Master in Artificial Intelligence
Master Thesis

Random Forest as a Tumour Genetic Marker Extractor

Raquel Leandra Pérez Arnal

Supervisors: Dario Garcia-Gasulla and Ulises Cortés

October, 2019

Universidad Politécnica de Catalunya (UPC)
Universitat de Barcelona (UB)
Universitat Rovira i Virgili (URV)

Facultad d'Informática de Barcelona (UPC)
Facultat de Matemátiques (UB)
Escola Técnica Superior d'Enginyeria (URV)

Gracias a Dario Garcia-Gasulla por ayudarme a corregir y revisar la tesis, a pesar de que ello implicase leerla una respetable cantidad de veces. A Davide Cirillo por su ayuda respecto a como evaluar los resultados. Al Doctor Adrián Puche Gallego por ayudarme a entender parte de los papers médicos de contenido altamente específico.

Abstract

Identifying tumour genetic markers is an essential task for biomedicine due to its applicability to cancer detection and therapy development. Several studies have shown that chromosome rearrangements can be a risk factor for tumour generation. Chromosome rearrangements are order or content alterations in the genetic material of an individual.

In this master thesis, we analyse a dataset of chromosomal rearrangements and present a methodology for extracting new genetic markers. We transform the genetic marker problem into a feature selection one, by engineering gene-related features from the data. These engineered features correspond to potential cancer genetic markers. To achieve so, we use a Random Forest Classifier. With it, we measure feature relevance and create a ranking of genetic marker candidates. To validate the performance of our model, we first validate that the aggregated ranking has consistent results. Independently we also validate the model by its discriminant capacity. Finally, we perform a validation with a query-based evaluation and present our conclusions.

Keywords

Cancer Research, Chromosomal Rearrangements, Tumour Genetic Markers, Random Forest

Contents

1. Introduction

Cancer is among the four current leading causes of death before the age of 70, having around 18.1 million deaths in 2018 [1]. For this reason, studying and understanding the biology of tumours constitutes a priority in biomedicine. One of the leading research lines on this field is the study of chromosomal rearrangements in solid tumour cells. Chromosomal rearrangements (or *breaks*) are changes in the basic structure of a chromosome. Examples of such alterations are the deletion, duplication or reordering of a subset of bases within a chromosome.

Several studies have shown that the presence of chromosomal rearrangements in tumours is often correlated with poor prognosis [2, 3], and some of them have been identified as hallmarks of certain tumour types. These studies imply that the presence of some specific gene expressions or DNA changes, like chromosomal rearrangements, can be used as tumour markers to characterise different types of cancer. Identifying these markers can be used, for example, to predict disease outcome or response to treatment. Some examples of chromosomal rearrangements which are known to be tumour genetic markers are mutations of chromosome 5q21 for colorectal cancer [4] or deletions on chromosome 3p for lung cancer [5].

Usually, these genetic markers are found using medical or genetic experimentation, which is slow and expensive. Machine Learning (ML) is one of the fields that can better help at this task, given its capacity to find and exploit patterns in data. With this idea in mind, in this thesis we present a new methodology to find potential tumour genetic markers from a dataset of chromosomal rearrangements. This data consists of a sequence of chromosomal rearrangements for every patient within a set of 2,586 cancer patients. Each rearrangement is represented as triplets *source base pair*, *destiny base pair* and *rearrangement type*. The complete dataset, which contains 2,586 patients with an average of 115 rearrangements per patient, is analysed and explained in §5.

Working at a base-pair level is unfeasible given the number of samples that this dataset contains, as there are lots of base-pairs in a genome (around 3 billion) for only a few samples (298,104 total rearrangements within all the samples). In §6 we handle this issue by engineering features at a more general level and by pre-processing the dataset.

After cleaning the dataset, we train a Random Forest classifier using the germinal layer of the cancer type in each patient as a target variable, and the engineered features as input variables. The germinal layers are the first categorical differentiation of cells during embryonic development. As explained in §4, the germinal layers can be used as a generalisation of the cancer type (the cancer classification based on the organ where the tumour has been formed). The discussion and selection of the machine learning model can be found in §7. In §8 we present the training configuration of the chosen model. Considering the inherent stochasticity of Random Forest model training, we aggregate the results of 500 independent trainings on a single feature ranking. The best features from the unified ranking are new potential genetic markers found by the methodology. In our experiments, we extract more than thirty potential markers (see Table 7).

An experimental evaluation of the potential markers identified in this work is unfeasible, due to its associated costs. Also, there were no studies on the same dataset to cross-check our findings, as it has only been made available a few months after the writing of this thesis. As an alternative, we perform a query-based evaluation using a database of medical papers. This evaluation is shown in §9. Results indicate that the potential genetic markers found by our proposed methodology are related to some already known genetic markers.

It is thus possible that the markers not found in the literature through our queries could be not known by the current state of the art. These could be analysed experimentally to find genetic markers at a gene level in a fastest and cheaper way (as the searching space would be several times smaller than the searching on the whole genome). Finally, all the results obtained in this thesis are explained in §10 and we expose our final conclusions in §11.

2. Document Overview

In this section, we will do a summary of the document contents. First, in §3 and §4 we present the basic concepts of machine learning and biology needed to understand the domain of the thesis. §3 includes the definitions of supervised learning, feature selection, decision tree and random forest. §4 contains the definitions of chromosomal rearrangements, cancer (with its different classification criteria) and the germinal layers.

In §5 we present the dataset used in the thesis and a detailed analysis of all the patient-related features it contains. These features are the cancer type, the germinal layer, the gender and age of the patient, and the tumour stages in two different granularities.

Cancer type and germinal layer are the candidates to target variable. In the subsection dedicated to these variables, is explained which of them is chosen as target and why. Gender and age of the patient are variables with a strong relationship with the cancer type. In their subsection, we explain some examples of their relationship with cancer epidemiology and their distribution in the data with respect to the germ layers. The tumour stages correspond to how much has spread the tumour within the body, some cancer types are more prone to metastasise than others, so this variable is also important to characterise the cancer samples. In their subsection, we show some examples of their relationship with the cancer types and their distribution within our data.

After studying the data, in §6 we show how we pre-process the dataset to be exploitable with a machine learning algorithm. Among other problems, the original dataset contained very sparse data and missing values. In this section, we solve these problems and engineer several new features related to the genetic information of the samples.

We present genetic marker finding as a feature selection problem with respect to the features extracted in the pre-processing section. We also perform a classification with the selected features to validate our results. For performing the classification we had different options, in §7 we present the model we chose, a Random Forest, and explain why we decided to use it instead of any of the other possible options, like Decision Trees, Neural Networks or SVM.

In §8 we present the experiments performed to find the genetic markers and the classification results used to validate them. This section is separated in: Hyperparameter tuning (§8.1), Feature ranking generation (§8.2), Germ layer specific ranking generation (§8.3) and classification results(§8.4). In Hyperparameter tuning, we present the methodology used to select the hyperparameters and ones used in the rest of the experiments. In feature ranking generation, we present the feature ranking obtained by aggregating a 500 independent trainings of a Random Forest. In Germ layer specific ranking generation, we perform a feature ranking specific for every germ layers following the same methodology than in feature ranking generation. Finally, we present the classification results obtained by using different sets of features; this results show that the classifier can discriminate the samples by using the selected features.

Given some difficulties for validation our results, we performed a query-based evaluation over the medical literature. This validation is presented in §9 and a discussion of its results is shown in §10. Finally, in §11 we present the conclusions extracted of the thesis.

3. Machine Learning Background

This section introduces the Machine Learning nomenclature used in the remaining of the document.

Supervised Learning

The first step when dealing with a machine learning task is to define the typology of problem to be solved. A task can be supervised, unsupervised or semi-supervised depending on the data available and the kind of relationships or patterns you want to extract from it. We say a task is supervised when you use an algorithm to build a model from a set of labelled training data. This model maps a set of inputs with the desired outputs in the training stage and is able to generalise and infer new outputs from new data. One example of supervised learning would be a spam classifier trained with labelled emails (spam or not spam). In this thesis, we have a supervised learning task. The samples of our task correspond to the engineered and metadata features and the labels correspond to the germ layers of each sample.

Feature selection

Feature selection is the process of selecting relevant features for model construction. The objectives of feature selection are usually to avoid the curse of dimensionality, shorter training times, improve the interpretability of the models or reduce overfitting. In this thesis, we use it as a knowledge discovery tool, as in our case, finding the most relevant features (*i.e.*, the genetic markers) is the main objective. In this thesis we use a Random Forest as a feature selection method.

Decision Tree and Random Forest

A Decision Tree is a model based on inferring simple rules from the features of the training dataset. Decision trees operate by separating the samples through these rules in a tree-like structure. The rules generated by the decision tree are the ones that maximise the difference of the samples within a node with respect to one of the features. The feature is selected tipically by using either the information gain, the gini impurity or the entropy measures [6]. Figure 1 shows a simple example of a decision tree we trained for illustrative purposes with the iris dataset (four features and three classes). As we can see in the figure, each node contains a rule, an entropy value, the number of samples of the node and the distribution of the samples within the three classes. The root of the tree is the first decision rule; it separates the samples depending if their second feature is greater or smaller than $2,45$. In the first node the samples are distributed equally among the three classes.

Some of the advantages of this model its capacity to train using very few samples. Does feature selection during the training and is one of the most interpretable models. The main problem with this model is that it is unstable (a small change in the data can lead to a substantial change in the final tree structure) and prone to do overfitting in the training data.

One way of avoiding or at least mitigating these problems is to use an ensemble of decision trees. A Random Forest Classifier is a voting ensemble of decision trees. To avoid overfitting, the trees used in a random forest must significantly differ from one another. One way of obtaining different trees is to train them with different subsets of the dataset both in terms of features and samples. Then the classification is performed as a voting of the individual classification results of all trees.

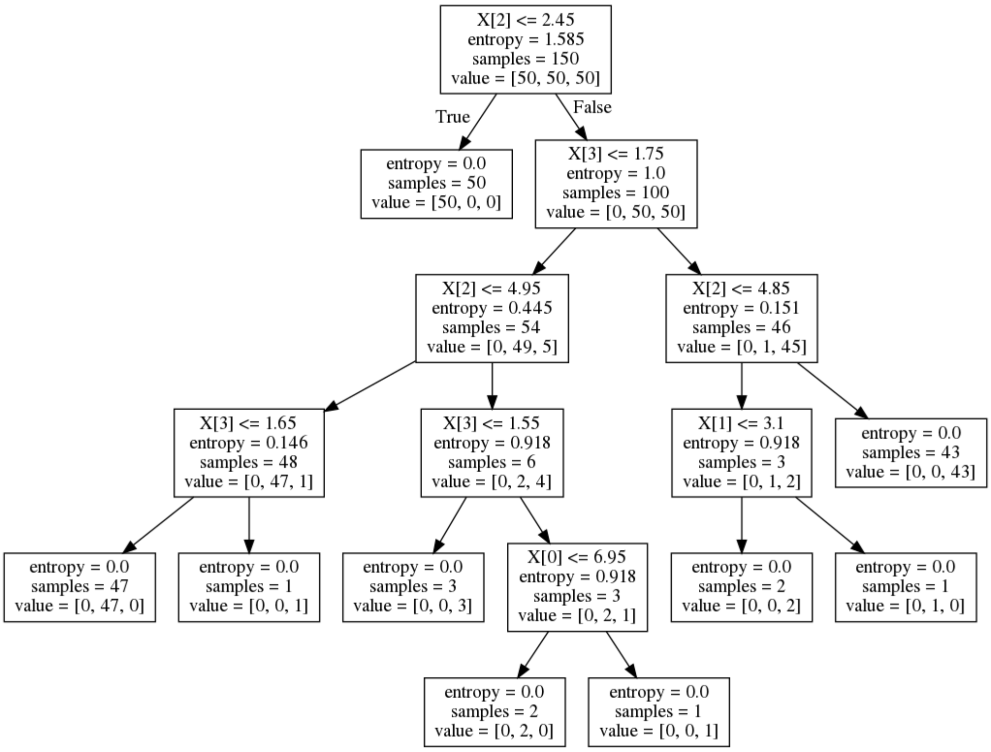


Figure 1: Example of a decision tree trained with the iris dataset. This model classifies into iris setosa, iris virginica and iris versicolor.

4. Biological Background

This thesis is strongly related to genetics. To understand its motivation, methodology and relevance properly, some genetic and biological background is needed. In this section, we explain three of the basic concepts that the reader needs to understand the remaining of the document.

Chromosomal rearrangements

The dataset that we use in this thesis contains a list of chromosomal rearrangements of several cancer patients. To understand the dataset, and the problem we are dealing with in this thesis, it is essential to understand what a chromosomal rearrangement is, and how it affects the generation of tumours. Let us start with the first.

A chromosomal rearrangement is a chromosomal disorder that involves a change in the structure or order of a chromosome. These rearrangements can appear, for example, as an error when a cell splits. There are four different types of abnormalities. Next, we give a brief description of those. For a simplified illustration of their effect, see Figure 2.

- Deletion: A section of the chromosome is erased. Deletion of a chromosomic part containing tumour suppressor genes might lead the cell to generate a tumour.
- Duplication: A section of the chromosome is duplicated. If the duplicated section contains an oncogene (*i.e.*, gene that has the potential to cause cancer), this rearrangement can generate a tumour.
- Inversion: A section of the chromosome is split and reattached to the chromosome with inverted order. There are two kinds of chromosomal inversions, paracentric and pericentric, which are represented as *t2tinv* and *h2hinv* in the data.
- Translocation: Two sections of two different chromosomes split and reattach into the incorrect chromosome.

There are two main ways in which chromosomal rearrangement can cause cancer. One is by the disappearance of a cancer suppressor. The other one is with the appearance of an oncogene. Deletion and Duplication are directly related with both scenario (deletion on cancer suppressors, duplication of oncogenes). Inversion and Translocation on the other hand, typically have an indirect effect. They may break a cancer suppressor, or they can create an oncogene, depending on the final ordering of the bases after the chromosomal abnormality has taken place.

Cancer and Cancer Types

In this thesis, we work with cancer samples from a genetic point of view. We call Cancer to the set of diseases related to the abnormal growth of the cells. These diseases can spread to other parts of the body, in a process known as metastasis. Several studies relate the appearance of cancer with chromosomal rearrangements in the tumour cells, both as a cause that generated the tumour or as a consequence of the abnormal splitting of the tumour cells [7, 8, 9].

There are different ways of classifying a tumour based on its characteristics. These classifications can be, for example the cell where the tumour was generated (*e.g.*, carcinoma, sarcoma, lymphoma *etc.*

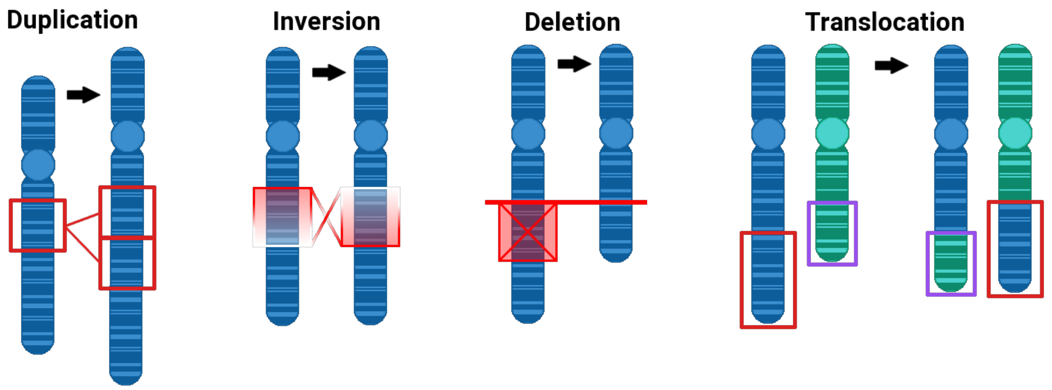


Figure 2: Visual representation of chromosomal rearrangements.

), the tumour stage (in simple terms, the medical measure of how much has advanced and spread the tumour), or just the body tissue where it was generated (*e.g.*, breast cancer, lung cancer *etc.*), in further references, we will refer to the latter classification as *tumour type* (or *cancer type*). In practice, the classification criteria used depends on the context. For example, when performing a diagnosis over a patient it might be useful for the oncologist to use the cancer type, but for the pathologist, it might be better to use tumour stage to have in mind how much the tumour has spread.

This work is strongly based on the study of the genetic differences of the samples for different kinds of tumours. The more fine-graned classification criteria that we could use as provided by the original dataset would be the tumour type. Unfortunately, the small number of samples (2,586), respect to the huge feature dimensionality (3 billion base-pairs if we used the data as given) classified into so many classes (21 cancer types in our dataset) would make that an intractable problem. To handle this issue, one of the measures that we performed was to use the germ layers as a generalisation of the cancer types. A cell is derived from only one germinal layer of the four possible; So, this generalisation is biologically coherent. Another reason to use this generalisation instead of other possibilities is that the cells that come from a germinal layer, in genetic terms, should have more in common between them than with the cells that are from another germ layer, in the same way, that the cells of an organ are more similar between them than with the cells of another organ.

Germinal layers

On the previous section, we introduced our decision of using the germinal layers as a target variable because of dataset size limitations. The germinal layers (or germ layers) are the first differentiation of cells that forms during embryonic development. When the zygote starts its division, cells start to differentiate. The three layers with the newly differentiated cells are known as the germinal layers; Endoderm, Ectoderm and Mesoderm. The subsequent differentiation within the Ectoderm layer turns into the Neural Crest layer. This layer, although it is derived from Ectoderm cells, is sometimes considered a fourth germinal layer because of its relevance [10]. This will be our case.

Further on the fetus development, the cells of a germ layer get more specialised and end up becoming the parts of the body. For example, the brain is generated by Ectoderm cells, and the skin is generated

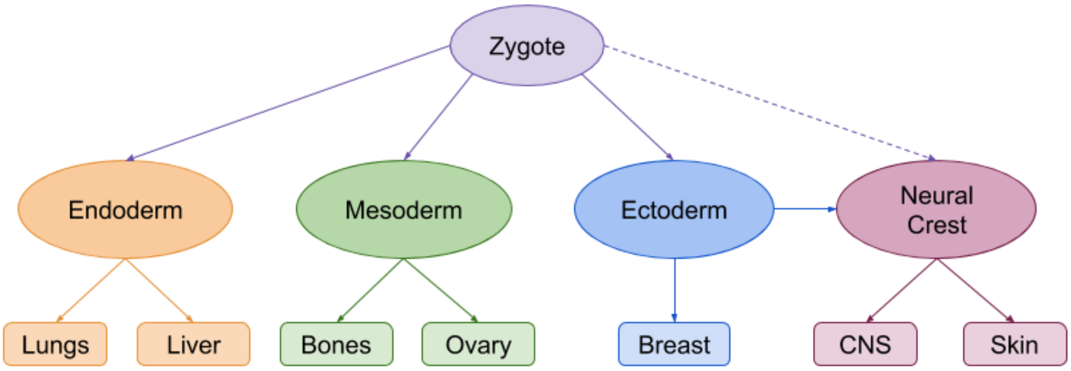


Figure 3: A visual representation of the four germinal layers with some examples of the organs developed from them.

by Neural Crest cells, the bones are generated by Mesoderm cells, and the stomach is generated by Endoderm cells.

The germinal layers are the first differentiation that appears on the cell development, for this reason, we expect this classification of the cells to be related in a stronger way with the cell genetic traits than other possible classification criteria.

5. The Dataset

There are several projects with the goal of saving and documenting cancer-related data for research purposes, and one of the more recent and ambitious of these projects is the Pan-Cancer Analysis of Whole Genomes project (PCAWG) [11, 12]. The PCAWG contains the genome sequencing from over 2,800 tumours from the 33 most frequent types of cancer, which makes it an valuable tool for researchers around all the world studying cancer.

The genome sequencing of cancer can be used for several research purposes, for example, reclassifying tumour types based on molecular similarity [13, 14] or studying the relationship between somatic mutations and cancer appearance and progression [15]. This thesis will be focused on the second type of study.

Working with the human genome, which has 3,088,286,401 base pairs is a challenging task. There are around 115 breaks per patient, on average in our dataset. Considering the huge number of bases makes our data extremely sparse. In order to work with the full genome sequencing, we would need a huge number of samples. A number fitting the feature space size. If we were to have as many samples as features, we would need 3,088,286,401 samples (which would be almost half the world population). Clearly the domain space must be simplified. To try to reduce the feature space, we use a dataset containing only chromosomal breaks instead of the whole genome sequencing. This dataset was generated from the Pan-Cancer genome sequencing data, by applying an *in-house* pipeline from the Life Sciences department at the Barcelona Supercomputing Center (BSC). This pipeline uses the Burrows-Wheeler Aligner (BWA) [16] to identify and extract breakpoints which point to genomic and chromosomal rearrangements.

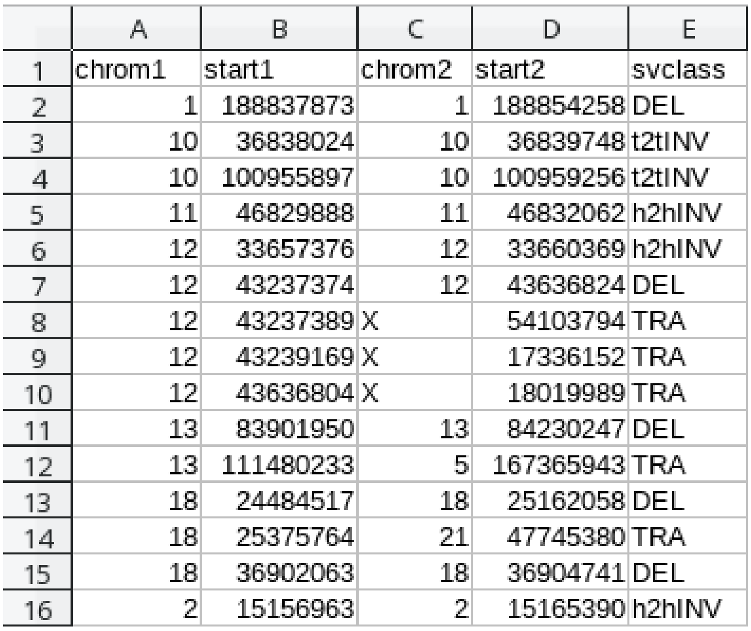
Our starting dataset is the output of the commented pipeline. The dataset is composed by a set of 2,586 files, each one containing one or more chromosomal rearrangements found in a tumour cell sample. Each file corresponds to a different patient. A chromosomal rearrangement is represented by its type (*svclass*), the starting base-pair (start) and chromosome (*chrom*1), and the ending base-pair (*start*2) and chromosome (*chrom*2). We can see an example of these files in Figure 4. Additionally, there is patient metadata available for each file. This includes sex, age, tumour type and corresponding germ layer. Table 1 shows a brief description of the patient metadata.

Variable	Type	Description
donor_sex	Categorical	The sex of the patient (male or female)
donor_age_at_diagnosis	Numerical	The patient age.
histology_tier1	Categorical	The germinal layer of the tumour tissue.
histology_tier2	Categorical	The type of the tumour.
tumour_stage	Categorical	The stage of the tumor (Primary, Metastatic etc.)

Table 1: Variables contained in the metadata.

5.1 Data Analysis

Our first step to begin this work was to perform an exhaustive analysis of the data, to understand it properly and define a set of objectives and a hypothesis. On this analysis, variables were analysed independently and also in the context of cancer, concerning the cancer types and germinal layers. This



	A	B	C	D	E
1	chrom1	start1	chrom2	start2	svclass
2	1	188837873	1	188854258	DEL
3	10	36838024	10	36839748	t2tINV
4	10	100955897	10	100959256	t2tINV
5	11	46829888	11	46832062	h2hINV
6	12	33657376	12	33660369	h2hINV
7	12	43237374	12	43636824	DEL
8	12	43237389	X	54103794	TRA
9	12	43239169	X	17336152	TRA
10	12	43636804	X	18019989	TRA
11	13	83901950	13	84230247	DEL
12	13	111480233	5	167365943	TRA
13	18	24484517	18	25162058	DEL
14	18	25375764	21	47745380	TRA
15	18	36902063	18	36904741	DEL
16	2	15156963	2	15165390	h2hINV

Figure 4: Example of chromosomal rearrangements of a patient.

way, we obtained a general characterisation of the dataset. The dataset characterisation will be explained in this section.

Usually, when you study a dataset, one of the ways to validate it is to compare its features with known statistics and facts about the data it contains. A sample of humans is expected to have age values within a given range, 0-100, for example. A value of 300 should be noticed, as well as a saturation of values on a given age range. If most samples were within a small range, for example, 15 to 20, this would imply a bias, which would be a warning about the representativity of the data. This dataset includes genetic samples from different cancer types. Our dataset has some limitations in this aspect. Some of the ones that we have found when studying it are:

- Working with a dataset containing more than one type of cancer is uncommon. Usually, cancer is studied from a clinical point of view, which implies that most of the cancer-related papers are focused on only one specific type of cancer. This issue is explained by the fact that different cancer types behave differently. Furthermore, different cancer types are differently treated. This fact makes the process of statistic extraction hard and painstaking since we would need to independently gather and aggregate sets of data for dozens of cancer types.
- We do not know the origin country of the patients. The PCAWG is an international project, which implies that the samples can be from patients from all the globe. Environmental and experimental factors are very relevant to cancer. Two of the main ones being the eating habits and the contamination level. These factors are strongly related to the country of residence of the patient [1], so not having a feature with the country of residence of the patient might be a limitation of the dataset. As it makes unfeasible to validate our data concerning the known cancer epidemiology.

Regardless of these difficulties, we studied the cancer literature, and we even use it to validate our results, through paper query matching. Also, there are some cancer-type specific facts that can be expected to be satisfied within this dataset, like for example, having zero or very few breast cancer samples corresponding to male patients.

Germinal Layer and Cancer Type

In our data, we have two variables which correspond to cancer types at different levels. These are Histology_tier1 and histology_tier2 which specify the germinal layer and cancer type of the sample. Figure 5 and Figure 6 show the distribution of samples within these two variables. Notice that samples are strongly unbalanced for both features.

To illustrate the relationship between both features, Table 2 shows the distribution of data samples between both variables. We can observe that the different germinal layer contain a different number of cancer types. For example, Ectoderm only contains one cancer type, breast cancer, while Endoderm contains ten different cancer types. This difference is significant enough because the genetic variability between samples of several cancer types is expected to be higher than the one between samples of the same cancer type.

Since our goal is to find a categorisation of cancer types based on chromosomal breaks, these two variables are the candidates to be used as a target. If we were seeking more specific results, the best option would be using the cancer type. This way, we could obtain genetic markers related to specific cancer types, which are closer to the kind of genetic markers used in clinical research (for example, for making cancer gene therapies). Using cancer type as target variable, however, was unfeasible, for the following reasons:

- There are too many cancer types with respect to the number of samples (21 cancer types respect to 2,586 samples). Also, as we can see on Figure 6 the dataset is very unbalanced with respect to the cancer type, having around 25 samples of the smaller class (cervix cancer) and more than 300 samples of the bigger one (liver cancer).
- The huge dimensionality of the feature space, which in our case corresponds with the 3 billion base-pairs of the human genome, forces us to simplify the data. One of the simplifications that we decided to use is selecting the germ layer as a target variable instead of the cancer type.

Having more samples would relax the necessity of simplifications, which would allow using the cancer type as the target variable.

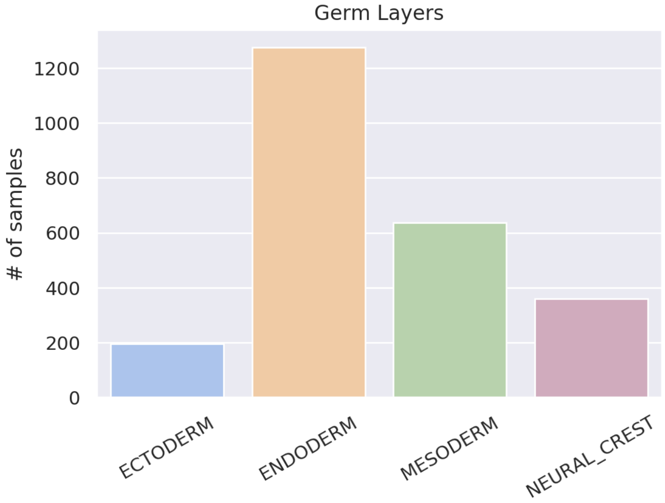


Figure 5: Number of samples in the data for every cancer type.

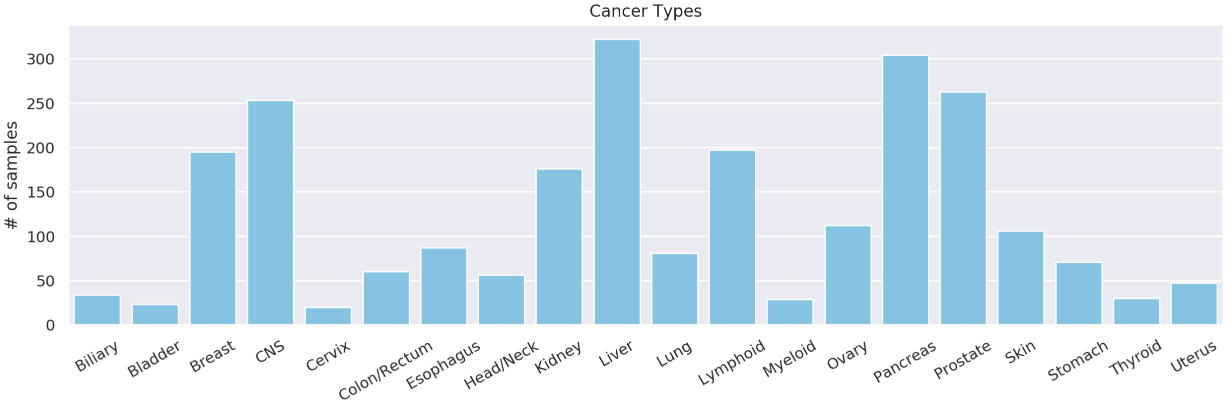


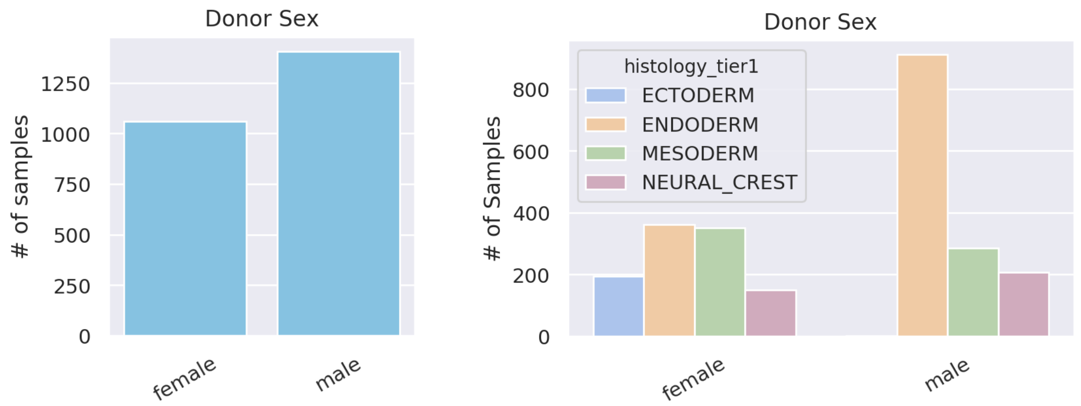
Figure 6: Number of samples in the data for every cancer type.

	Germ Layers			
Cancer Types	Ectoderm	Endoderm	Mesoderm	Neural Crest
Biliary	0	34	0	0
Bladder	0	23	0	0
Bone/SoftTissue	0	0	92	0
Breast	209	0	0	0
CNS	0	0	0	261
Cervix	0	0	20	0
Colon/Rectum	0	60	0	0
Esophagus	0	87	0	0
Head/Neck	0	0	56	0
Kidney	0	0	176	0
Liver	0	322	0	0
Lung	0	84	0	0
Lymphoid	0	0	197	0
Myeloid	0	0	29	0
Ovary	0	0	112	0
Pancreas	0	306	0	0
Prostate	0	263	0	0
Skin	0	0	0	106
Stomach	0	72	0	0
Thyroid	0	30	0	0
Uterus	0	0	47	0
Total	209	1,281	729	367

Table 2: Number of samples of each cancer type with respect to its germinal layer. This table shows *how* unbalanced is the dataset respect the two variables and *how* are samples distributed among them.

Gender

The patient's sex is known to be strongly related to cancer type [17]. In one hand, cancer incidence is a bit higher in males. Coherently, Figure 7a shows that our dataset contains more samples of male patients than female patients (57% vs 43%) . On the other hand, some cancer types can only appear in females (*e.g.*, uterus and ovary cancer) or males (*e.g.*, prostate cancer). Beyond the particular types, there are types of cancer that are more common in one gender than in the other, even though it is biologically possible for it to appear on both genders (*e.g.*, breast cancer is more common in females). As Figure 7b shows, in our specific data, there are no male patients with cancer from a tissue generated by cells of the Ectoderm germinal layer. This germ layer contains only breast cancer (see Table 2), which is very uncommon in males. On the other hand, we can see that between male patients, the most common cancer types are the ones from Endoderm germ layers, which includes prostate cancer.



(a) Distribution of samples among genders.

(b) Distribution of samples among genders and germ layers.

Figure 7: Distribution of samples with respect to the gender variable.

Age

The patient's age is a critical variable when talking about cancer, as it can affect tumour incidence, the tumour behaviour once it has appeared and even the treatment that will be used to cure the patient. The next examples illustrate these effects:

- Two examples of how cancer incidence is affected by patient age are [18, 19]. In these papers authors show that leukaemia and Central Nervous System (CNS) are the most common cancer types among children. While other cancer types have more prevalence within older adults.
- An example of how cancer behaviour is affected by patient age is breast cancer. In [20] the authors show that breast cancer has more risk of metastasis and cancer recurrence in young patients (less than 35 years old) than in older ones.
- Foster *et al* study how the age of the patient affects oncologist choice of the treatment [21]. In this paper, authors show that intensive cancer therapy was significantly less recommended for older patients (around 80 years) than for younger ones (around 65 years).

There are different ways of defining age groups in medicine. In this thesis we decided to use the clustering-based one presented in [22]. This categorisation has biological meaning, and does not have overlapping classes. The age group definition was extracted by applying different clustering algorithms on a database of literature-derived entries describing age and phenotype. The age groups are defined as: Infant (birth-2), preschool child (3-5), child (6-13), adolescent (14-18), young adult (19-33), adult (34-48), middle aged (49-64), aged (65-78) and "79 and over" ($\geq$79).

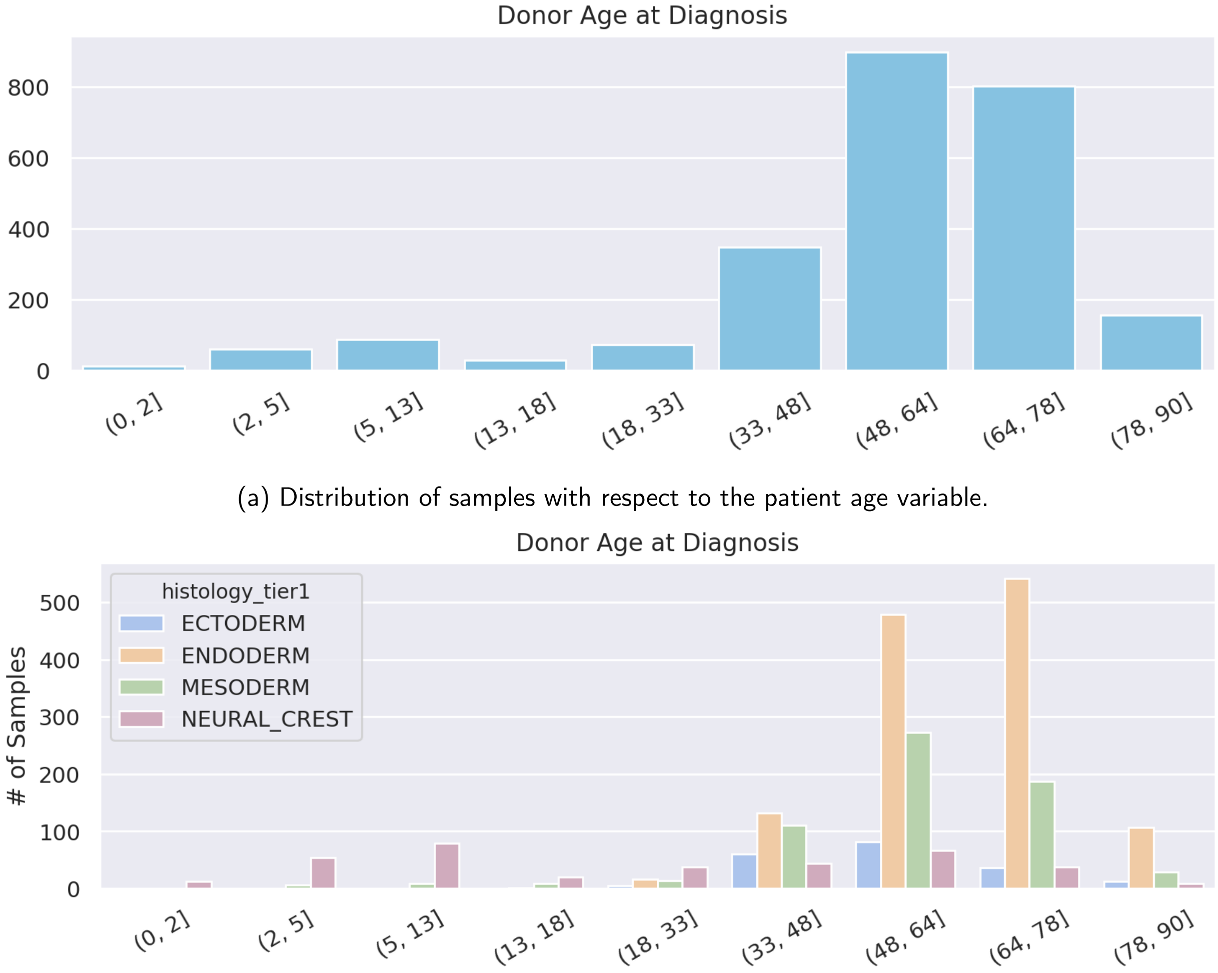
Using as a reference the commented age groups, we can see in Figure 8a and Figure 8b that all the age groups are minimally represented on our data. First, we can see some samples of the infant and pre-school child groups, these patients mostly have Neural Crest cancers, this is coherent with cancer epidemiology, as discussed above. Within the young adult group we have very few samples (72), which is to be expected since the young adult group usually has the smallest cancer incidence. We can also notice that within this age group most of the samples are from the Neural Crest germ layer (37), some samples belong to Endoderm and Mesoderm (16 and 14 respectively) and very few of Ectoderm (5).

Most of our patients are from the adult, middle aged and aged groups. Within the patients of these three groups we can see a greater prevalence of Endoderm cancer types, followed by Mesoderm. Notice that the difference between the number of samples of Endoderm and Mesoderm increases within the older age groups, having almost the same number of samples on each of the adult groups. It is important to take into account that Ectoderm and Mesoderm are the two germ layers with the largest number of samples. However, the cancer types within these two germ layers are also the ones with the highest dependency of external risk factors, like lung [23], liver [24] or kidney [25] cancer. Age has also a strong relationship with hormone levels. Some of the most known examples of these hormonal changes are adolescence or menopause. Ectoderm, Mesoderm and Endoderm germ layers contain some hormone-related cancer types in these age groups, like breast [26], ovary [27] and uterus [28]. These three cancer types have a strong relationship with the hormone changes related to menopause, which usually happens between the 45 and 55 years.

The last age group to comment is 79 and over. We do not have much samples from this age group. Treating cancer usually implies invasive surgeries or strong and dangerous treatments, this kind of therapies are a lot more dangerous in an old patient, so depending on the case the doctor might decide not doing them. These differences on the treatment imply an extra difficulty on obtaining cancer cell samples, so it is possible that this age group might be under-represented. Regarding the distribution of germ layers on this age group, we can see a similar behaviour to the aged group, but with a greater difference within the number of samples of Endodoerm and Mesoderm.

Of all features considered for this study, age is the only one that contains missing values. From the 2,586 valid samples (the ones that have metadata), 119 do not have an age value, representing a 5% of our data. The distribution of these missing values with respect to the germ layers are: 92 in Mesoderm, 14 in Ectoderm, 8 in Neural Crest and 5 in Endoderm. Significantly, within Mesoderm the missing values account for a 12% of all samples in the class.

Even though 5% is a small percentage of missing values, this is something important to take into account considering the relevance of age for cancer characterisation. To fill these missing values, we used the Multivariate Imputation by Chained Equations (MICE) [29]. MICE uses the values of similar samples to the ones with missing values to predict the missing values. As a result, the quality of these new values will highly depend on the variety of non-missing samples. In our case, this is a inconvenient on our methodology, since most of the missing values correspond to the Mesoderm germ layer.



(a) Distribution of samples with respect to the patient age variable.

(b) Distribution of samples with respect to the patient age variable by germ layer.

Figure 8: Distribution of samples with respect to the patient age variable, and labelled by germ layer.

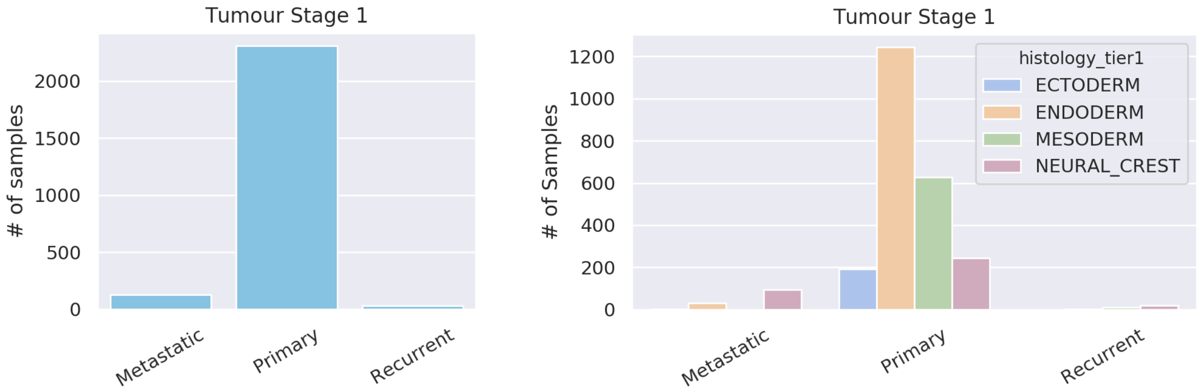
Tumour Stage

The tumour stage represents how much has cancer spread within the body. We have two different features with tumour stage information. The first one, *tumour_stage1*, contains the most general information, which tells us if this tumour is Primary (*i.e.*, has not spread to other parts of the body), Metastasic (*i.e.*, has spread to other parts of the body) or Recurrent (*i.e.*, the tumour was not completely removed from the patient and re-appeared later). Figure 9b shows the distribution of the samples within these classes with respect to the germinal layers. On the figure, we can see clearly that most of the samples (99%) came from primary tumours. There are some known facts regarding the organ that contains the tumour characteristics that we can relate to our dataset. For example:

- Breast cancer (Ectoderm) is usually found in the early stages, which makes it less probable to metastasise or to re-appear once it is removed.
- Some of CNS cancers (like Medulloblastoma) and skin cancers (like Melanoma), both from Neural crest category, are known to be very likely to metastasise.

The second categorisation of tumour stage, *tumour_stage2*, has more specific information. Its classes represent the type of tissue where the tumour has spread. This kind of knowledge is used by doctors to measure how dangerous is a specific tumour and decide the *best* medical procedure to remove it.

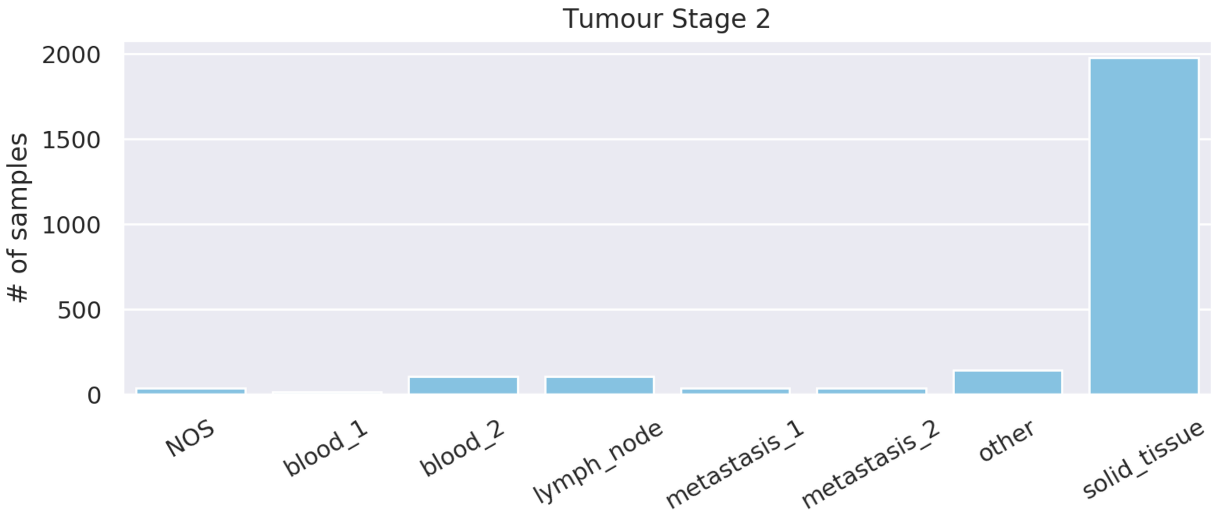
Figure 10b shows the distribution of the samples within the categories of this feature with respect to the germ layers. Given the specificity of this data, it is difficult to interpret this plot without extensive medical knowledge, but we can see that some of the classes are strongly specific of some of the germ layers. For example Mesoderm tumours seem to have a tendency to spread to lymph nodes, which is coherent with the fact that *some* of the cancer types from Mesoderm germ layers have lymph nodes nearby (like the neck, the ovaries or the uterus). Also, Mesoderm germ layer contains Myeloid cancer, which is a kind of leukaemia (blood cancer). This is also coherent with spreading into the blood.



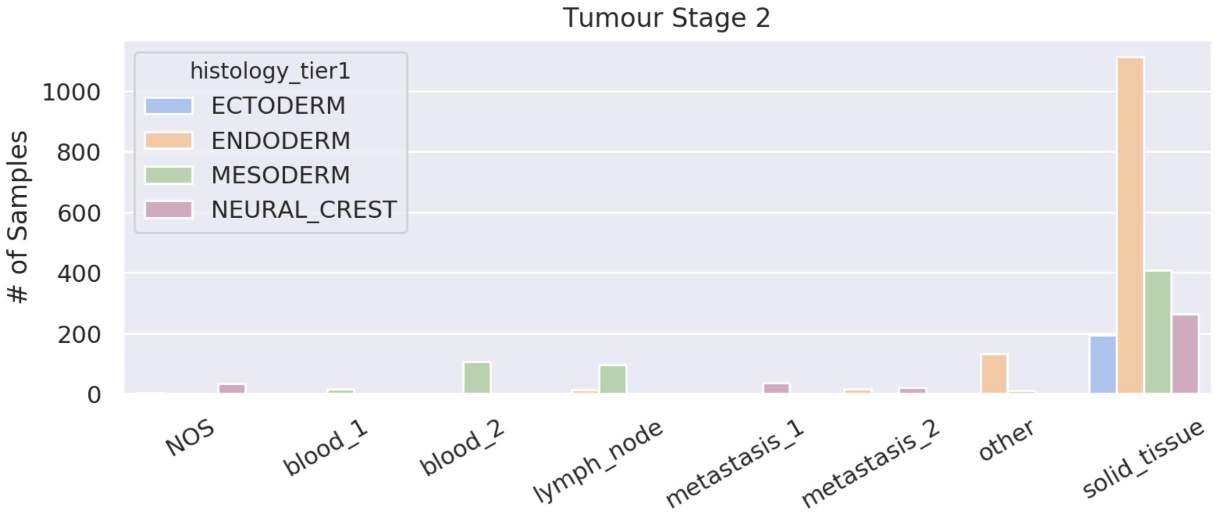
(a) Distribution of samples with respect to its tumour stage.

(b) Distribution of samples with respect to the tumour stage, separated by germinal layer.

Figure 9: Distribution of samples with respect to the Tumour Stage 1 feature.



(a) Distribution of samples with respect to the Tumour Stage 2 variable.



(b) Distribution of samples with respect to the Tumour Stage 2 variable, labelled by germ layer.

Figure 10: The values of the categories correspond to: Not Otherwise Specified (NOS), metastasis local to lymph node (metastasis 1), metastasis to a distant location (metastasis 2), blood-derived in bone marrow (blood 1) and blood-derived in peripheral blood (blood 2) .

6. Pre-processing

In this section, we explain how we transform the original data, to create a dataset exploitable by machine learning algorithms. The results shown in our data analysis suggests that the dataset ought to be pre-processed before it can be fed to a machine-learning algorithm. Among the particularities that should be dealt with, we consider the following:

- Genetic abnormalities are too sparse for a model to be able to extrapolate relationships from them. The human genome contains around 3 billion DNA base pairs. The breakages from the data were given from one specific base-pair to another. The mean number of breaks per patient was around 115, which makes this an *extremely* sparse dataset.
- Some samples were missing their metadata. For these samples, we did not know the cancer type nor the germ layers.
- The dataset contained categorical variables. In Table 1, we can see that most of the metadata variables are categorical. Some models can handle categorical data without the need for processing it in any way, but Random Forest is not one of them, so we needed to transform these variables into numerical ones.
- The dataset contained missing values on the Age variable. As shown in Section 5.1 the age has a strong relationship with tumour-genesis, and with the cancer type of the patient. For example, most of the cancer types are very uncommon in young people. In our dataset 119 of the age values where missing, respect to the number of samples, it represents a 5% of the values of the age.

The dataset contained 196 samples without metadata values. This means that these samples did not have either cancer type or germ layer label, together with the patient information (*i.e.*, age, gender and tumour stages). Since we cannot use samples without class label, we discard these 196 instances. As commented in §4, we selected as a target variable the germ layer not using the cancer type features in our experiments.

After removing these unlabelled samples, we focus on preparing the features. From the metadata we will extract a list of metadata features, such as the age of the patient, its sex and the two kinds of tumour stages (see the bottom part of Table 3). In Section 5 we analysed all these variables and saw that they could be important for classifying the samples.

From the genetic data we engineer a second set of features. These are the most interesting ones for the context of this thesis. Genetic features are related to the kind of breaks or the chromosome that contains them. Features include, among others, the number of rearrangements on every chromosome or the number of rearrangements of each type. We decided to use this kind of features because it might be useful in a medical level for finding genetic markers or genetic patterns related to cancer. Also, they represent an abstraction of the data which makes them more usable in a sparse domain The final set of features (engineered and metadata features) as shown in Table 3 accompanied by a brief description.

For the metadata features, we perform a one-hot-encoding over the categorical features (both tumour stages). Furthermore, we impute 119 missing values for the age feature using MICE. Data was split into two partitions. One for training the model and another one for testing the classification results. The partition was stratified *w.r.t.* the germ layers, in order to try to maintain their original distribution (Table 2). This way, we obtained a training partition with 2,068 samples and a test partition with 519 samples.

Genetic Data		**Num. Features**
#_of_breaks	No. of breaks of the sample.	1
DUP, DEL, TRA	No. of breaks per break type.	3
h2hINV, t2tINV	No. of breaks per inversion type.	2
chr_1, ..., chr_Y	No. of breaks per chromosome.	24
DEL_1, ..., DEL_Y	No. of deletions per chromosome.	24
DUP_1, ..., DUP_Y	No. of duplications per chromosome.	24
TRA_1, ..., TRA_Y	No. of translocations per chromosome.	24
h2hINV_1, ..., h2hINV_Y	No. of h2h inversions per chromosome.	24
t2tINV_1, ..., t2tINV_Y	No. of t2t inversions per chromosome.	24
prop_{chr_n, ..., t2tINV_n}	For each break type, for each chromosome, proportion of breaks over total breaks in patient.	149
Patient Metadata		
female	Gender of the patient, 1 if female and 0 otherwise.	1
donor_age	Age of the patient.	1
ts_1_category	Metastatic, Primary or Recurrent	3
ts_2_category	NOS, bone marrow, periphleal blood, derived from tumour, metastasis to lymph node, metastasis to distant location, other or solid tissue	9
Total number of features		313

Table 3: Genetic features (top) and metadata features (bottom) extracted from the dataset.

Both partitions have 313 features, including boolean features from the one-hot-encoding. Most of these features are genetic-related information extracted from the chromosomal rearrangements, among them we have the total number of breaks of the sample, the number of breaks of the different rearrangement types (Deletions, Duplications *etc.*), the number of rearrangements of each type by chromosome. The percentage of patient breaks on each chromosome was also added.

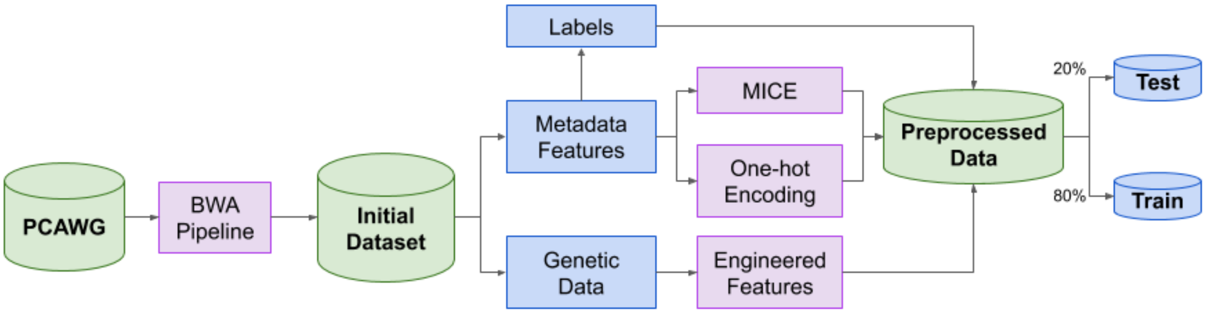


Figure 11: Diagram showing the data-flow from the PCAWG to our final dataset.

7. Methodology

Finding genetic markers from our chromosomal rearrangement dataset corresponds to a Supervised Feature Selection problem (in Machine Learning terms); we are trying to find what are the most relevant features to characterise the data with respect to the target variable, *i.e.*, the germinal layers. In order to validate our results, we would need to check that the extracted features are correct. However, there is no repository of genetic markers to check against. Furthermore, the genetic markers we are using (the ones defined by our features) are not common in the medical domain, for the reasons explained in §5. As we will further develop in §9, some of the markers found by our algorithm would be expected to be on the medical cancer literature. We could therefore use such literature to evaluate our results. Another approach that we use to evaluate our results is to classify the samples using the selected features. This way, we can show that at least the features extracted are characteristic enough for a model to be able to discriminate between the germinal layers. If a model can classify with the given information, this means that the data we are feeding it has meaningful relationships with respect to the target classes.

To generate our methodology, we would need or a feature selection model with a classifier or a classifier that performs feature selection during the training. However, as we are dealing with a very specific dataset for solving a particular problem, we need a methodology that fulfils the next characteristics:

- The method needs to be explainable. We are dealing with a medical problem; doctors can only use a model if there is any way of explaining why the model has decided something so that it would be useless a black-box model in a medical context.
- The method needs to be scalable. The number of genetic cancer samples is expected to increase with time, as it is an open research field. Having more samples would allow using a target variable with smaller granularity (*i.e.*, the cancer types instead of the germ layer), which would give genetic markers for specific cancer types. If the methodology used is not scalable, it could not adapt to this increment of the number of samples.
- The method needs to be able to generalise well with respect to the classification results. If the classification results change between trainings, it means that the model has found new/different relationships between the input and the target variables. If there are too many differences among trainings, this would imply a lack of consistency and reliability, making our results hardly usable.
- The method needs to be able to work with a limited number of samples. We have only 2,586 in our dataset, and we need to select a methodology capable of extracting relationships among those few samples. with a limited number of samples. We have only 2,586 in our dataset, and we need to select a methodology capable of extracting relationships among those few samples.

Starting for the classifier, there are three primary families of models from which we can choose our classifier; Neural Networks, Support Vector Machines (or Kernel-based models) and Decision Trees. Table 4 shows a comparison of these model families with respect to the characteristics we need in the model. As we can see, there is no a perfect option, although the one closest would be the Decision Tree classifier, as it is scalable, explainable and it can be trained with a small dataset. The only problem with this model is that it is prone to overfitting. Fortunately, this can be solved by using an ensemble of Decision Trees, like a Random Forest.

A Decision Tree Classifier [30] is based on inferring simple decision rules from the training data, it produces linear separations on the data using optimised thresholds on individual features. There are many

	Neural Networks	SVM	Decision Trees
Explainable	No	No	Yes
Scalable	Yes	No	Yes
Able to generalise	Yes	Yes	No
Small dataset	No	Yes	Yes

Table 4: Comparison of the most used families of supervised classification models.

Decision Tree algorithms, within them, we decided to use (Classification and Regression Trees) CART [31] following the study of [32]. In this paper authors compare the different Decision Tree algorithms performance in the medical domain and conclude that CART is the best option while classifying medical data.

A Random Forest Classifier [33] is a voting ensemble of uncorrelated Decision Trees. To ensure that the Decision Trees of the forest are uncorrelated, each Tree is trained with a random sample of the data with replacement. Additionally, each of the trees can only use a random subset of the features. The predictions of a Random Forest are made by majority voting over the classification results of the Decision Trees, overcoming this way the overfitting tendencies of the Decision Trees. In general, a Random Forest has the advantages of a Decision Tree, but are a more robust model.

8. Random Forest as Feature Extractor

On this study, a Random Forest model implementation [34] was used to identify which chromosomal rearrangements, and in which location, are correlated with the germ layers of the corresponding tumour. The correlated features will then serve as genetic markers of that particular germ layer. Random Forest has shown to have good performance over many applications, is one of the more interpretable models among the current machine learning state of the art, and it is capable of providing feature importance after training it. These properties make it a *good* candidate for the computational biology field, both as a classification or feature selection tool [35, 36].

We trained the Random Forest using the set of 313 features extracted from the chromosomal rearrangement data and the patient metadata (see Table 3) targeting the prediction of the corresponding germ layer. We focus on two aspects of the trained model, the feature selection and the methodology evaluation.

The feature importance ranking shows which features did the model take into account for classifying the samples and to what degree. The first features in this ranking are the most relevant ones, which means that they contain any type of differentiation with respect to the target classes. For example, if we where classifying cats and lions, the size might be one of the first features in the ranking while having hair would be one of the bottom ones, as both categories have abundant hair. This feature ranking is what we use to generate our aggregated ranking and contains the potential genetic markers.

The classification results show that effectively, features that the model considers that are important, are relevant for classifying the samples in the target classes. If we had very low accuracy, it would mean that the model is not able to differentiate the samples based on the data it has, so, the feature ranking would not be relevant to extract the genetic markers.

The Random Forest model has a certain level of stochasticity while training, which means that two different trainings might generate different feature rankings. To ensure that our feature ranking is robust, we aggregate the feature ranking of 500 Random Forest training processes, each one of them composed by 100 Decision Tree classifiers. This large number provides additional robustness to the aggregated feature ranking. The aggregated ranking was used to extract the *most* important features for the task of discriminating the target classes. These will be our potential genetic markers.

Once we had a general ranking, we wanted to look for more specific markers for every one of the germ layers. To obtain them we performed a second experiment. This time we performed four different classifications, classifying each of the germinal layers against the other three (*i.e.*, one vs all). This way, we obtained one ranking specific for each of the germ layers. The best features of these rankings are the corresponding potential genetic markers for each one of the germ layers. Simply put, these new rankings contain the features that can differentiate a germ layer with respect to the other three.

8.1 Hyperparameter Tuning

A Random Forest classifier has several hyperparameters that need to be estimated. Among these hyperparameters, we fixed the class weight (*class_weight*) and the number of trees (*n_estimators*) of the model. The class weight used was *balanced*, which means that the values of the weights were calculated to adjust inversely proportional to the class frequencies in the input data (*i.e.*, the classes with more samples were multiplied by smaller weights). We fixed this parameter for mitigating the effect of having an unbalanced number of samples of each of the target classes. We also fixed the number of trees to 100. This variable controls the stochasticity of the model, the most trees we add the most robust tends to be the model, but it also is slower to train. We expected that with 100 trees, the model would gain robustness while not being computationally prohibitive. Notice that the training process was performed 500 times for every experiment, and every training took roughly 10 seconds while running on a commodity computer.

The rest of the hyperparameters of the model were estimated using Random Search Cross-Validation [37] over the four-class classification task (*i.e.*, all vs all). This kind of cross-validation uses random distributions instead of trying all the combinations of a list of values for every hyperparameter. This method is a lot faster than the traditional grid-search cross-validation and has shown to obtain better results than the grid-search option. Table 5 shows the probabilistic distributions used to search in the hyperparameter space, as well as the best values found by the methodology. We fixed these hyperparameters on all further experiments.

8.2 Feature ranking generation

Since Random Forest feature selection has a certain level of stochasticity, we first assess the stability of the method by performing 500 independent runs. The rankings resultant from each of these runs are then aggregated feature-wise. After the aggregation, the new order of the features is the final feature ranking. We compute the mean ranking of all features as an indicator of robustness. Results indicate a solid consistency among runs, which speaks for the relevance of all further experiments. As Table 6 shows, the top 3 features are the same in the 500 experiments, while the top 15 features are very stable. The aggregated ranking implies, in the worst case, moving a feature two position higher or lower. This indicates that features were not generated at random and had relevance for the problem we are trying to solve.

8.3 Germ layer specific ranking generation

To obtain a feature ranking specific for each germ cell, the Random Forest was trained to discriminate each germ layer from the remaining three, transforming the original multi-class problem with four germ layers into four binary problems (*e.g.*, *Endoderm* vs *Non-Endoderm*).

Hyperparameter	Distribution	Best Value
max_depth	$unif(2, 20)$	13
min_samples_split	$unif(2, 11)$	5
min_samples_leaf	$unif(1, 20)$	3
bootstrap	*unif*([*True*, *False*])	*True*
criterion	*unif*([*gini*, *entropy*])	*entropy*
max_features	*unif*([*auto*, *log2*, *None*])	*None*
class_weight	-	*balanced*
n_estimators	-	100

Table 5: Distributions used on the random search cross-validation and the best hyper-parameters selected for the Random Forest.

Ranking	Mean position	Features
1	1.000	donor_age_at_diagnosis
2	2.000	female
3	3.000	tumour_stage1_Primary_tumour
4	4.018	tumour_stage2_solid_tissue
5	4.992	DEL
6	7.200	tumour_stage2_other
7	7.478	chr_8
8	7.500	TRA
9	7.914	number_of_breaks
10	10.518	proportion_DUP
11	10.910	proportion_DEL
12	12.448	tumour_stage2_lymph_node
13	12.728	proportion_chr_9
14	15.136	t2tINV
15	16.184	proportion_DEL_14

Table 6: The 15 *most* characteristic features among the 313 according to 500 Random Forest runs. The second column contains the mean position of each feature over 500 executions.

We train 500 Random Forests for each germ layer and aggregate these rankings (see the details in Section 8.2). This way, we obtain four rankings, with the most discriminating features for each one of the germ layers. The results obtained (see Table 7) show different feature rankings for every classification experiment, especially on the chromosome related features. These results suggest that the presence or the type of rearrangements on specific chromosomes are related with the germ layer of the cancer cell, and also that our proposed model is able to find those characterisations.

As mentioned before, we have two types of features. Metadata features, those associate with patient specific information; and chromosomal features, those associated with specific genetic variations (*e.g.*, the number of deletions in chromosome 8). We focus on chromosomal features, as this kind of features could help, for example in research lines related to find oncogenes or cancer suppressor genes.

If we compare the results of the different germ layers from Table 7, we can see that germ layers containing more cancer types tend to have more general features at the top of the ranking. For example,the Endoderm germ layer, which is the germ layer with the most cancer types, does not contain specific chromosomal features (like proportion of translocations on the chromosome 5) until the eleventh position. While Ectoderm (which only contains breast cancer samples) has a chromosomal feature in the fifth position. We can see a similar phenomenon in the all *vs* all classification, where the first chromosomal layer is on the seventh position. These results may indicate that chromosomal features are specific per cancer type, and that by aggregating them into germ layers we are losing their specificity. Furthermore, if we could perform the experiment classifying into the cancer types instead of the germinal layers we might obtain more specific markers for every one of the cancer types. As the cells would be a lot more different from one class with respect to another.

Rank	All Germ	ECTODERM	NEURAL_CREST	MESODERM	ENDODERM
1	donor_age	female	donor_age	donor_age	donor_age
2	female	donor_age	ts1_Primary	ts2_blood	female
3	ts1_Primary	ts2_solid_tissue	prop_DUP	ts2_lymph_node	DEL
4	ts2_solid_tissue	TRA	chr_21	DEL	ts2_solid_tissue
5	DEL	chr_8	prop_chr_9	prop_TRA_5	#_of_breaks
6	ts2_other	prop_h2hINV_19	chr_5	female	TRA
7	chr_8	TRA_17	prop_chr_2	#_of_breaks	ts2_lymph_node
8	TRA	prop_DEL_4	prop_DUP_12	chr_19	ts2_other
9	#_of_breaks	t2tINV	ts2_solid_tissue	prop_chr_3	ts1_Primary
10	prop_DUP	prop_TRA_17	chr_6	ts1_Primary	ts2_blood
11	prop_DEL	prop_DEL	prop_chr_5	prop_chr_9	prop_TRA_5
12	ts2_lymph_node	prop_chr_9	ts2_lymph_node	prop_t2tINV	prop_DEL
13	prop_chr_9	prop_chr_5	chr_1	ts2_solid_tissue	prop_chr_1
14	t2tINV	prop_chr_4	DEL_1	ts1_Metastatic	prop_DUP
15	prop_DEL_14	prop_TRA_9	prop_chr_21	prop_TRA	prop_chr_4

Table 7: Best 15 features found for each classification experiment. The second column (All Germ) shows the best features for the all vs all classification task; this column corresponds with the results of Table 6. Third to sixth columns shows the best features for the one vs all classification task.

8.4 Classification results

In order to further validate the features found by the random forest, we look at their performance when classifying the data instances. The model was tested using different sets of the best features of the ranking. This gives some intuitions about the possible relevance of the features extracted by the model and their relation with the data. We report classification results using the top 5, 15, 25, 50, 100, and all 313 features according to the importance ranking produced (reported in Table 8). Due to the unbalanced number of samples of the germ layers, we decided to use the F1 measure instead of the accuracy or any of the other commonly used measures. F1 is a good option for measuring the performance of unbalanced datasets because is the harmonic mean of the precision (the ratio of true positive predicted classes with respect to the predicted number of positive classes) and recall (the ratio of true positive predicted classes with respect to the true number of positive classes). Best F1 measure is obtained by using the top 25 features in most of the experiments. These results show that selecting the best features affects the classification performance significantly, which means that features extracted are relevant for the characterisation of germ layers. Also, it means that low-level features where adding noise to the data, making the task more difficult for the classifier.

	F1 All Germ	F1-ECTO.	F1-NEURAL.	F1-MESO.	F1-ENDO.
All	0.694	0.420	0.877	0.649	0.830
100	0.709	0.494	0.839	0.664	0.840
50	0.722	0.543	0.859	0.660	0.826
25	**0.741**	**0.556**	**0.887**	0.681	**0.841**
15	0.737	0.551	0.879	**0.682**	0.834
5	0.630	0.404	0.818	0.565	0.731

Table 8: Classification results using the top 5, 15, 50 and 100 features, or all of them. The second column (F1 All Germ) shows the mean F1 measure for the all vs all classification task. Third to sixth columns show the F1 measure for the one vs all classification task. Best results in bold.

9. Query-based evaluation

The validation of the top chromosomal features obtained in the previous sections is not straight-forward. It will require thorough analysis from medical experts, to validate the existence of genetic markers associated with them. This process can take from months to years of lab experimentation.

To produce a first coarse evaluation of the results, we use a crowd-based approach based on state of the art on cancer research. The well-known Medical Subject Headings [38] was used, which indexes medical papers from PubMed [39]. PubMed allows querying over 29 million medical abstracts from MEDLINE, life science journals, and online books. Through its search engine, it is possible to find the number of papers mentioning both a cancer type (*e.g.*, Pancreas) and a chromosome on the same text. The result of this search gives an approximate idea of the current medical knowledge concerning the relationship between a type of cancer and a chromosome.

Such query-based evaluation has a limitation. The obtained features are trained to discriminate germ layers. However, as discussed in §5, most medical papers do not work at this granularity. As a result, queries including germ layer terms produce limited results, too few to be considered crowd-based validation.On the other hand, queries including cancer type terms (*e.g.*, breast) provide lots of results, which could be interpreted as consistent evidence. For this reason, we only use this evaluation method to validate the genetic markers found for the Ectoderm germ type, which only contains one cancer type (breast). For this germ layer we query on breast cancer instead of Ectoderm. The other germ types contain several cancer types on variable proportions, and are therefore excluded from the query-based evaluation.

Finding the appropriate terms to query with regards to the features is not straight-forward either. Engineered features are often too specific (*e.g.*, proportion of *h2h* inversions on chromosome 19). For this reason, we limit ourselves to evaluate the relation found with whole chromosomes, disregarding any further particularity of the feature (*e.g.*, chromosome 19, instead of the proportion of h2h inversions on chromosome 19). This loss of detail will not allow us to validate the specific genetic markers we found, but it will still provide evidence regarding the consistency of the methodology.

In particular, 24 queries were performed. One for each chromosome, together with the term for *breast cancer*. For example, one of the queries performed is "Chromosomes", "Human", "Pairs", "Breast" and "Neoplasms". Since not all chromosomes are expected to be related to breast cancer, not all queries will be relevant. We focus on the top three chromosomes related to breast cancer by the number of returned results (Table 9). These are chromosomes 17, 11 and 8. At the same time, we find the top three chromosomes associated with a feature discriminating breast cancer on our results (Table 7 column Ectoderm). These are chromosomes 8, 19 and 17 (chr_8, prop_h2hINV_19 and TRA_17).

Remarkably, two of the three most mentioned chromosomes on papers related to breast cancer, are involved in two of the three most relevant features we found for discriminating the Ectoderm germ layer (*i.e.*, breast cancer). This combination has a random statistical probability of roughly 1%.

Query Chromosomes	Number of Results
17	714
11	250
8	231
1	185
16	165
13	129
3	123
6	106
7	73
10	64

Table 9: Top 10 of the chromosomes with the most results for the Breast cancer queries.

10. Discussion

The results that are shown in this work open up several questions. To start with, the *query-based evaluation* finds two of the top three chromosomes related to breast cancer to be consistent with the literature (see §9). However, what is happening with the missing one?

The chromosome found by the model, but not in the literature, is chromosome 19. This could be either a mistake by the model or relation not yet discovered by the medical community. This is precisely the sort of result with potential impact on the medical domain, as previously unknown genetic markers at gene level could be contained in this chromosome.

The chromosome found in the literature, but not by the model, is chromosome 11. This miss behaviour could be a consequence of its relevance for certain Mesoderm cancer types (Kidney, where it is the first chromosome in several papers, or Ovarian, where it is the second). This, in turn, affects the Random Forest model, since this chromosome will not be discriminative for Ectoderm, even though it may be representative.

The absence of chromosome 11 in our results brought to our attention a couple of limitations of our approach. The first is related to the use of a classifier for extracting feature information. A classifier focuses on the discriminability of features. This might cause the model to oversee features that, while being representative for a particular type of cancer, are not discriminant in the context of several cancer types. This problem might be mitigated by doing pairwise classification instead of one vs all, comparing pairs of cancer types. By doing so, all features that are discriminant for our target cancer type with regards to any of the other cancer types would be identified. This will remain as future work.

The complete, and unfeasible, solution to this problem would be to have a healthy person sample to compare against. However, healthy genomes do not have chromosomal rearrangements. As such, a healthy sample would be empty and impossible to compare against.

11. Conclusions

The results presented in this thesis target the identification of genetic markers. Given the large granularity of features used in our approach, this is not a straight-forward process from the medical perspective. To provide some evidence on the consistency of our approach, we performed three different evaluations: validating the consistency of the aggregated ranking, checking the discriminant capacity of the classifier and a query-based validation against an extensive database of medical papers. In this evaluation, we found that, out of the top three chromosomes identified with breast cancer in the literature, two are also found by the method. This coincidence has a random statistical probability of roughly 1%. This gives us a proportional confidence in asserting that features found by our models are useful guidelines for cancer genetic markers. There are other evidence highlighting the medical consistency of our findings. For example, the most reliable feature for discriminating Ectoderm (*i.e.*, breast cancer) is gender.

Another interesting insight from Table 7 is that different germ layers seem to be related to different break types. While translocations are the most relevant break type for Ectoderm, for Mesoderm and Endoderm deletions seem to be more relevant. The Neural Crest case deserves a specific commentary. The cancer types from the Neural Crest Germ layer are frequently related to children, in particular, Central Nervous System cancer (CNS). Children develop cancer differently when compared to adults.

The same Table 7 also displays a remarkable correlation between the specificity of a germ layer (*i.e.*, how many different cancer types it contains) and the specificity of the features found by our model. On the one hand, the most specific germ layers (Ectoderm and Neural Crest, with only one and two cancer types) have between 5 and 6 chromosome specific features among the top 10 ranked. On the other, the most generic germ layers (Endoderm and Mesoderm) and the all vs all classification (All Germ) have between zero and two chromosome specific features on the top 10. This seems to indicate that specific cancer types could be characterised further if more data became available for analysis.

Randomised Decision Trees build inside the Random Forest algorithm, are among the fastest machine learning models for classification, with a complexity of *O(KNlogN)* [40]. An essential feature of our model and methodology is thus its high scalability. If more data becomes available, we could extract more specific markers for one or more cancer types with minimal computational cost. Beyond being scalable, the method is also robust to high-dimensional and sparse domains, since we treated these appropriately. Notice the actual train data set has 313 features for 2,068 samples, with a 92% of zero values. In this case, the model design was tuned explicitly for this setting, including a large number of decision trees on each random forest, and a large number of random forests to be aggregated.

Summarising the results obtained on this paper; we have obtained potential general markers that could be related to tumour-genesis on the four basic types of germ layers. We have found specific potential markers for each one of the germ layers, obtaining coherent results with respect to the literature on the subject. Finally, we have obtained a general method for genetic marker mining, that could be generalised when more data becomes available. The continuation of this work requires extensive experimentation by medical experts in order to test and validate our many hypotheses.

References

[1] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[2] Joe W. Albertson, Donna G., Collins, Colin, McCormick, Frank, Gray. Chromosome aberrations in solid tumors. *Nature Genetics*, 34(4):369–376, 2003.

[3] Monique Nicole Helena Luijten, Jeannie Xue Ting Lee, and Karen Carmelina Crasta. Mutational game changer: Chromothripsis and its emerging relevance to cancer. *Mutation Research - Reviews in Mutation Research*, 777(March):29–51, 2018.

[4] I Nishisho, Y Nakamura, Y Miyoshi, Y Miki, H Ando, A Horii, K Koyama, J Utsunomiya, S Baba, and P Hedge. Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science*, 253(5020):665–669, 1991.

[5] J Whang-Peng, CS Kao-Shan, EC Lee, PA Bunn, DN Carney, AF Gazdar, and JD Minna. Specific chromosome defect associated with human small-cell lung cancer; deletion 3p(14-23). *Science*, 215(4529):181–182, 1982.

[6] Y.-S. Shih. Families of splitting criteria for classification trees. *Statistics and Computing*, 9(4):309–315, Nov 1999.

[7] Torben R Kasparek and Timothy C Humphrey. Dna double-strand break repair pathways, chromosomal rearrangements and cancer. In *Seminars in cell & developmental biology*, volume 22, pages 886–897. Elsevier, 2011.

[8] Scott A Tomlins, Bharathi Laxman, Saravana M Dhanasekaran, Beth E Helgeson, Xuhong Cao, David S Morris, Anjana Menon, Xiaojun Jing, Qi Cao, Bo Han, et al. Distinct classes of chromosomal rearrangements create oncogenic ets gene fusions in prostate cancer. *Nature*, 448(7153):595, 2007.

[9] Peter J Campbell, Philip J Stephens, Erin D Pleasance, Sarah O'Meara, Heng Li, Thomas Santarius, Lucy A Stebbings, Catherine Leroy, Sarah Edkins, Claire Hardy, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics*, 40(6):722, 2008.

[10] Brian K Hall. The neural crest as a fourth germ layer and vertebrates as quadroblastic not triploblastic. *Evolution & development*, 2(1):3–5, 2000.

[11] Kun Zhang and Hong Wang. Cancer Genome Atlas Pan-cancer analysis project. *Chinese Journal of Lung Cancer*, 18(4):219–223, 2015.

[12] Pan-cancer atlas. `https://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html`.

[13] Katherine A Hoadley, Christina Yau, Toshinori Hinoue, Denise M Wolf, Alexander J Lazar, Esther Drill, Ronglai Shen, Alison M Taylor, Andrew D Cherniack, Vésteinn Thorsson, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304, 2018.

[14] Ashton C Berger, Anil Korkut, Rupa S Kanchi, Apurva M Hegde, Walter Lenoir, Wenbin Liu, Yuexin Liu, Huihui Fan, Hui Shen, Visweswaran Ravikumar, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer cell*, 33(4):690–705, 2018.

[15] Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385, 2018.

[16] Michael Burrows and David J Wheeler. A block-sorting lossless data compression algorithm. 1994.

[17] Michael B Cook, Sanford M Dawsey, Neal D Freedman, Peter D Inskip, Sara M Wichner, Sabah M Quraishi, Susan S Devesa, and Katherine A McGlynn. Sex disparities in cancer incidence by period and age. *Cancer Epidemiology and Prevention Biomarkers*, 18(4):1174–1182, 2009.

[18] Robert W Miller, John L Young Jr, and Biljana Novakovic. Childhood cancer. *Cancer*, 75(S1):395–405, 1995.

[19] James G Gurney, Richard K Severson, Scott Davis, and Leslie L Robison. Incidence of cancer in children in the united states. sex-, race-, and 1-year age-specific rates by histologic type. *Cancer*, 75(8):2186–2195, 1995.

[20] A J Nixon, D Neuberg, D F Hayes, R Gelman, J L Connolly, S Schnitt, A Abner, A Recht, F Vicini, and J R Harris. Relationship of patient age to pathologic features of the tumor and prognosis for patients with stage i or ii breast cancer. *Journal of Clinical Oncology*, 12(5):888–894, 1994. PMID: 8164038.

[21] Jill A Foster, Gregory D Salinas, Dorcas Mansell, James C Williamson, and Linda L Casebeer. How does older age influence oncologists' cancer management? *The oncologist*, 15(6):584–592, 2010.

[22] Nophar Geifman, Raphael Cohen, and Eitan Rubin. Redefining meaningful age groups in the context of disease. *Age*, 35(6):2357–2366, 2013.

[23] Jonathan M Samet, Erika Avila-Tang, Paolo Boffetta, Lindsay M Hannan, Susan Olivo-Marston, Michael J Thun, and Charles M Rudin. Lung cancer in never smokers: clinical epidemiology and environmental risk factors. *Clinical Cancer Research*, 15(18):5626–5645, 2009.

[24] Mimi C. Yu and Jian-Min Yuan. Environmental factors and risk for hepatocellular carcinoma. *Gastroenterology*, 127(5, Supplement 1):S72 – S78, 2004.

[25] Wong-Ho Chow, Linda M Dong, and Susan S Devesa. Epidemiology and risk factors for kidney cancer. *Nature Reviews Urology*, 7(5):245, 2010.

[26] The Endogenous Hormones and Breast Cancer Collaborative Group. Endogenous Sex Hormones and Breast Cancer in Postmenopausal Women: Reanalysis of Nine Prospective Studies. *JNCI: Journal of the National Cancer Institute*, 94(8):606–616, 04 2002.

[27] Annekatrin Lukanova and Rudolf Kaaks. Endogenous hormones and ovarian cancer: epidemiology and current hypotheses. *Cancer Epidemiology and Prevention Biomarkers*, 14(1):98–107, 2005.

[28] Robert Hoover, Richard Everson, JosephF Fraumeni JR, and MaxH Myers. Cancer of the uterine corpus after hormonal treatment for breast cancer. *The Lancet*, 307(7965):885–887, 1976.

[29] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

[30] P. H. Swain and H. Hauska. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3):142–147, July 1977.

[31] Leo Breiman, JH Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. wadsworth, 1984. *Google Scholar*, 1993.

[32] D Lavanya and K Usha Rani. Performance evaluation of decision tree classifiers on medical datasets. *International Journal of Computer Applications*, 26(4):1–4, 2011.

[33] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[35] Yanjun Qi. Ensemble Machine Learning. *Ensemble Machine Learning*, pages 307–323, 2012.

[36] Ramón Díaz-Uriarte and Sara Alvarez de Andrés. Supplementary material for "Gene selection and classification of microarray data using random forest". *BMC Bioinformatics*, 7:1–73, 2005.

[37] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. 13:1–25, 2012.

[38] Mesh: The nlm controlled vocabulary thesaurus used for indexing articles for pubmed. `https://www.ncbi.nlm.nih.gov/mesh`.

[39] Pubmed: Citations for biomedical literature from medline, life science journals, and online books. `https://www.ncbi.nlm.nih.gov/pubmed`.

[40] Gilles Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.