May 2019

Random Forest as a Tumour Genetic Marker Extractor

Raquel PÉREZ-ARNAL [a], Dario GARCIA-GASULLA [a], David TORRENTS [a], Ferran PARÉS [a], Ulises CORTÉS [a,b], Jesús LABARTA [a,b] and Eduard AYGUADÉ [a,b]

[a] *Barcelona Supercomputing Center (BSC)*
({raquel.perez, dario.garcia}@bsc.es)
[b] *Universitat Politècnica de Catalunya - BarcelonaTech (UPC)*

Abstract. Finding tumour genetic markers is essential to biomedicine due to their relevance for cancer detection and therapy development. In this paper, we explore a recently released dataset of chromosome rearrangements in 2,586 cancer patients, where different sorts of alterations have been detected. Using a Random Forest classifier, we evaluate the relevance of several features (some directly available in the original data, some engineered by us) related to chromosome rearrangements. This evaluation results in a set of potential tumour genetic markers, some of which are validated in the bibliography, while others are potentially novel.

Keywords. cancer research, chromosomal rearrangements, tumour genetic markers, Random Forest

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 genes of the 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 several tumour types. This implies that the presence of some specific gene expressions or DNA changes, like chromosomal rearrangements, can be used as tumour markers to characterize different types of cancer. Finding these markers can be useful in several ways, like predicting disease outcome or response to treatment. Some examples of chromosomal rearrangements as tumour genetic markers are mutations of chromosome 5q21 on colorectal cancer [4] or deletions on chromosome 3p on lung cancer [5].

In this paper, we present a new methodology to find potential tumour genetic markers from a dataset of chromosomal rearrangements. This data consists of a set of files, each one of them from a patient, where we have a sequence of rearrangements, repre-

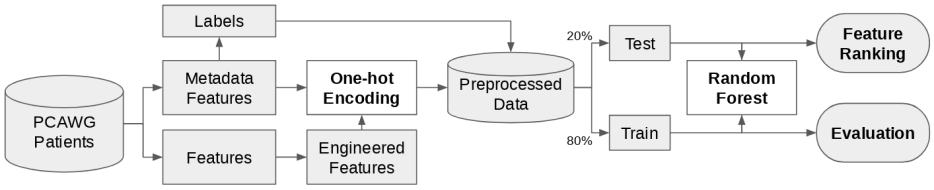

Figure 1. Summary of the data pipeline.

sented as triplets *source base pair*, *destiny base pair* and *rearrangement type*. From these sequences of triplets, we engineer several features related to the chromosomal rearrangements, and then, use a Random Forest as a feature extractor. In this way, we generate a ranking of the features by their importance. The best features of the ranking are new potential genetic markers found by the methodology. In our experiments, we extract more than thirty potential markers. Figure 1 shows the basic data pipeline used on this work.

2. Data

The data used in this study comes from genetic cancer data from the PanCancer Analysis of Whole Genomes (PCAWG) project [6]. The PCAWG study is an international collaboration to identify common patterns of mutation in more than 2,800 cancer whole genomes from the International Cancer Genome Consortium. Let us remark that scientific works with a primary focus on pan-cancer are under publication embargo until July 25, 2019. For this reason, there is no comparison of our results with similar approaches. This situation also guarantees the novelty of all our experiments.

Using an *in-house* pipeline designed to analyse tumour genomes in a clinical and research context, we have identified breakpoints that inform of sites of genomic and chromosomal rearrangements in the PCAWG data. This pipeline identifies breakpoints using the information of reads, and paired-end reads mapping from whole-genome sequencing, using Burrows-Wheeler Aligner (BWA). These predictions have passed strict filtering, ensuring a high-quality set of variants.

Our data is composed of 2,586 patients (samples), where each patient has a variable number of breaks (features). In this dataset there are 21 possible types of cancer (*e.g.*, breast cancer, liver cancer, pancreas cancer, *etc.*). At the same time, every sample comes from one of four possible germ layers (*i.e.*, basic cell types). The germ layers are *ECTODERM*, *ENDODERM*, *MESODERM* and *NEURAL CREST*. In this study we consider 4 types of chromosomal breaks: deletions (*DEL*), translocations (*TRA*), duplications (*DUP*) and two kinds of inversions (*t2tINV* and *h2hINV*).

The total length of the human genome is over three billion base pairs. Those pairs are divided into 24 chromosomes. Characterizing breaks at a base pair level would end up with very sparse data on a high dimensional space (2,586 samples over 3 billion pairs). For this reason, we choose to reduce the granularity of the features used, working instead at chromosome level (*e.g.*, instead of considering a deletion on the gene 15p5, we consider a deletion on chromosome 15). As a result, we end up with 2,586 samples over the 24 chromosomes.

We use the germ layers as a generalization of the cancer types because of the small number of data samples, but with more data, this methodology could extract the markers

	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	1281	729	367

Table 1. Number of samples for every cancer type on the dataset, with their corresponding germ layer.

for every cancer type. Table 1 shows how many samples are in the dataset for each cancer type, and how are these samples distributed over the germ layers.

2.1. Preprocessing

The first pre-processing step performed on the data was to remove all the patients with no germ layers nor cancer type labels available. After removing these unlabelled samples, we extracted a set of features from additional metadata available in the dataset (see the bottom part of Table 2) and engineered another set (see the top part of Table 2). The engineered features are related to the kind of breaks or their position in the DNA. Features include, among others, the number of rearrangements on every chromosome or the number of rearrangements of each type.

Metadata include gender, age, tumour_stage1 and tumour_stage2 as features and germ layer as target variable (i.e., label). The last two features stand for the clinical stage of the tumour (non-genetic information). We initially had two options to select as target variable: the cancer type and the germ layer. Ideally, we would have selected cancer type but, the number of samples (2,586) *w.r.t.* the high number of cancer types (21) makes the problem unfeasible. For this reason, we end up selecting the germ layer instead of as a generalization of the cancer type. The final set of features (engineered and metadata features) as shown in Table 2 accompanied by a brief description.

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	Methastasic, Primary or Recurrent	3
ts_2_category	NOS, bone marrow, periphleal blood, derived from tumour, methastasis to lymph node, methastasis to distant location, other or solid tissue	9
Total number of features		313

Table 2. Genetic features (top) and metadata features (bottom) extracted from the dataset.

After the extraction of these features, we perform a one-hot-encoding over all categorical features (both tumour stages). Furthermore, we impute 119 missing values for the age feature using the Multivariate Imputation by Chained Equations (MICE) [7].

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 1). This way, we obtained a training partition with 2,068 samples and a test partition with 519 samples. Both of them containing 313 features, including boolean features from the one-hot-encoding.

3. Approach

On this study, a Random Forest model [8] was used to identify which chromosomal rearrangements, and in which location, contain genetic markers. 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 [9,10].

We trained the Random Forest using the set of 313 features extracted from the chromosomal rearrangement data and the patient metadata (see Table 2). The result of training this model is the feature importance order (where the best features are positioned at the beginning, and the worse are positioned at the end) and the classification results.

In the first experiment, the feature importance orders obtained from training 300 Random Forest were used, each one containing 100 Decision Tree classifiers, to gen-

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 3. Distributions used on the random search cross-validation and the best hyperparameters selected for the Random Forest.

erate our aggregated feature ranking. This large number provides additional robustness to the aggregated feature ranking. The aggregated ranking was used to extract the *most* important features. These will be our potential genetic markers.

In the second experiment, this process was repeated four times, discriminating each germ layer type against all the rest (*i.e.*, one vs all). Four new rankings representing the best features (*i.e.*, potential genetic markers) for each one of the specific germ layers were obtained.

4. Experiments

All Random Forest models target to classify the patients by the germ layer associated with their cancer. This classification has two outputs of interest: the feature importance and classification results. The feature importance order extracts the relationships between the features and the germ layers, while classification results prove that these results are relevant.

To estimate the hyperparameters of the Random Forest, Random Search Cross-Validation [11] was used, with three validation partitions and the parameter distributions presented in Table 3. Those hyperparameters were tuned for the four class classification task (*i.e.*, all vs all), and fix their value on all further experiments.

4.1. Feature ranking generation

Since Random Forest feature selection has a certain level of stochasticity, we first assess the stability of the method by performing 300 independent runs. We compute the mean ranking of all features as an indicator of robustness. Results (shown in Table 4) indicate a very strong consistency among runs, which speaks for the relevance of all further experiments.

4.2. Germ layer specific ranking generation

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

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 4. The 15 *most* characteristic features found by the 300 Random Forest runs. The second column contains the mean position of each feature over the 300 executions.

After transforming the data labels, we train 300 Random Forests for each germ layer and aggregate these rankings (see the details in Section 4.1). This way, we obtain four rankings, with the *most* discriminating features for each one of the germ layers.

The results obtained (see Table 5) 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.

4.3. Classification results

To have some intuition about the importance of the features extracted by the model and their relation with the data, the model was tested using different sets of the best features of the ranking. We report classification results using the best 5, 15, 25, 50, 100, and all 313 features based on its importance ranking (reported in Table 6). Best F1 measure is obtained by using the best 25 features in most of the experiments.

5. Query-based evaluation

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

In order to produce the first evaluation of the produced results, we use a crowd-based approach based on state of the art on cancer research. The well-known Medical Subject Headings [12] was used, which indexes medical papers from PubMed [13]. 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

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 5. 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 on Table 4. Third to the sixth column shows the best features for the one vs all classification task.

	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 6. 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 the sixth column show the F1 measure for the one vs all classification task. Best results in bold.

mentioning both a cancer type (*e.g.*, Pancreas) and a chromosome. The result of this search can give an idea of the current medical knowledge regarding the relationship between a type of cancer and a chromosome.

The obtained features are trained to discriminate germ layers. However, most medical papers do not work at this granularity. Instead, cancer type queries produce larger, and thus more representative results. For this reason we can only use this evaluation method reliably on the ECTODERM germ type, which only contains one cancer type (breast). The other germ types contain several cancer types on variable proportions. On the other hand, the engineered features are not appropriate for a straight-forward query on medical papers. These 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).

In particular, 24 queries were performed. One for each chromosome, together with the term for *breast cancer*. 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. 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. These are chromosomes 8, 19 and 17 (chr_8, prop_h2hINV_19 and TRA_17).

Remarkably, two of the three 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%.

6. 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 §5). However, what is happening with those missing?

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

The chromosome found in the literature, but not by the model, is chromosome 11. This behaviour can 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 on 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.

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.

7. Conclusions

The results presented in this paper target the guidance 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 a partial, 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 proportional confidence in asserting that the features found by our models are useful guidelines for cancer genetic markers. There is 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 5 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. This cancer type is frequently related to children, in particular, Central Nervous System cancer (CNS). Children develop cancer differently when compared to adults. Our intuition is that chromosomal duplication is more consistent with the growth patterns of children, which would explain the findings of our model.

The same Table 5 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 characterized further if more detail became available for analysis.

Randomized Decision Trees build inside the Random Forest algorithm, are among the fastest machine learning models for classification, with a complexity of *O(KNlogN)* [14]. 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 a 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.

Summarizing the results obtained on this paper; we have obtained potential general markers that could be related to tumour-genesis on the four basic germ layer types. We have found specific potential markers for every one of the significant germ layers, obtaining coherent results respect to the known bibliography on the subject. Finally, we have obtained a general method for genetic marker mining, that could be generalized by the growth of the dataset. The continuation of this work requires extensive experimentation by medical experts in order to test and validate our many hypotheses.

Acknowledgements

This work is partially supported by the Joint Study Agreement no. W156463 under the IBM/BSC Deep Learning Center agreement, by the Spanish Government through Programa Severo Ochoa (SEV-2015-0493), by the Spanish Ministry of Science and Technology through TIN2015-65316-P project, and by the Generalitat de Catalunya (contracts

2017-SGR-1414). Pro. U. Cortés is a member of the Sistema Nacional de Investigadores (Level III) (SNI-III). México. We would like to thank MD. Adrián Puche Gallego and PhD. Davide Cirillo for useful discussions and guidance.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] J. W. Albertson, Donna G., Collins, Colin, McCormick, Frank, Gray, "Chromosome aberrations in solid tumors," *Nature Genetics*, vol. 34, no. 4, pp. 369–376, 2003.

[3] M. N. H. Luijten, J. X. T. Lee, and K. C. Crasta, "Mutational game changer: Chromothripsis and its emerging relevance to cancer," *Mutation Research - Reviews in Mutation Research*, vol. 777, no. March, pp. 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*, vol. 253, no. 5020, pp. 665–669, 1991.

[5] J. Whang-Peng, C. Kao-Shan, E. Lee, P. Bunn, D. Carney, A. Gazdar, and J. Minna, "Specific chromosome defect associated with human small-cell lung cancer; deletion 3p(14-23)," *Science*, vol. 215, no. 4529, pp. 181–182, 1982.

[6] K. Zhang and H. Wang, "Cancer Genome Atlas Pan-cancer analysis project," *Chinese Journal of Lung Cancer*, vol. 18, no. 4, pp. 219–223, 2015.

[7] S. v. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, pp. 1–68, 2010.

[8] 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*, vol. 12, pp. 2825–2830, 2011.

[9] Y. Qi, "Ensemble Machine Learning," *Ensemble Machine Learning*, pp. 307–323, 2012.

[10] R. Díaz-Uriarte and S. Alvarez de Andrés, "Supplementary material for Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, pp. 1–73, 2005.

[11] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," vol. 13, pp. 1–25, 2012.

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

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

[14] G. Louppe, "Understanding random forests: From theory to practice," *arXiv preprint arXiv:1407.7502*, 2014.