1* column and will only be filled if the latter contains a value, otherwise staying blank.
- Another important thing to note is the **important number of features**: there are 122 diagnostic columns in the dataset, which means we should be careful about the **curse of dimensionality**. As such, reducing the number of features via feature selection and feature engineering is important.
- Lastly, it should be noted that for the sake of medical privacy, the **dataset is anonymous**. This is relevant for the computer vision tasks performed on microscopy slides, as a link had to be made between the slides (on which the name of the patient appears) and the database (on which only the initials of the patient as well as their birthdate appears). We will develop this point more when explaining the data processing steps we applied to the slides.

In the light of these preliminary remarks, we can already intuitively apprehend a problem of **sparsity** in the dataset. The acquisition of data through time, the fact that certain columns are recent or, conversely, obsolete and the fact that many columns are conditioned by the value of other columns can participate in this effect.

3.1.2 • Sparsity of the data and correlations between variables

Sparsity of the data

As we mentioned in our preliminary remarks about the data, the data presents some degree of sparsity. As we can see in Figure 2, some columns present a very high percentage of missing data, to the point of being mostly blank. The tendency is that overall, the columns are better filled with the years, as can be seen in Figure 3. In certain cases, columns have started being used recently and can not be retroactively filled. As such, patients diagnosed a long time ago do not have values for these columns, which can participate to the observed effect.

Furthermore, even when considering the possibility of dropping rows to keep a non-sparse dataset, another problem arises. Indeed, the blank values of each column don't necessarily

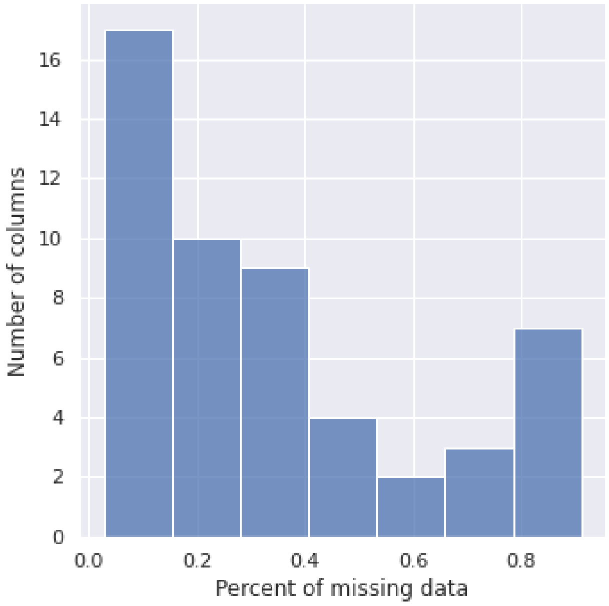


Figure 2: Missing data for the diagnostic variables

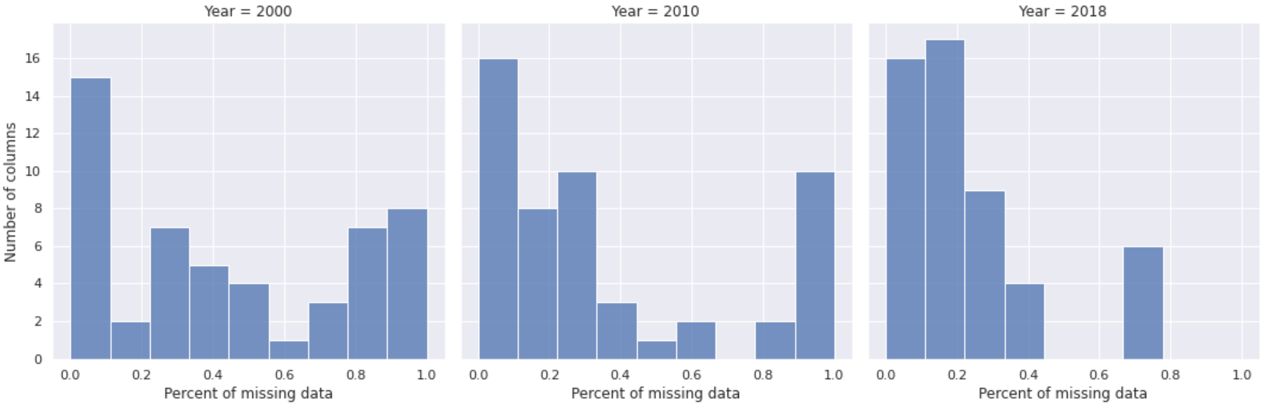


Figure 3: Missing data for the diagnostic variables, for patients diagnosed in 2000, 2010, 2018

occur for the same patients, meaning the blank values for the different columns cover, in the mathematical sense, the set of patients. Even when considering only the subset of the k most populated (or least sparse) columns, we still cover a high percentage of the rows as can be seen in Figure 4.

This is a difficulty since the balance between dropping columns —and then the rows with missing values— to increase density of data or keeping them but with a higher sparsity will be delicate to find. Indeed, when dropping columns, we are potentially dropping useful features, but having sparse data could also harm the models. Finally, even if we drop columns and then the rows with missing data, we will still end up with only a fraction of the patients, which

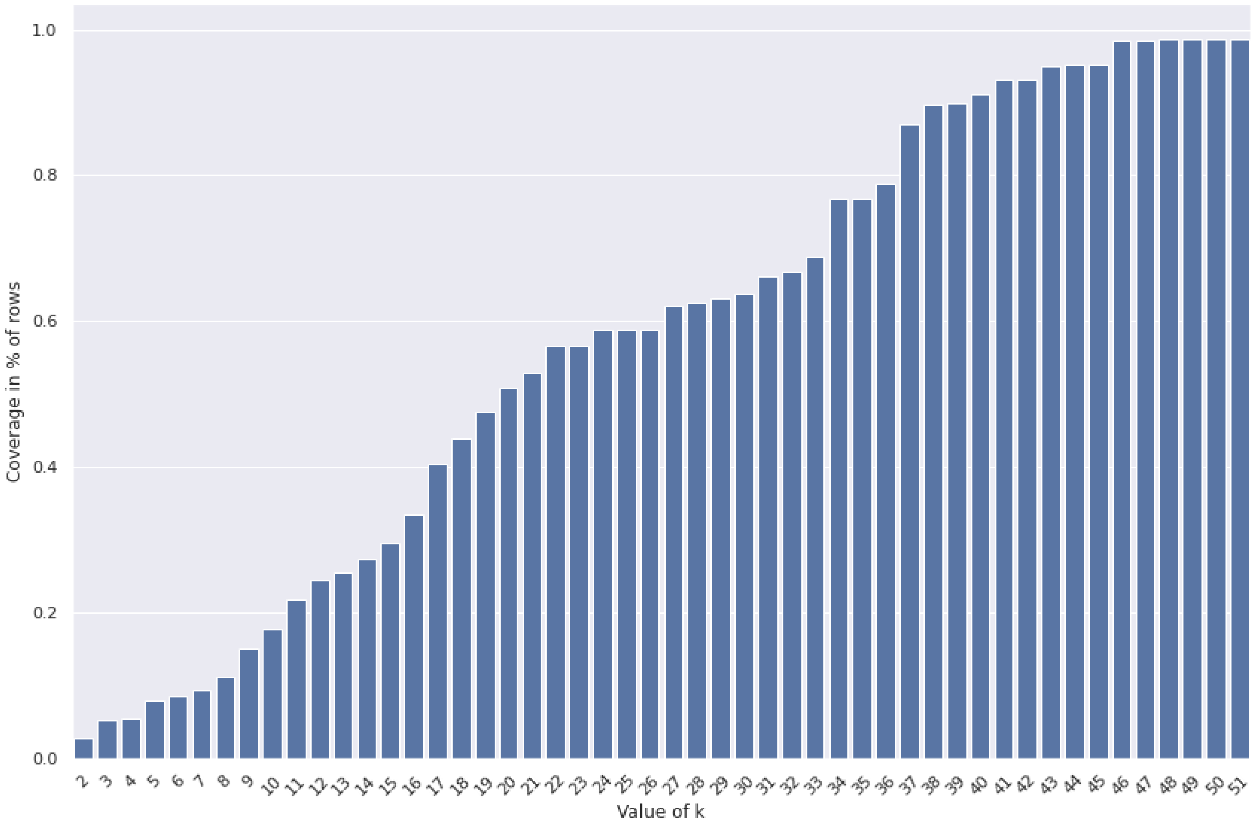


Figure 4: % of rows presenting a missing value for union of k least sparse columns

means training on a smaller dataset, that could furthermore be a non-representative subclass of patients.

Correlation between features

Looking into the correlations between variables, we can compute a heatmap, as seen in Figure 5.

As we can observe, some variables show a high correlation (the NPMC mutation and CD34 marker, or blasts levels in the blood and blasts levels in the bone marrow for example), already known to the doctors. Apart from these cases, variables are generally not strongly correlated (correlation absolute value below 0.5).

We can also plot the evolution of the explained variance after a PCA, to try and get an idea of the underlying dimensionality of the data, as seen in Figure 6.

As we can observe, we don't have a drop in explained variance for any principal component, which makes it hard to define a threshold to project on the underlying vector space.

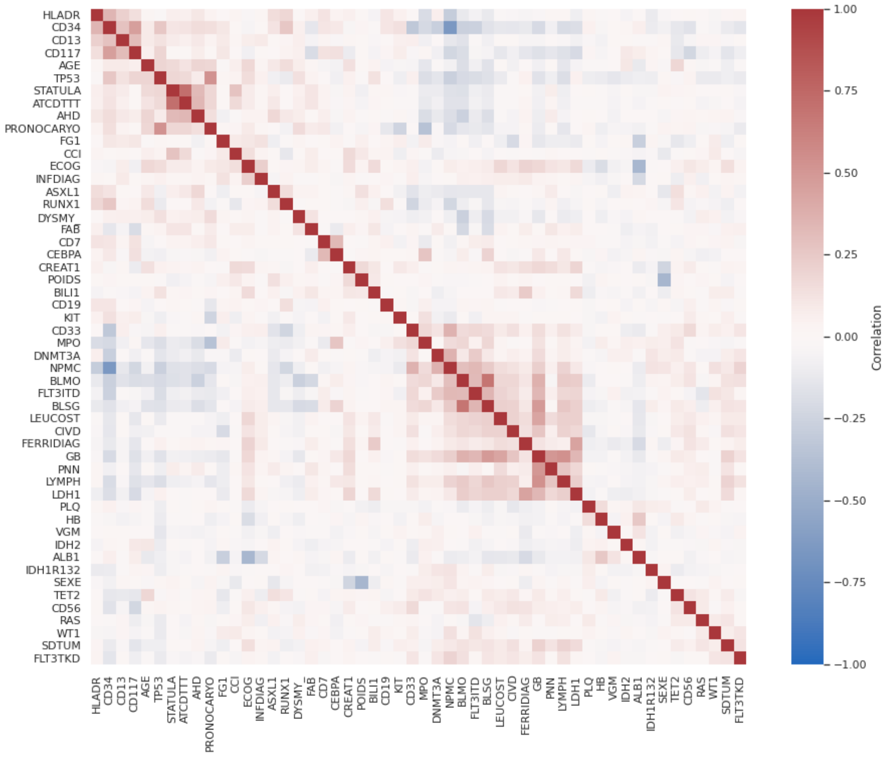


Figure 5: Correlation between diagnostic variables

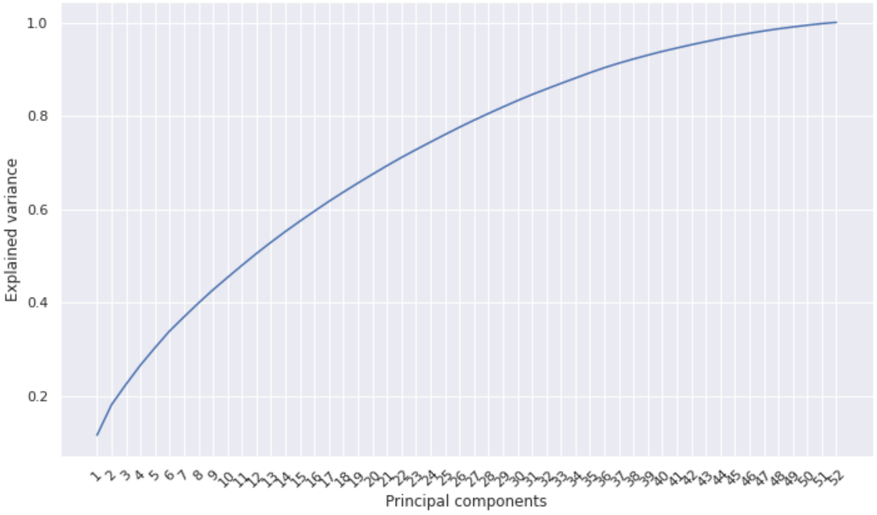


Figure 6: Cumulative explained variance for k principal components

3.1.3 • Different behaviors for different clusters of patients

The different treatments given to patients are not equivalent. We worked with the two most frequent choices of treatment: intensive chemotherapy (ICT) and demethylating agent Azacitidine (AZA).

ICT is more efficient in combating the disease than AZA. When considering patients older than 70, the mortality rate among patients treated with ICT was 79.5% at three years after diagnosis as compared to 94.7% with AZA. However, ICT is also more toxic as a treatment, and some patients can die from this toxicity in the first stages of treatment. As a result, ICT is only given to patients for which the doctors estimate that the benefits outweigh the risks. In practice, ICT is almost systematically given to younger patients up to 75 years old and AZA is only considered for patients older than this threshold or with comorbidities. Even then, if the doctor think the patient can handle ICT, this treatment may be preferred.

Consequently, the cohorts of patients that have received ICT and AZA are very different both in demographics (population age, overall health of the patient at diagnostic) as well as in response to treatment (overall survival is higher among the group that receives ICT) as can be seen in Figures 7, 8. In particular, the distributions of deaths in function of time are very different for ICT and AZA: patients treated with AZA tend to die earlier than those treated with ICT (remark that the first two months are deadlier among the ICT cohort, were the toxicity of ICT can prove fatal to the patients), as can be seen in Figure 7.

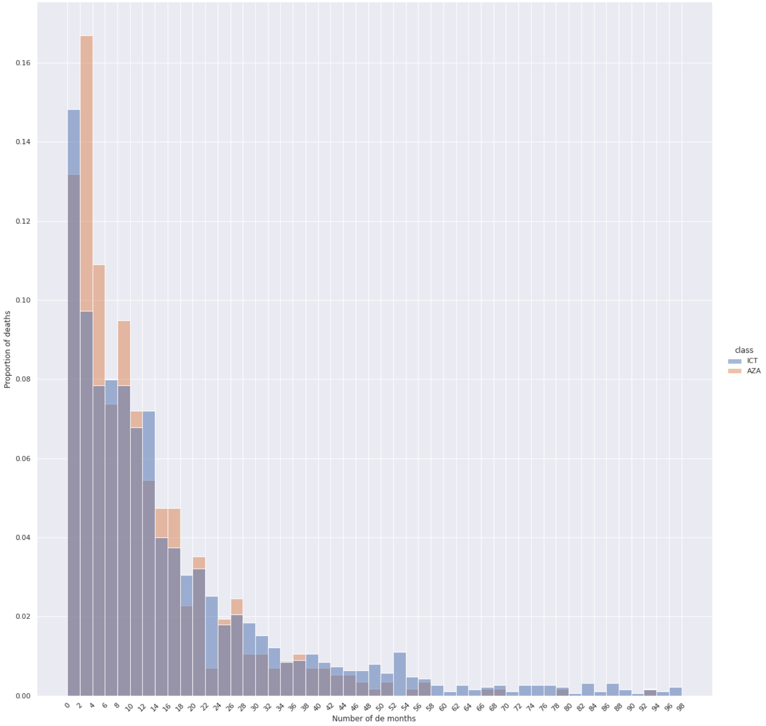


Figure 7: Distribution of deaths among ICT and AZA cohorts with time

These elements led us to distinguish between patients that have received ICT and AZA, and train specific models for each population.

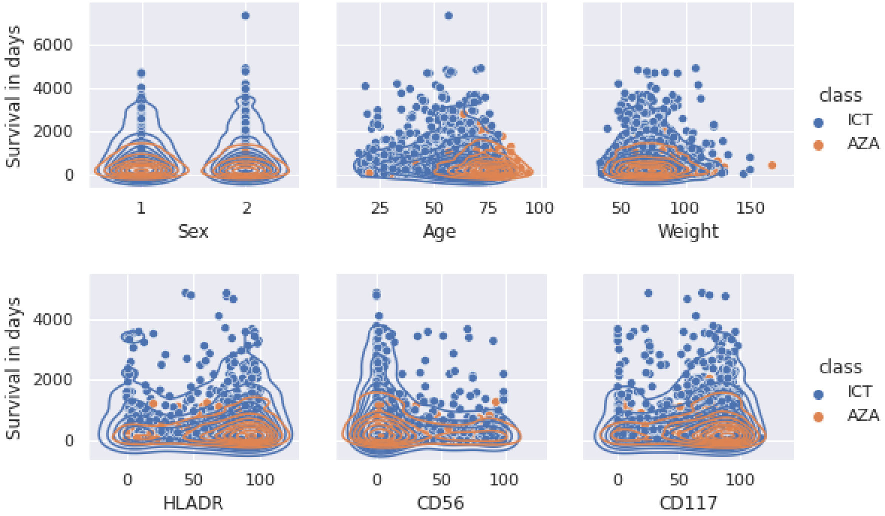


Figure 8: Scatter plot of correlation between some variables and survival time. HLADR, CD56 and CD117 are cell surface markers, given in % of expression. Kernel density estimations are also represented.

3.2 Data pre-processing

Given the precedent remarks regarding the structure and properties of the data as well as the important number of features, a particular importance had to be put in the data pre-processing step. The work that we conducted can be roughly split in three complementary approaches to the data:

- feature and patient selection
- feature engineering
- image processing on the microscopy slides

3.2.1 • Feature and patient selection

First was a work on selecting the important features. Indeed, with 122 columns, it was important to reduce the number of features if possible to reduce the impact of the curse of dimensionality, and to avoid harming the overall quality of the model by risking having many correlated features.

To select the features, we dug in the data and made the following observations:

- some columns where duplicated (appeared twice in the data, although they were meant to appear only once in the thesaurus)
- a few columns were empty
- some categorical columns contained only one value
- some columns contained redundant information (birth date and age for example)
- some columns contained information irrelevant for our purposes (e.g. patients' initials)

These columns were dropped from the data.

Furthermore, not all patients were kept for our analysis. Indeed, as we choose to work with ICT and AZA, we dropped the patients that received neither of these treatments. In addition, some patients had a rare form AML (called type 3 AML) of excellent prognostic and where survival rates are very high (near 95%). These patients were also dropped from the data as they are outliers.

3.2.2 • Feature engineering

The next step was curating these diagnostic variables. Indeed, some of these variables presented high sparsity that needed reducing and many variables were conditioned by others. For example, the columns for *history of therapeutic treatments 2* and *history of therapeutic treatments 3* were only filled if the column *history of therapeutic treatments* was itself filled and there was indeed more than one therapeutic treatment in this patient's medical history.

To circumvent the problem, we discussed pertinent ways to combine columns with Dr. Bertoli. For example:

- The categorical columns **NRAS** and **KRAS** both contained information about mutations of the RAS gene. We chose to fuse these columns in a single column **RAS** that simply contained the information about whether or not one of the RAS mutation could be found in this patient's genes (0- no RAS mutation, 1- NRAS or KRAS mutation)
- For the columns **SDTUM** (4 columns): these columns contained information about the tumoral syndrome of the patient. We kept a single column with a new label *2=multiple*. In other words, if the patient had several tumoral syndromes (in the columns SDTUM_1, SDTUM_2, SDTUM_3, SDTUM_4), we gave him the label 2, otherwise we give him label 1 (only 1 syndrome, previously from 1-8) or 0 if he had 0. We obtained the encoding 0-no, 1- one tumoral syndrome and 2- multiple.
- The columns **CIVDCLINDIAG** and **TYPECIVDCLINDIAG** contained information about the presence of a disseminated intravascular coagulation (DIC) and the type of DIC. We merged them using now 3 labels: 0- no, 1-Hemorrhage, 2-Thrombosis (i.e. replacing 1-yes by the type of DIC)
- The columns **DYSMY** and **MRCCYTO** both contained information about dysmyelopoiesis in the cells. MRRCYTO is a preferred indicator by the doctors. We merged them as fol-

lows. If MRCCYTO was filled in, we kept this value (0-no, 1-yes), otherwise we kept the value DYSMY (0-no, 1-yes)

- The columns **ATCDTTT** (3 columns) contained information about history of therapeutic treatment. We merged the three columns as follows. If ATCDTTT_1=0, we assigned label 0. If one of the three columns contained a value among 1,2,3,6, we assigned label 1 as they pertain to similar types of treatments. Otherwise we assigned label 2 (or NaN when the three columns were NaN).

This feature engineering allowed us to reduce the number of features to 52 while increasing the overall quality of the columns (decreasing the average sparsity from 43% to 29%).

3.2.3 • Image processing on microscopy slides

Finally, we discussed the possibility of augmenting our data by adding the information of microscopy slides of the patients' bone marrow, a sample of which is represented in Figure 9. As our data was tabular, we had to devise a pipeline to take into account the visual information from the slides.

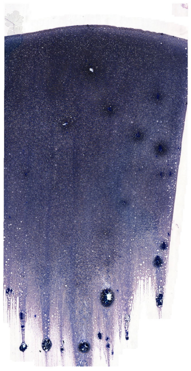


Figure 9: Microscopy slide of a patient's blood

For this purpose, we first had to do some preprocessing. Indeed, the slides are digitized in a proprietary format by the scanner. Each file is of a size of several gigabytes (up to 15Gb per image) as the resolution is very high. Furthermore, each slide is accompanied by the name of the patient and not by any other information that could link the slide to our database. However, as we previously stated, the database is anonymous. As such, the preprocessing has to be done in several steps, and the protocol that we chose to apply was the following:

1. We used the open-source viewer Pathomation to extract the etiquette associated to the slide, which contains the patient's name

2. We used the Tesseract engine [7] to perform optical character recognition (OCR) on this image and automatically extract the name of the patient

3. A member of the medical staff, who is authorized to look up the patients by name, made the correspondence between the slides and the dataset entries by matching *slide_id* and *patient_id* via the name

We can then concentrate on extracting the information from the images. As the sheer size of the images doesn't allow to directly train a neural network on them, we have to devise another protocol.

1. Using the PathAIA library, we extract from each image a patch (sub-image) that contains an exploitable density of cells (not too high, unlike the top of Figure 9). This selection is performed on the thumbnail of the slide (in low resolution) for speed and efficiency
2. We extract in full resolution the patch selected at the previous step
3. We match the patches and labels and load them in a custom dataset created with PyTorch, implementing the dataset class

A limitation for this work is the time and money needed to scan a slide. As of the time of this writing, the hospital has provided us with 100 slides, which was enough to devise the pipeline and test it, but is too limited to perform training and learning. The objective was in this case to give all the algorithms ready-for-use to the hospital for the time when the rest of the slides are scanned.

3.3 Models

We compared the performances of a gradient boosting algorithm (XGBoost) and two neural networks architectures: a multilayer perceptron (MLP) and a neural oblivious decision ensemble model (NODE).

We used 5-fold cross-validation on the dataset to evaluate all the models.

The Shapley Additive Explanations method (SHAP) was used to showcase the importance and influence of variables on the predictions. The Boruta algorithm was then used to extract the most important features for prediction.

3.3.1 • XGBoost

The first model we trained and evaluated was XGBoost, a gradient boosting model [8].

Gradient boosting works by building an ensemble model out of weak learners. At each step, a weak learner (for example a decision tree) is fit in a way to minimize the error of the ensemble when added to the model, as illustrated in Figure 10. It applies the principle of gradient descent in the space of functions, given that the weak learner that is added is fit to approach the gradient of the loss over the predictor function.

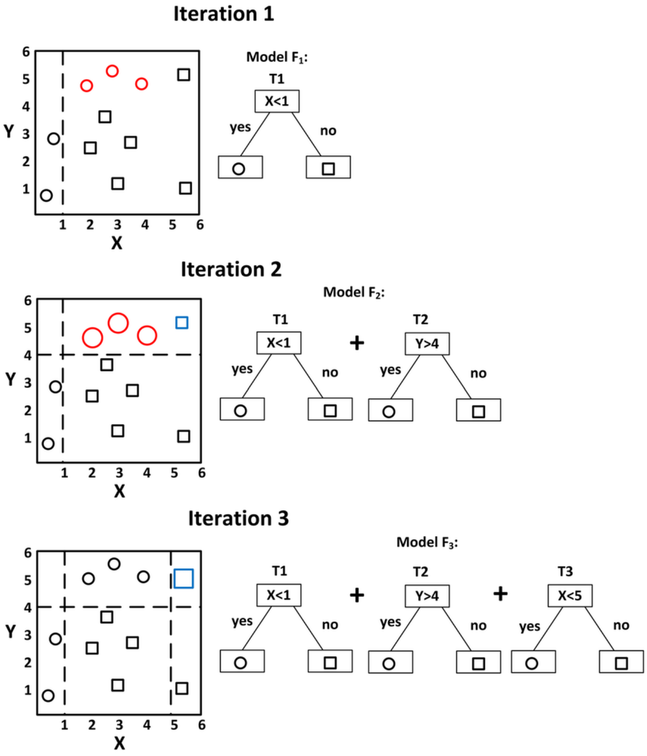


Figure 10: Illustration of the principle of Boosting algorithms

We choose to first evaluate the performances of such a model because gradient boosting is often the state-of-the-art when it comes to machine learning on fully tabular data [9].

To optimize the choice of hyperparameters, we used a grid-search algorithm, and calibrated XGBoost by modifying the following parameters:

- n_estimators: the number of gradient boosted trees. Equivalent to the number of boosting rounds
- max_depth: maximum tree depth for base learners
- learning_rate: the boosting learning rate
- gamma: the minimum loss reduction required to make a further partition on a leaf node of the tree. The larger gamma is, the more conservative the algorithm
- subsample: subsample ratio of the training instances. Setting it to 0.5 means that XGBoost would randomly sample half of the training data prior to growing trees. This is to prevent overfitting. Subsampling occurs once in every boosting iteration
- scale_pos_weight: controls the balance of positive and negative weights, useful for unbalanced classes

The *max_depth* and *gamma* parameters allow to directly control model complexity to avoid overfitting.

The *scale_pos_weight* parameter allows to account for an imbalanced dataset and improve model performance in this case.

3.3.2 • Neural networks

We tried two different neural networks architectures: a multilayer perceptron (MLP) and a Neural Oblivious Decision Ensemble model (NODE).

MLP

The first architecture we used was a multilayer perceptron, illustrated in Figure 11. We empirically settled for a configuration consisting of three linear layers with ReLU non-linear activation and dropout layers between each linear layer (except for the last layer).

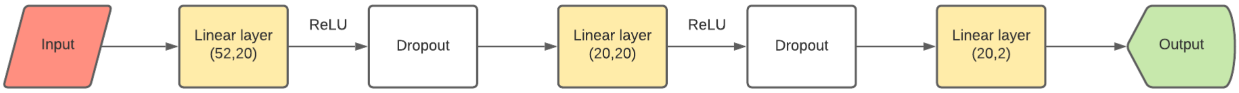


Figure 11: Schema of the multilayer perceptron

NODE

The second architecture we implemented was the Neural Oblivious Decision Ensemble architecture [10]. This architecture has been designed to generalize ensembles of oblivious decision trees, while benefiting from both end-to-end gradient-based optimization and the power of multi-layer hierarchical representation learning. In their paper, the authors showed results that approached or outperformed the state-of-the-art gradient boosting algorithm XGBoost on several different tabular datasets. We used an open-source PyTorch implementation. An illustration of a NODE layer is presented in Figure 12.

3.4 Feature extraction and importance

We used the Boruta algorithm [11] to extract the most relevant features from our set of 52 diagnostic variables, and retrained our models with the subsets of features selected by Boruta for comparison with training on all features. The Boruta algorithm proceeds in the following way:

- Each feature is replicated, and the values of the replicated variables are then randomly permuted. These new variables are called shadow features. All correlations between the shadow features and the labels to predict are random by design

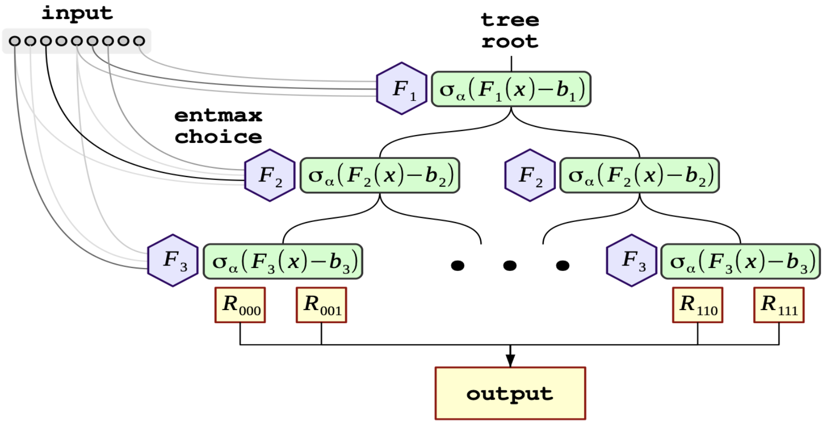

input
tree root
entmax choice
F_1
$\sigma_\alpha(F_1(x)-b_1)$
F_2
$\sigma_\alpha(F_2(x)-b_2)$
F_2
$\sigma_\alpha(F_2(x)-b_2)$
F_3
$\sigma_\alpha(F_3(x)-b_3)$
F_3
$\sigma_\alpha(F_3(x)-b_3)$
R_{000}
R_{001}
R_{110}
R_{111}
output

Figure 12: Illustration of a NODE layer, with its single oblivious decision tree. The splitting features and the splitting thresholds are shared across all the internal nodes of the same depth. The output is a sum of leaf responses scaled by the choice weights.

- Several classification runs are performed, and for each run the importance of all features is computed. The shadow features are randomized before each run, and therefore the random part of the system is different for each run
- A feature is deemed important for a single run if its importance is higher than the maximal importance of all shadow features
- A statistical test is performed for all features. The null hypothesis is that the importance of the variable is equal to the maximal importance of the shadow features. For each feature we count how many times the importance of the feature was higher than this max (a hit is recorded for the variable). Variable is deemed important (accepted), when the number of hits is significantly higher than the expected value, and is deemed unimportant (rejected), when the number of hits is significantly lower than the expected value
- Variables which are deemed unimportant are removed from the information system, usually with their randomized mirror pair. The procedure is performed for a predefined number of iterations, or until all attributes are either rejected or conclusively deemed important, whichever comes first.

To showcase feature importance, we used the Shapley Additive Explanation method, or SHAP [12]. Shapley values are a concept of the cooperative game theory field, whose objective is to measure each player's contribution to the game. The method for obtaining Shapley values was proposed by Lloyd Shapley in 1953 [13]. In game theory, a Shapley value is the average marginal contribution (or added contribution) of player among all possible coalitions initially excluding this player. The formula for calculating the Shapley value is thus:

$$\varphi_i(v) = \frac{1}{n_{players}} \sum_{S \subseteq N \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

Where v is a function such that $v(S)$ is the total expected sum of payoffs the members of S can obtain by cooperation and $n_{players}$ is the number of players.

In the case of machine learning, players are the features. Hence, the Shapley value is the average marginal contribution of a feature to the prediction, and each Shapley value represents the impact that the feature generates in the prediction [1].

[1] In practice, the v function is a conditional expectation function of the model. The Shapley values attribute to each feature the change in the expected model prediction when conditioning on that feature.

4
RESULTS

4.1 Results for predicting overall survival

For prediction of overall survival, we chose to train and evaluate models for different time points. For a given time T, each model was trained to predict if a patient would be dead or alive after T, starting from their time of diagnosis.

In our cohort, 3030 patients (82.2%) received ICT and 657 (17.8%) AZA as first line treatment. Median overall survival (OS) was 19 and 9 months, respectively.

We compared the results obtained with those of a naive predictor, that systematically predicts the majority issue (Figure 13, 14).

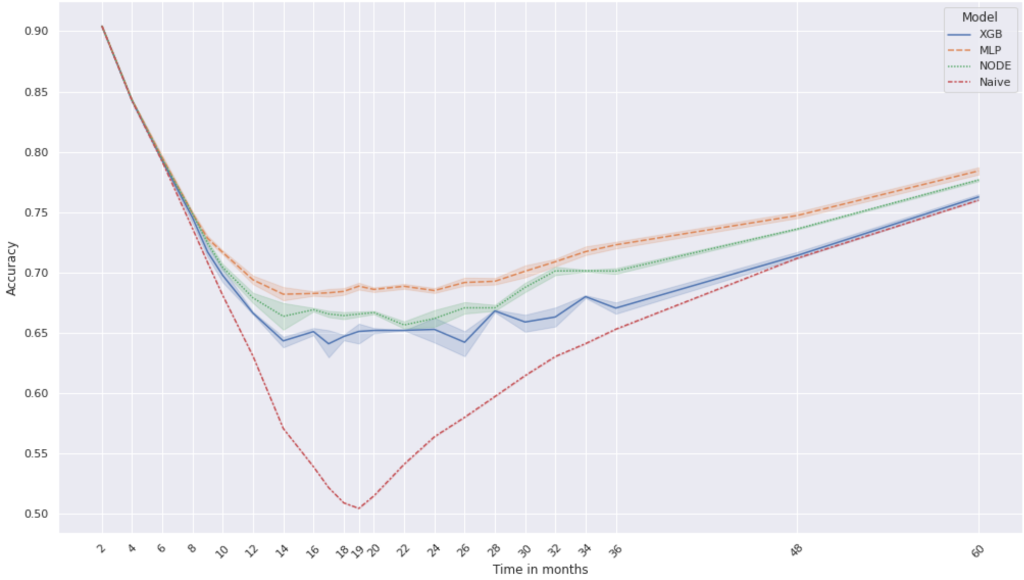


Figure 13: Comparison of models accuracy with time for ICT

We observed that both for the ICT cohort and the AZA cohort, the improvements of our models over a naive predictor were maximum for the median time of survival. We found that neural networks outperformed XGBoost consistently on our dataset.

Regarding the ICT cohort, we achieved an accuracy of 68.5% for predicting OS at the 19-month mark, an improvement of 17.5% over a naive predictor. The AUC of our model was 67.7%.

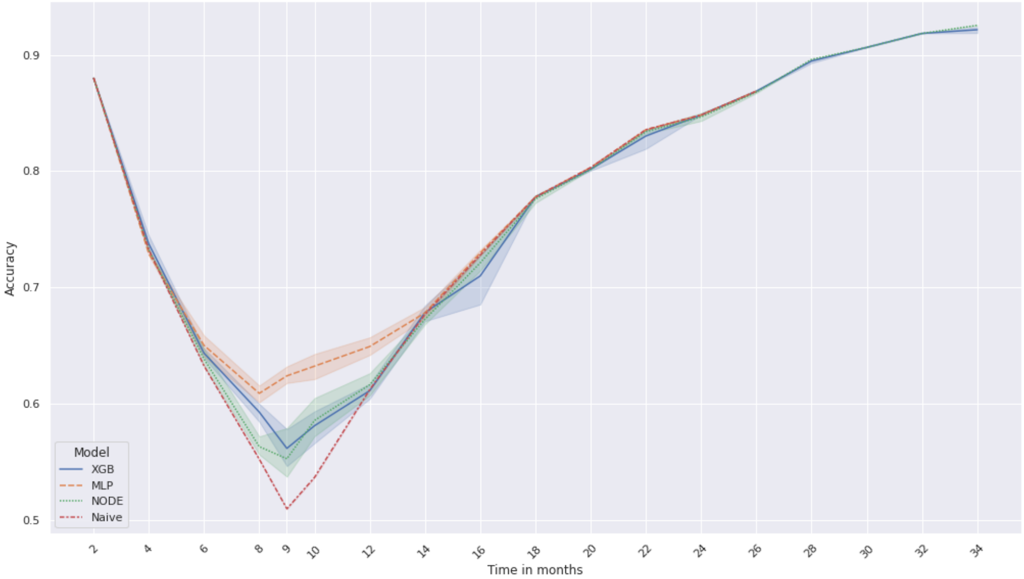


Figure 14: Comparison of models accuracy with time for AZA

The Boruta algorithm selected 13 variables as the most importants (Figure 15), with decreasing order of importance: age, cytogenetic risk, white blood cells counts (WBC), LDH, platelets count, albumin, MPO, mean corpuscular volume, CD117, NPM1 mutation, AML status, multilineage dysmyelopoiesis, ASXL1 mutation (Figure 15). When training with only these 13 variables, we achieved an accuracy of 67.8%.

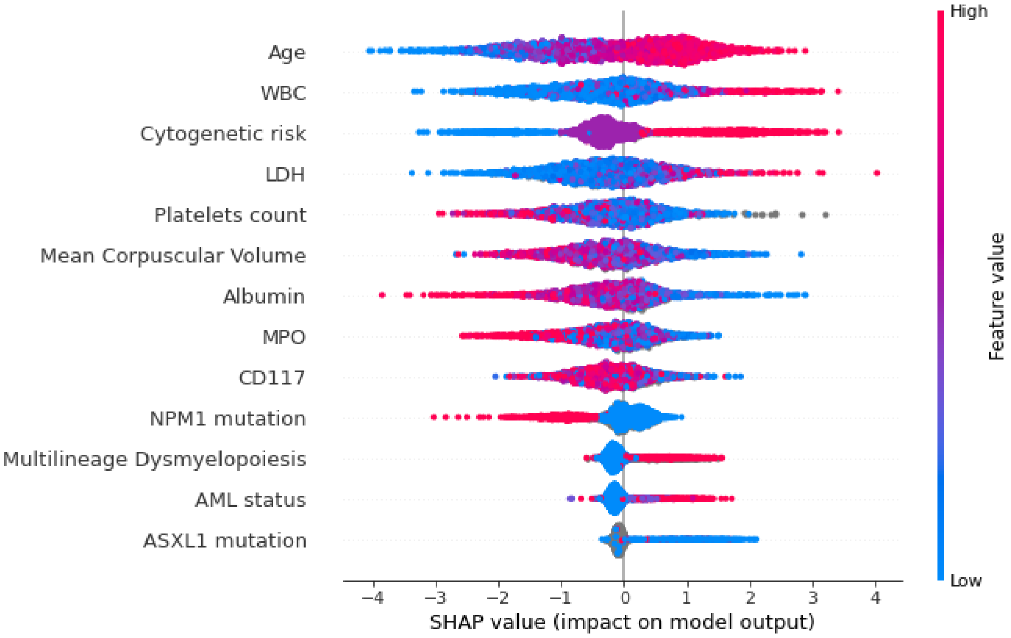


Figure 15: Shapley values for IC at median time

In the AZA cohort, we achieved an accuracy of 62.1% on predicting OS at the 9-month mark, an improvement of 11.1% over a naive predictor. The AUC of our model was 61.2%. Here the Boruta algorithm selected only 7 variables: blood blasts, serum ferritin, CD56, LDH, hemoglobin, CD13 and the presence of a disseminated intravascular coagulation. When training with only these 7 variables, we achieved a 61.9% accuracy.

4.2 Results for predicting treatment

We then designed models to predict the best treatment between ICT and AZA for the 1032 patients older than 70 years.

We observed that again, neural networks outperformed XGBoost on our dataset (Figure 16).

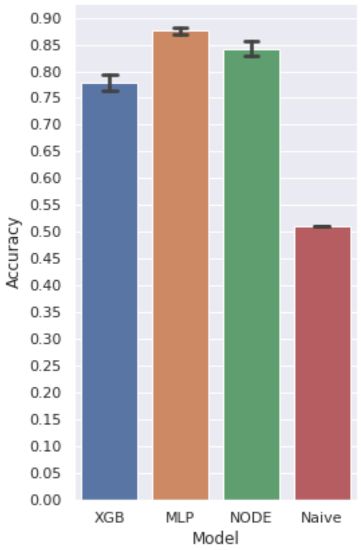


Figure 16: Accuracy for prediction of treatment

We achieved a 88% accuracy, which is 37% more than a naive predictor given the distribution of the cohort: 51% having received ICT and 49% having received AZA. The AUC for this model was 87.6%.

For this model, 12 features out of 54 were selected by the Boruta algorithm as the most important: age, TP53 mutation, bone marrow blasts, AML status, disseminated intravascular coagulation, blood blasts, cytogenetic risk, IDH2 mutation, IDH1 mutation, presence of an infection at diagnosis, ASXL1 mutation and presence of leukostasis.

The accuracy when training with these 12 variables was 87%.

4.3 Results for predicting relapse, remission

For prediction of remission and relapse, the task proved more difficult for the algorithms we tested. When predicting from the diagnostic variables between remission or non-remission death, we did not improve accuracy over a naive predictor.

We then tried to consider supplementary variables for this task. For predicting remission, we considered variables measured during induction in addition to the diagnosis variables. Indeed, the induction phase happens before remission or non-remission in the AML stages. Similarly, for predicting relapse, we considered variables measured during induction and after remission in addition to the diagnosis variables. This is because relapse happens after remission in the stages of AML (Figure 1). With these new sets of variables, the accuracy we obtained was 66.4% for remission. This accuracy is 2.4% more than the naive predictor. For relapse, we obtained an accuracy of 71.2%, which is 7.2% better than the naive predictor.

5
DISCUSSION AND ROADMAP FOR CONTINUING THIS WORK

5.1 Discussion

Using state-of-the-art models and a systematic approach (grid-search for XGBoost, k-fold cross-validation for all models), we have encountered a ceiling in the accuracy for the prediction of overall survival. Our approach could be tested on other datasets in order to confirm whether this is a common problem when trying to predict OS for AML or if it is pertaining to our dataset. In the first case, many explanations could be envisioned:

- The dataset cannot and does not cover all relevant information about a patient (lifestyle, healthiness, diet. . .) which could affect overall survival
- The evolution of practices and treatments over the last 20 years mean the patients in the dataset did not have equal chances at the time of their diagnostic
- There might a part of inherent unpredictability regarding survival

Furthermore, the fact that accuracy only slightly changes after feature selection with Boruta hints that all the variables may not be relevant for predicting OS, and that a small subset of optimized variables can allow for almost the same results.

In their paper, Gerstung and al. obtained a concordance of 72% for overall survival at 3 years. Concordance is defined as the probability that the survival times of two individuals are concordant to the observation. A pair of observations i, j is considered concordant if the prediction y and the data x go in the same direction, i.e.:

$$((y_i > y_j) \wedge (x_i > x_j)) \vee ((y_i < y_j) \wedge (x_i < x_j))$$

It would be interesting to compute and compare the accuracy obtained by Gertsung and al. with their statistical baseline or naive predictor.

Awada and al. integrated cytogenetic and gene sequencing data from a multicenter cohort of 6,788 AML patients and identified 4 unique genomic clusters of distinct prognoses by applying Bayesian Latent Class method. Extracting invariant genomic features driving each cluster, they obtained a 97% cross-validation accuracy when used for genomic subclassification. Subclasses of AML defined by molecular signatures overlapped current pathomorphological and clinically defined AML subtypes. It would be interesting to use their classification of AML subtypes as a feature for our cohorts, to see if it improves our models.

5.2 Multimodal model

At the time of the writing of this report, a number of 100 microscopy slides have been scanned and matched to the database. This limited number, even with data augmentation, means that to fully assess the potential of the multimodal model, the remaining few thousands slides will need to be scanned and processed.

It will be necessary to assess the performance of different neural architectures, and to settle for the one yielding the best results.

The work on the sub-image selection algorithm can be continued to refine the selection of the patches of tissue. We have chosen a simple algorithm that works with the "intensity" of color in the considered zone and computes its distance to an objective in order to select the target zone. We could refine the process and do image segmentation to evaluate the density of cells in a zone and match it to a desired cell density.

It will also be important to consider at which level to look for information. At the time of this writing, we have chosen the level of maximum resolution, at which the cells appear most clearly. It can be interesting to implement an algorithm to consider both a low-level perspective (micro-information) and a higher-level perspective where cells do not appear clearly but where macro-information could be seen more easily. An idea could be to feed both images to neural networks and to concatenate the features extracted before feeding them to a classification layer or to train fully separate neural networks and to ensemble the predictions.

5.3 Accounting for the newest practices in the field

Treatments in the field of medicine are constantly evolving in accordance with advances and discovery of new drugs. In our case, the treatment proposed for patients for which chemotherapy is not suitable has seen an evolution since 2020. A combination of Azacitidine and Venetoclax is now preferred to the exclusive use of Azacitidine, as it has been noted to increase chances of survival and remission.

It will be necessary to proceed with data exploration and analysis and re-train our predictive models to account for this shift when the database will contain enough patients treated with Azacitidine + Venetoclax.

Specifically, the question of treatment for patients over 70 does no longer consist of choosing between ICT and AZA but between ICT and AZA+VEN. Furthermore, this question is more relevant than ever given the improvements of combining AZA+VEN, bringing this less toxic treatment closer to the efficiency of ICT.

6
CONCLUSION

During this internship, we have used the DATAML database and designed models for predicting overall survival and the treatment choice of an expert hematologist for AML patients. Our models show that the best results are obtained at median time for OS, and that a reduced subset of variables approaches the performances of the models trained on all features.

On a personal level, this third-year internship was an opportunity for me to familiarize myself with the functioning of a research institution, and with the challenges of the research profession. Through the various team meetings in which I participated, I was able to observe and learn about the operating processes of such an organization. The lessons I learned from this experience - whether in terms of communication, project management or resilience - will be useful in my future life.

I have the satisfaction to have acquired skills that will be useful for the continuation of my studies as well as for my professional life. Working with Git in a collaborative environment, systematizing the methods used to increase the robustness of our models as well as avoiding arbitrary choices (of hyperparameters for example) and the formative experience of writing an abstract and working on a paper are valuable experiences that I am glad to take away from this internship.

Perhaps most importantly, I have the gratifying feeling that I have progressed as a person, in contact with extremely caring and competent colleagues, and that I have grown from this third year internship.

REFERENCES

[1] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan, “Openslide: A vendor-neutral software foundation for digital pathology,” *Journal of pathology informatics*, vol. 4, p. 27, 09 2013.

[2] J.-N. Eckardt, M. Bornhäuser, K. Wendt, and J. M. Middeke, “Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects,” *Blood Advances*, vol. 4, no. 23, pp. 6077–6085, 2020.

[3] S. Warnat-Herresthal, K. Perrakis, B. Taschler, M. Becker, K. Baßler, M. Beyer, P. Günther, J. Schulte-Schrepping, L. Seep, K. Klee, *et al.*, “Scalable prediction of acute myeloid leukemia using high-dimensional machine learning and blood transcriptomics,” *Iscience*, vol. 23, no. 1, p. 100780, 2020.

[4] K. Morita, F. Wang, H. Makishima, Y. Yan, T. Yoshizato, K. Yoshida, B. P. Przychodzen, K. Patel, C. E. Bueso-Ramos, C. Gumbs, *et al.*, “Pan-myeloid leukemia analysis: machine learning-based approach to predict phenotype and clinical outcomes using mutation data,” *Blood*, vol. 132, p. 1801, 2018.

[5] N. S. Siddiqui, A. Klein, A. Godara, C. Varga, R. J. Buchsbaum, and M. C. Hughes, “Supervised machine learning algorithms using patient related factors to predict in-hospital mortality following acute myeloid leukemia therapy,” 2019.

[6] M. Gerstung, E. Papaemmanuil, I. Martincorena, L. Bullinger, V. I. Gaidzik, P. Paschka, M. Heuser, F. Thol, N. Bolli, P. Ganly, *et al.*, “Precision oncology for acute myeloid leukemia using a knowledge bank approach,” *Nature genetics*, vol. 49, no. 3, pp. 332–340, 2017.

[7] R. Smith, “An overview of the tesseract ocr engine,” in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, pp. 629–633, 2007.

[8] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[9] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, “An up-to-date comparison of state-of-the-art classification algorithms,” *Expert Systems with Applications*, vol. 82, pp. 128–150, 2017.

[10] S. Popov, S. Morozov, and A. Babenko, “Neural oblivious decision ensembles for deep learning on tabular data,” *arXiv preprint arXiv:1909.06312*, 2019.

[11] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, “Boruta - a system for feature selection,” *Fundam. Informaticae*, vol. 101, p. 271–285, Dec. 2010.

[12] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777, 2017.

[13] L. S. Shapley, *17. A Value for n-Person Games:*, pp. 307–318. Princeton University Press, 2016.