NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

FACULTY OF BIOLOGY

DEPARTMENT OF CELL BIOLOGY AND BIOPHYSICS

POSTGRADUATE PROGRAMME

«BIOINFORMATICS-COMPUTATIONAL BIOLOGY»

MASTER THESIS

“Gradient Boosted Trees on Transcriptomics: Diagnosis of Acute Myeloid Leukaemia”

Ioanna Soulioti

Athens, 2021

I dedicate this thesis to my parents,
Zafeiria and Miltiadis,
who have given me invaluable educational opportunities.

Acknowledgments

I would first like to thank the supervisor of my thesis, Associate Professor Michael Filippakis, who accepted to be my supervisor, for his valuable help and guidance during my studies and my thesis.

I would like to acknowledge and give my warmest thanks to Dr. Athanasios Angelakis who made this work possible, as well as his valuable assistance, cooperation and advice throughout this work.

I am also grateful to Associate Professor Vassiliki Ikonomidou and Professor Ioannis Trougakos for reading my work and being part of my three-member Advisory Committee.

Finally, I must express my very profound gratitude to my parents, Zafeiria and Miltiadis, my sister Maria and my long-term partner Theodoros for their continuous support and love over the years as well as their understanding when undertaking my research and writing my thesis.

Abstract

Acute myeloid leukaemia (AML) is a type of cancer which mostly occurs in adults with occasionally lack of symptoms. It is characterized by proliferative, abnormally differentiated and infrequently poorly differentiated hemopoietic cells (Döhner et al., 2015). It affects 0.3 to 5.3 per 100000 people around the world each year. If the AML patients do not undergo treatment, the median survival is approximately 2 months (Sekeres et al., 2004). Its quick diagnosis is very important and can benefit the overall survival (Mottal et al., 2020).

The current diploma thesis describes the work of Angelakis & Soulioti, 2021. Machine learning techniques are applied on transcriptomics data in order to develop a new screening tool which could predict if an individual has AML or is healthy. More specifically, a state-of-the-art machine learning algorithm which belongs to the category of gradient boosted trees, CatBoost, is applied on gene expression microarray datasets that consist of 3374 individuals, AML patients and healthy subjects, and 34 probe sets as features (CatBoost34) and a subset of 3374 subjects which consists of 2177 individuals, their age and 26 probe sets as features (CatBoost26). The performance of CatBoost26 model is the best one in the literature as regards the prediction of AML using similar or not data.

Table of Contents

Figures

Tables

Table 9: Performance of the *k*-NN of Roushangar & Mias, 2019 of the 10CV on 80% training set and on the 20% validation set of 3374 individuals and 984 probe sets.

Abbreviations

AML	acute myeloid leukaemia
ALL	acute lymphocytic leukaemia
CML	chronic myeloid leukaemia
CLL	chronic lymphocytic leukaemia
T-ALL	T-cell acute lymphoblastic leukaemia
MPAL	mixed phenotype acute leukaemias
BM	bone marrow
PB	peripheral blood
WHO	World Health Organization
ICU	intensive care unit
FAB	French-American-British
MRC	myelodysplasia-related changes
t-MN	therapy-related myeloid neoplasms
NOS	not otherwise specified
DS	Down syndrome
PM	perfect match
MM	mismatch
EST	expressed sequence tag
cDNA	Complementary DNA; DNA copy of a messenger mRNA (mRNA) molecule produced by reverse transcriptase
bp	base pairs
ML	machine learning
GEP	gene expression profiling
DT	decision tree
RF	random forest
GBT	gradient boosted tree
GOSS	Gradient-based One-Side Sampling
EFB	Exclusive Feature Bundling
*k*NN	*k*-Nearest-Neighbors
LASSO	least absolute shrinkage and selection operator
SVM	support vector machine
DNN	deep neural network
tp	true positive
tn	true negative
fp	false positive
fn	false negative
NPY	not published yet
10CV	ten fold cross validation
PVC	prediction values change
LFC	loss function change

1. Hematological malignancies

Hematological malignancies are diseases originating from cells of the bone marrow (BM) and the lymphatic system and have been categorized by The World Health Organization (WHO) into two basic groups: myeloid and lymphoid neoplasms (Bahakeem & Qadah, 2020; Rodriguez-Abreu et al., 2007). They are separated further into three main categories: leukaemia, lymphoma and plasma cell neoplasms (Rodriguez-Abreu et al., 2007).

1.1. Leukaemia

There are two forms of leukaemia, acute and chronic form, and each form has two types, lymphogenic and myelogenic. Leukaemias are broadly separated into four major classes; acute lymphocytic leukaemia (ALL), chronic lymphocytic leukaemia (CLL), acute myeloid leukaemia (AML) and chronic myeloid leukaemia (CML). Their clinical characteristics and prognosis differ. It is estimated that more than 250000 people worldwide are diagnosed with leukaemia each year (Pejovic et al., 2002; Rodriguez-Abreu et al., 2007).

1.1.1. Acute leukaemias

Acute leukaemias, AML and ALL, are rare types of cancer and accounts of less than 3% of all malignancies. Although their infrequency, they are one of the major causes of death in young age. The incidence of AML is approximately 2.5 per 100000 persons and for ALL 1.3 per 100000. The average age of ALL diagnosis is 10 years when the median age of AML diagnosis is 66 years.

In most cases, the cause of acute leukaemias is unknown. Some cases of AML are related to previous chemotherapy or radiation therapy. There are also genetic and immunologic factors which contribute to developing the disease (Pejovic et al., 2002; WHO, 2014).

1.1.1.1. AML

AML is the most commonly diagnosed type of leukaemia in adults with an incidence of 20000 cases per year in the United States. Despite the fact that can occur in any age, from newborns

to elderly people, is more often reported in older adults with a median age at 66 years, with 54% patients over 65 years and 33% over 75 years. The incidence of AML increases dramatically with age (De Kouchkovsky & Abdul-Hay, 2016; Johansson & Harrison, 2021; Kumar, 2011; Short et al., 2018; WHO, 2014). AML also accounts for 15-20% of childhood leukaemia (Agaian et al., 2014).

Its etiology is the result of repeated somatic mutations that give a proliferative advantage, therefore they increase the risk of malignant tumors. Sequencing efforts have identified mutations in FLT3, NPM1, KIT, CEBPA, TET2, DNMT3A and IDH1 genes but one quarter of AML patients carry no mutations in these leukaemia-associated genes (Sekeres et al., 2004; Welch et al., 2012).

Survival of AML patients depends on a lot of conditions such as age, genetic subtype, performance status, coexistent medical condition or previous diseases and sex, but if the patients do not undergo the necessary treatment the average survival is 2 months (Röllig & Ossenkoppele, 2021; Sekeres et al., 2004).

AML symptoms vary and some patients experience no symptoms until the diagnosis. As in many types of cancer, its quick diagnosis even in an intensive care unit (ICU) is essential for the recovery of patients, particularly in the case of children (Mottal et al., 2020).

1.1.1.2. Classification of AML

In 1970, the French-American-British (FAB) Cooperative Group proposed an AML classification system. In this system, AML is classified to its subtypes (M0-M7) according to morphologic and cytochemical characteristics despite the fact that it is unclear if this system enhances the already available prognostic information from cytogenetics. Morphologic and cytochemical characteristics misleading in about 10-15% of cases, hence immunophenotypic, cytogenetic and molecular genetic analysis are essential. The scope of this classification system is to compare easier the different AML cases (Pejovic et al., 2002; Walter et al., 2013).
In 2017, the 4th edition of the WHO Classification of Tumors of Hematopoietic and Lymphoid Tissue has been published. AML is categorized by WHO into 6 groups: AML with recurrent

genetic abnormalities; AML with myelodysplasia-related changes (MRC); therapy-related myeloid neoplasms (t-MN); AML, not otherwise specified (NOS); myeloid sarcoma; and myeloid proliferations related to Down syndrome (DS) (Hwang, 2020).

Table 1. Classification of AML according to FAB.

FAB subtype	Name
M0	Undifferentiated AML
M1	AML without maturation (poorly differentiated)
M2	AML with maturation (more differentiated)
M3	Acute promyelocytic leukaemia (APML)
M4	Acute myelomonocytic leukaemia (AMML) Subtype: **1. M4 eos: Acute myelomonocytic leukaemia with >5% eosinophils**
M5	Acute monocytic leukaemia Subtypes: **1. M5a: Acute monoblastic leukaemia - poorly differentiated** 2. **M5b: Acute monocytic leukaemia – more differentiated**
M6	Acute erythroblastic leukaemia
M7	Acute megakaryoblastic leukaemia

Table 2. Classification of AML according to WHO.

Categories
1. AML with recurrent genetic abnormalities
AML with a translocation between chromosomes 8 and 21 [t(8;21)]
AML with a translocation or inversion in chromosome 16 [t(16;16) or inv(16)]
Acute promyelocytic leukaemia (APL) with the *PML-RARA* fusion gene
AML with a translocation between chromosomes 9 and 11 [t(9;11)]

AML with a translocation between chromosomes 6 and 9 [t(6:9)]
AML with a translocation or inversion in chromosome 3 [t(3;3) or inv(3)]
AML (megakaryoblastic) with a translocation between chromosomes 1 and 22 [t(1:22)]
AML with the *BCR-ABL1* (*BCR-ABL*) fusion gene
AML with mutated *NPM1* gene
AML with biallelic mutations of the *CEBPA* gene
AML with mutated *RUNX1* gene
2. AML with myelodysplasia-related changes
3. Therapy-related myeloid neoplasms
4. AML, not otherwise specified
AML with minimal differentiation
AML without maturation
AML with maturation
Acute myelomonocytic leukaemia
Acute monoblastic/monocytic leukaemia
Pure erythroid leukaemia
Acute megakaryoblastic leukaemia
Acute basophilic leukaemia
Acute panmyelosis with fibrosis
5. Myeloid sarcoma
6. Myeloid proliferations related to Down syndrome

1.1.1.3. Diagnosis of AML

The AML diagnostic procedure demands a variety of laboratory examinations. Firstly, the doctor will focus on the patient's symptoms and his medical history. Later will go through a physical examination, which means the doctor pays attention on the patient's eyes, lymph nodes, mouth and skin as well as he looks for possible signs of infection and bleeding areas. After that, a peripheral blood (PB) microscopic morphological examination is required. The incidence of excess number of blast cells in a PB sample is a characteristic of leukaemia. Depend on the results of the previous process, BM sample will be examined for additional morphological, immunophenotyping and cytogenetic analysis. The FAB classification system requires the presence of at least 30% in PB or BM sample, while the WHO classification system requires 20% blasts in PB or BM (American Cancer Society, 2012; Percival et al., 2017).

The manual examination of blood and bone marrow samples has a variety of drawbacks. Since it is done manually is a time-consuming procedure. It is also based on pathologists experience hence it is prone to human errors. On the other hand, machine learning (ML) diagnostic approaches have the potential to low the cost (MoradiAmin et al., 2016; Warnat-Herresthal et al., 2020) Therefore the need for automatic diagnosis, which could help hematologists to identify easier and earlier AML and low the cost of diagnosis, arises (Goutam & Sailaja, 2015).

1.1.2. Chronic leukaemias

Chronic leukaemias are separated into CML and CLL.

1.1.2.1. CML

CML presentation increases with age but affects all age groups. The median age of CML diagnosis is 53 years. It is diagnosed by a reciprocal translocation, in which the Abelson murine leukaemia (ABL) proto-oncogene moves from its normal site on 9q34 to 22q11. The 9q34 is called ABL and the 22q11 is called the breakpoint cluster region (BCR). The product of the chimeric gene is a tyrosine kinase that enhances cell division and diminish apoptosis of mature myeloid cells (Burke & Startzell, 2008; Pejovic et al., 2002; Sawyers, 1999).

1.1.2.2. CLL

CLL is the most common type of leukaemia and it is estimated that 1.8 to 3 per 100000 persons diagnosed in the United States. The median age of CLL diagnosis is 72 years and most patients are older than 60 years. It is not considered to be curable but can be controlled for many years. The patients may need to follow a clinical course for a long time (Baumann et al., 2020; Pejovic et al., 2002).

2. Transcriptomics

Transcriptomics is the study of the total transcriptome, corresponds to all RNA transcripts, of an organism. It enables the study of gene expression changes in different tissues and conditions. Understanding the transcriptome means understanding the functional elements of genome and the molecular elements of cells and tissues, as well as the development and disease.

Transcriptomic technologies have multiple applications in biomedical field. They are used for diagnosis and profiling of a variety of diseases and they have been characterized by two basic methods: microarrays and RNA sequencing (RNA-Seq) (Lowe et al., 2017; Wang et al., 2009).

2.1. Microarrays

Microarray technology facilitates the analysis of thousands of transcripts at the same time. Microarrays consist of probes, which are short oligonucleotides which are attached to a solid surface. Transcript abundance determined by the fluorescently labeled transcripts that are hybridized to these probes (Lowe et al., 2017).

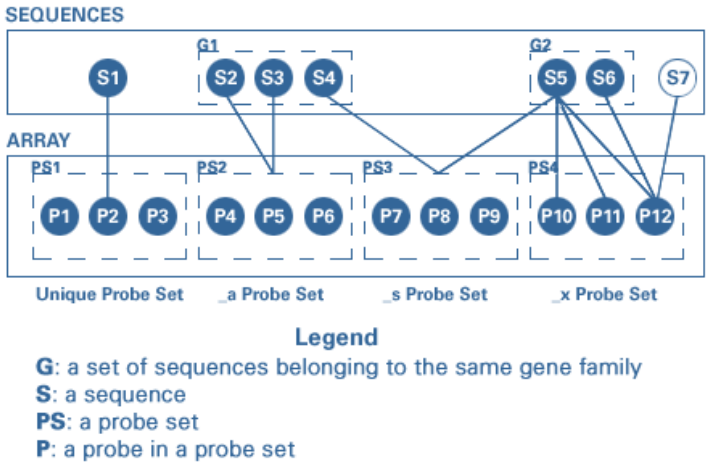
Affymetrix GeneChip microarrays are the most popular ones. Each gene on an Affymetrix microarray correspond to a probe set, which composed of 11 different pairs of 25-bp oligonucleotides. These oligos cover parts of the transcribed region of the gene. Each pair consists of a perfect match (PM) and a mismatch (MM) nucleotide oligomer. The PM is designed to match exactly the sequence of interest and the MM is used for distinguishing the noise due to non-specific match. Not all the probes in a probe set match one known transcript. Some probes hit alterative transcripts from the same gene while others match transcripts from different genes (N. Jiang et al., 2007). Each probe set has an identifier which consist of digits followed by '_a', '_s', '_x' and ending in '_at', which indicates that the probe set is designed in such a way to detect the antisense strand of the gene. Probes in a gene family probe set '_a' match to the same set of sequences that belong to the same gene family. An identical '_s' probe set is designed in a way that all probes in the probe set cross-hybridize to the same set of sequences which are not defined as from the same gene family. A mixed '_x'

probe set has at least one probe that cross-hybridize to other sequences while a unique probe set does not match any other sequence (Thermo Fisher Scientific Inc., 2017).

Probes in DNA microarrays are most of the times, cDNA clones. The majority of these clones are expressed sequence tags (EST's) or cDNA clones. A basic problem with these clones is that they lack of reliable annotations of their sequence data (Liu et al., 2010).

EST's are small reads (200-800 bp) come from one-shot sequencing of randomly selected cDNA clones, thus they represent the expressed part of the genome and are used to identify gene transcripts (Parkinson & Blaxter, 2009).

Figure 1. Representation of the different probe set types.



(Thermo Fisher Scientific Inc., 2017)

2.2. RNA-Seq

RNA-seq is a high-throughput sequencing method which is used to map and reveal the quantity of transcripts in a sample. The first step is the preparation of the cDNA fragments with adaptors attached to one end or both ends. Each molecule is used to generate nucleotide sequences which are commonly 30 to 400 bp depending on the sequencing approach. Later, the reads are aligned to the reference genome or transcripts or assembled *de novo* when the reference genome is not available or lack of annotations. *De novo* assembly is used to produce

a map which contains transcripts' structure and/or level of expression for each gene (Robertson et al., 2010; Wang et al., 2009).

3. Machine learning (ML)

According to Naqa and Murphy, "ML is an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment". ML algorithms are used to teach computers how to deal with enormous data and interpret the information produced from them (Naqa & Murphy, 2015; Rao, 2016). It combines computer science and statistics and its goal is to improve computers in order to execute a task through experience, and thus make the human life easier (Jordan & Mitchell, 2015).

3.1. Categories of ML

ML is divided into four main categories: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. Supervised ML is used for classification and regression tasks and is more commonly used for medical purposes than unsupervised ML. In supervised ML the inputs are labeled, and the goal of this ML technique is to create a function that maps the input to the output. In unsupervised ML the inputs are not labeled, and the models are tasked to find interesting structures in the data. It is mostly used for clustering and feature reduction. The semi-supervised learning uses a combination of labeled and unlabeled inputs. In reinforcement learning, the algorithms receive feedback from the environment when selecting an output for a given input (Muhsen et al., 2020; Rao, 2016; Simeone, 2018).

3.2. ML in the diagnosis of AML

ML has multiple potential implications in medicine, including diagnosis, risk status assignment, prognosis and treatment planning of hematological malignancies (Radakovich et al., 2020). It has also been applied to gene expression analysis in order to classify tumors, predict patients' survival and identify new therapeutic targets (Gal et al., 2019).

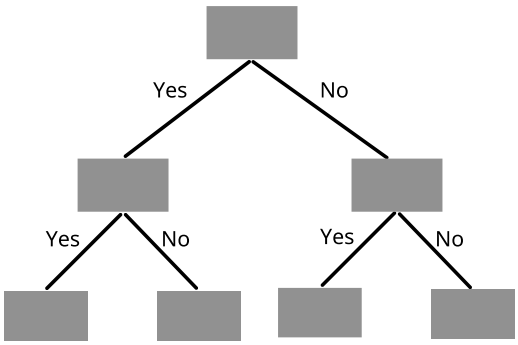
Leukaemias are characterized by strong transcriptomic signals, thus ML techniques and transcriptomic data are used over the last two decades in order to define leukaemia subtypes and find important gene signatures (Golub et al., 1999; Warnat-Herresthal et al., 2020). Specifically in AML diagnosis, some methods use gene expression profiling (GEP) data to predict if an individual has AML or is healthy (Angelakis & Soulioti, 2021; Roushangar & Mias, 2019) or exclude the possibility that an individual has AML (Warnat-Herresthal et al., 2020), while other approaches use microscopic data such as histopathology slides to identify AML and its subtypes (Kazemi et al., 2016).

4. Models

4.1. Decision trees (DT's)

DT's are models which represent logical rules. One new field where DT's are applied is microarray analysis. Their architecture compromises of nodes and branches. DT's are designed with the root on the top and the leaves at the bottom. The root node represents a choice that its outcome will split all the observations into at least two subgroups. Each internal node depicts a test, each branch represents the result of the test and each leaf node depicts the final outcome. When a data instance enters the root node, the algorithm determines which node the data instance will follow next. This procedure is iterative and ends when the observation arrives at a leaf node so the instance is labeled with its class label (Hormann, 1964; Myles et al., 2004; Norton, 1989; Song & Lu, 2015; Xie et al., 2003).

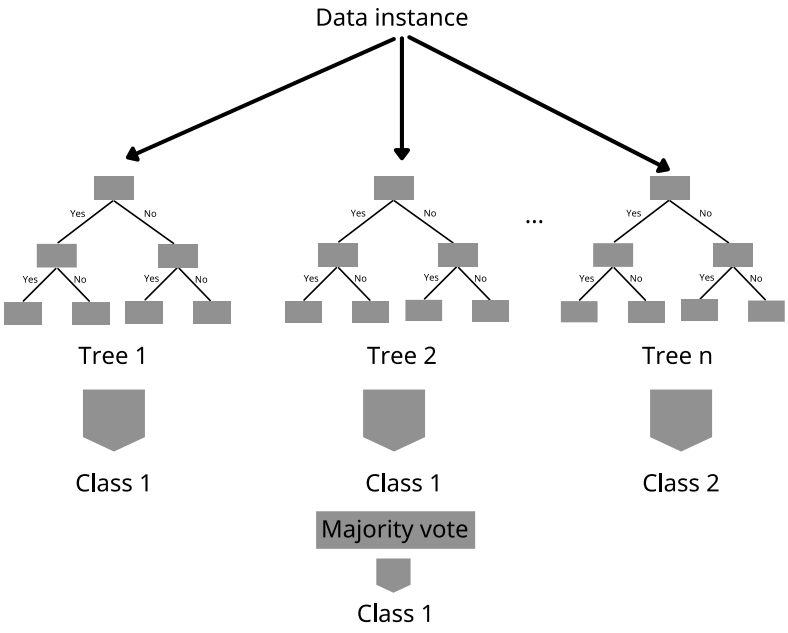
Figure 2. Representation of a DT.



4.2. Random forest (RF)

RF has wide applications in several fields such as face recognition, bioinformatics and medical image segmentation. This method is able to handle large feature space and works efficient on multi-class problems. It is also robust to overfitting and can deal with outliers. It is an ensemble approach which consists of a series of DT's, hence is called 'forest'. Each DT consists of a randomly selected subset of inputs, thus is called 'random'. Every DT votes for a class and at the end, each new data instance is classified to the class with the highest number of votes (Perner, 2012; Sachdeva & Kumar, 2021; Sarica et al., 2017). RF is also used for feature selection as it gives the features' importance (Kursa & Rudnicki, 2010).

Figure 3. Representation of RF.



4.3. Gradient boosted trees (GBT's)

GBT's is a state-of-the art Big Data analytics tool, due to its high performances as regards its efficiency, accuracy and interpretability in a variety of ML tasks, such as multi-class classification, regression, click prediction and learning to rank (Hancock & Khoshgoftaar, 2020; Ke et al., 2017). It has some implementations, including XGBoost, LightGBM and CatBoost (Anghel et al., 2018).

4.3.1. Extreme gradient boosting (XGBoost)

XGBoost is a scalable, open source package, tree ensembles ML tool for tree boosting. The main advantage of XGBoost in comparison with other gradient boosting models is the use of methods to control the overfitting problem. In addition, it can outperform other methods in terms of accuracy and F1-score, it needs less computational effort and it runs a lot faster than other algorithms in distributed or memory-limited settings. It is also a very effective technique when used on sparse data (Chen & Guestrin, 2016; Cui et al., 2016; Sheridan et al., 2016).

4.3.2. Light gradient boosting machine (LightGBM)

LightGBM includes most of the XGBoost's advantages but the main difference is that in LightGBM the decision trees are developed leaf-wise and not level-wise, which means row by row. In particular, is not required to examine the already existing leaves to create each new leaf (Al Daoud, 2019). It also utilizes Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) that accelerate the training process and maintain the accuracy (Ke et al., 2017).

4.3.3. Categorical gradient boosting (CatBoost)

CatBoost is a new kind of open source GBDT implementation that outperforms the existing state-of-the-art GBDT implementations. It differs from traditional GBDT algorithms for various reasons.

Firstly, it can handle efficiently noisy and heterogenous data. CatBoost can also solve problems with categorical features because it contains one-hot encoding process so it can convert categorical features to numbers at the preprocessing time or during training phase.

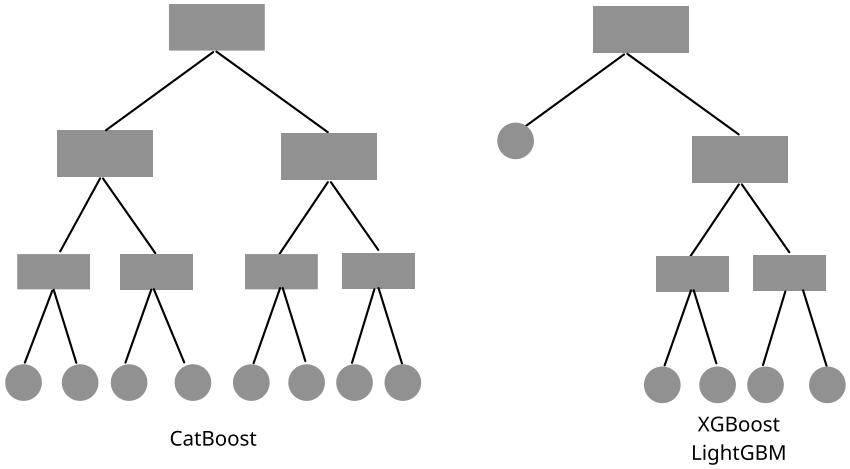
Moreover, it can overcome gradient bias, which causes prediction shift by using ordered boosting method. This approach improves the generalization ability of the model. In this method, CatBoost trains a model for each observation but the model is trained without that data instance. Then, the model is used to calculate the gradient estimation of the observation, which will be used to train the final model.

It controls the overfitting with two manners. It uses symmetric trees and the same features to split the instances, these trees are balanced and less prone to overfitting. It also uses random ordering of the training instances.

CatBoost provides a facile way to tune its hyper-parameters, including the depth and the learning rate of the trees, for achieving the highest performance.

It is an explainability approach since it makes use of Prediction Values Change (PVC) or Loss Function Change (LFC) to rank the developed model's features. Even though it is a black box algorithm, it can provide the impact of each feature so it can be used as feature selection approach (Dhananjay & Sivaraman, 2021; Hancock & Khoshgoftaar, 2020; G. Huang et al., 2019; S. Lee et al., 2021; Y. Zhang et al., 2020).

Figure 4. Representation of CatBoost, XGBoost and LightGBM.



4.4. Differences between RF and GBT's

RF and GBT's consist of DT's and they have two basic differences. RF constructs each DT independently while GBT's use boosting method. In boosting approach, a tree is developed at a time in order to improve by correcting the error of the previous ones. In addition, RF combine results by calculate the majority of votes but GBT's combine the results during the process (Downey, 2020).

4.5. *k*-Nearest-Neighbors (*k*NN)

The *k*NN is a non-parametric algorithm used in classification problems. For every data instance which needs to be classified, its *k* nearest neighbors are retrieved. Then, the majority of voting in the neighborhood is used to classify the data instance taking or not into consideration distance parameters. Very important in this method is to choose an appropriate value for *k* (Goos et al., 1999).

4.6. Least absolute shrinkage and selection operator (LASSO) regression

LASSO is a method used in regression models for feature selection and dimensionality reduction. It can minimize the prediction error by impose a constraint on the model parameters which decrease the regression coefficients to zero. Then, the features with a regression coefficient of zero are excluded (Ranstam & Cook, 2018).

4.7. Support vector machine (SVM)

SVM is a tool which consists of supervised learning methods used for classification and regression problems (Jakkula, 2011). Its decision function is a hyperplane that separates the samples which belongs to different classes (Salcedo-Sanz et al., 2014). The hyperplane is designed in such a way that the distance between the hyperplane and the closest data points of each class to be the maximum (S. Huang et al., 2018)

4.8. Deep neural networks (DNNs)

DNNs are powerful ML tools and are commonly used to solve large-scale real-word problems, including automated image classification, natural language processing and human action recognition tasks (Samek et al., 2017).They are artificial neural networks (ANNs) with multiple layers between the input and the output layers (Bengio, 2009).

4.9. Models' evaluation

Different metrics are used to evaluate if a model is accurate and valid.

Confusion matrix is a matrix in which the rows depict the actual class of the observations and the columns their predicted class. In binary classification, the confusion matrix is a 2 x 2 matrix.

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix}$$

The tp is the number of true positives, tn is the number of true negatives, fp represents the sum of false positive results and fn indicates the number of false negative ones.

Precision or Positive Predictive Value equals the number of true positive results divided by the total positive results and it shows the number of positives that are correctly identified by the model out of the total positive records.

$$precision = \frac{tp}{tp + fp}$$

Sensitivity or Recall is the ratio of the true positives predictions to true positives and false negative results and it presents the positive results that are correctly identified by the algorithm out of the actual positive ones.

$$recall = \frac{tp}{tp + fn}$$

Specificity is the ratio of the true negatives to the sum of true negatives and false positive predictions. Specificity indicates the ability of the algorithm to correctly identify negative results out of the real negative ones.

$$specificity = \frac{tn}{tn + fp}$$

Accuracy is the ratio of correctly classified samples (true positives and true negatives) to the total number of records and it measures the models' ability to make correct predictions.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

F1 score or F-measure is a metric who take into consideration the precision and the recall in order to measure the not correctly predictions.

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall}$$

The area under the curve (AUC) score is calculated using receiver operating characteristic curve (ROC), which plots sensitivity and 1-specificity. The value of AUC is the area under the

ROC curve, and it demonstrates the performance of the model for identifying positive and negative results. If the AUC > 0.9 the model has an excellent predictive ability.

PVC calculates the change in prediction when a value corresponding to the feature changes and LFC is used to rank a specific model to other models (Caelen, 2017; Dhananjay & Sivaraman, 2021; Muhsen et al., 2020; Sachdeva & Kumar, 2021; Vakili et al., 2020).

5. Data

The data used in this study are curated publicly available Affymetrix expression microarray one and consists of 34 datasets derived from 32 studies (Angelakis & Soulioti, 2021; Roushangar & Mias, 2019). It is an international multicentric dataset since it originated from 27 organizations, 25 cities, 15 countries and 4 continents. In addition, the data come from different microarray platforms: Affymetrix Human Genome U133 Plus 2.0 (GPL570), Affymetrix Human Genome U133A (GPL96) and Affymetrix Human Genome U133B (GPL97) and their sample source is either PB or BM. The dataset consists of 44754 probe sets which were common across all microarray platforms and 3374 arrays correspond to 3374 individuals. From 3374, 2668 (79.08%) are AML patients and 706 (20.92%) are healthy.

The final dataset consists of 2177 age-annotated individuals and 26 probe sets. From 2177, 1013 are female (46.53%), 943 are male (43.32%) and 221 (10.15%) are unknown-sex individuals. Moreover, 1629 (74.83%) subjects are labeled as AML and 548 (25.17%) are labeled as healthy. The mean and the standard deviation of age are 48.87 and 17.01 each. The final dataset was randomly divided into training and test set which are composed of 1740 (79.93%) and 437 (20.07%) subjects, respectively. Tables 3, 4 and 5 provide a detailed information about the entire dataset including the number of samples used, the sample source, the sex and the age of the individuals, the organizations from which the data originate, the total number of AML patients and healthy controls, the AML subtypes according to FAB or WHO classification system and statistics about the overall survival when available.

Table 3. GEO accesion, health status, Affymetrix platform, number of samples used and sample source and references of the train and validation set of 2177 individuals.

			Training set		Test set	
Author, Year	GEO accession	Disease state	Affymetrix platform id: Number of samples used & Sample source	Percentages	Affymetrix platform id: Number of samples used & Sample source	Percentages
Zatkova *et al.*, 2009	GSE10258	AML	GPL570: 6 BM	75%	GPL570: 2 BM	25%
Tomasson *et al.*, 2008, Walter *et al*., 2009	GSE10358	AML	GPL570: 245 BM	81.67%	GPL570: 55 BM	18.33%
Warren *et al*., 2009	GSE11375	Healthy	GPL570: 22 PB	84.62%	GPL570: 4 PB	15.38%
Metzeler *et al.*, 2008 Wang *et al*., 2021	GSE12417	AML	GPL570: 56 (52 BM & 4 PB)	71.8%	GPL570: 22 (21 BM & 1 PB)	28.2%
Wouters *et al.*, 2009, Taskesen *et al.*, 2011, Taskesen *et al.,* 2015	GSE14468	AML	GPL570: 412 (379 BM & 33 PB)	78.48%	GPL570: 113 (103 BM & 10 PB)	21.52%
Figueroa *et al.*, 2009	GSE14479	AML	GPL570: 14 BM	87.5%	GPL570: 2 BM	12.5%
Klein *et al.*, 2009	GSE15434	AML	GPL570: 194 (177 BM & 17 PB)	77.29%	GPL570: 57 (54 BM & 3 PB)	22.71%
Wu *et al*., 2012 (NPY)	GSE15932	Healthy	GPL570: 6 PB	75%	GPL570: 2 PB	25%
Karlovich *et al.*, 2009	GSE16028	Healthy	GPL570: 18 PB	81.82%	GPL570: 4 PB	18.18%
Krug *et al.*, 2011 (NPY)	GSE17114	Healthy	GPL570: 11 PB	78.57%	GPL570: 3 PB	21.43%
Kong *et al.*, 2012	GSE18123	Healthy	GPL570: 13 PB	76.47%	GPL570: 4 PB	23.53%
Sharma *et al.*, 2009	GSE18781	Healthy	GPL570: 20 PB	80%	GPL570: 5 PB	20%
Zhou *et al.*, 2010	GSE19743	Healthy	GPL570: 50 PB	79.37%	GPL570: 13 PB	20.63%
Li *et al.*, 2011	GSE23025	AML	GPL570: 22 (12 BM & 10 PB)	64.71%	GPL570: 12 (9 BM & 3 PB)	35.29%
Rosell *et al.*, 2011	GSE25414	Healthy	GPL570: 11 PB	91.67%	GPL570: 1 PB	8.33%
Schmidt *et al.*, 2006	GSE2842	Healthy	GPL570: 1 PB	50%	GPL570: 1 PB	50%
Lück *et al.*, 2011	GSE29883	AML	GPL570: 11 (9 BM & 2 PB)	91.67%	GPL570: 1 BM	8.33%
Xiao *et al.*, 2011	GSE36809	Healthy	GPL570: 28 PB	80%	GPL570: 7 PB	20%
Li *et al.,* 2013, Herold *et al*., 2014, Kuett *et al*., 2015, Herold *et al*., 2018	GSE37642	AML	GPL570: 120 BM	85.71%	GPL570: 20 BM	14.29%
Lauwerys *et al*., 2013 Ducreux *et al*., 2016	GSE39088	Healthy	GPL570: 37 PB	80.43%	GPL570: 9 PB	19.57%
Bullinger *et al.*, 2014 (NPY)	GSE39363	AML	GPL570: 12 (10 BM & 2 PB)	92.3%	GPL570: 1 BM	7.7%
Clelland *et al.*, 2013	GSE46449	Healthy	GPL570: 19 PB	79.17%	GPL570: 5 PB	20.83%
Opel *et al.*, 2015, Lueck *et al*., 2016	GSE46819	AML	GPL570: 9 (6 BM & 3 PB)	75%	GPL570: 3 (2 BM & 1 PB)	25%
Leong *et al.*, 2015 (NPY)	GSE68833	AML	GPL570: 148 BM	80.87%	GPL570: 35 BM	19.13%

Cao *et al.*, 2016	GSE69565	AML	GPL570: 12 PB	100%	-	0%
Meng *et al.*, 2015 (NPY)	GSE71226	Healthy	GPL570: 3 PB	100%	-	0%
Bohl *et al.*, 2016 (NPY)	GSE84334	AML	GPL570: 42 (23 BM & 19 PB)	93.33%	GPL570: 3 (2 BM & 1 PB)	6.67%
Tasaki *et al.,* 2017	GSE84844	Healthy	GPL570: 26 PB	86.67%	GPL570: 4 PB	13.33%
Tasaki *et al.,* 2018	GSE93272	Healthy	GPL570: 25 PB	71.43%	GPL570: 10 PB	28.57%
Leday et al., 2018	GSE98793	Healthy	GPL570: 49 PB	76.56%	GPL570: 15 PB	23.44%
Shamir *et al.,* 2017	GSE99039	Healthy	GPL570: 99 PB	81.82%	GPL570: 22 PB	18.18%
Green *et al.,* 2009 (NPY)	GSE14845	Healthy	-	0%	GPL570: 1 PB	100%
Lück *et al.*, 2011	GSE29883	AML	GPL570: 11 (9 BM & 2 PB)	91.67%	GPL570: 1 BM	8.33%
Xiao *et al.*, 2011	GSE36809	Healthy	GPL570: 28 PB	80%	GPL570: 7 PB	20%
Li *et al.,* 2013, Herold *et al*., 2014, Kuett *et al*., 2015, Herold *et al*., 2018	GSE37642	AML	GPL570: 120 BM	85.71%	GPL570: 20 BM	14.29%
Lauwerys *et al*., 2013 Ducreux *et al*., 2016	GSE39088	Healthy	GPL570: 37 PB	80.43%	GPL570: 9 PB	19.57%
Bullinger *et al.*, 2014 (NPY)	GSE39363	AML	GPL570: 12 (10 BM & 2 PB)	92.3%	GPL570: 1 BM	7.7%
Clelland *et al.*, 2013	GSE46449	Healthy	GPL570: 19 PB	79.17%	GPL570: 5 PB	20.83%
Opel *et al.*, 2015, Lueck *et al*., 2016	GSE46819	AML	GPL570: 9 (6 BM & 3 PB)	75%	GPL570: 3 (2 BM & 1 PB)	25%
Leong *et al.*, 2015 (NPY)	GSE68833	AML	GPL570: 148 BM	80.87%	GPL570: 35 BM	19.13%
Cao *et al.*, 2016	GSE69565	AML	GPL570: 12 PB	100%	-	0%
Meng *et al.*, 2015 (NPY)	GSE71226	Healthy	GPL570: 3 PB	100%	-	0%
Bohl *et al.*, 2016 (NPY)	GSE84334	AML	GPL570: 42 (23 BM & 19 PB)	93.33%	GPL570: 3 (2 BM & 1 PB)	6.67%
Tasaki *et al.,* 2017	GSE84844	Healthy	GPL570: 26 PB	86.67%	GPL570: 4 PB	13.33%
Tasaki *et al.,* 2018	GSE93272	Healthy	GPL570: 25 PB	71.43%	GPL570: 10 PB	28,57%
Leday et al., 2018	GSE98793	Healthy	GPL570: 49 PB	76.56%	GPL570: 15 PB	23.44%
Shamir *et al.,* 2017	GSE99039	Healthy	GPL570: 99 PB	81.82%	GPL570: 22 PB	18.18%
Green *et al.,* 2009 (NPY)	GSE14845	Healthy	-	0%	GPL570: 1 PB	100%
Opel *et al.*, 2015, Lueck *et al*., 2016	GSE46819	AML	GPL570: 9 (6 BM & 3 PB)	75%	GPL570: 3 (2 BM & 1 PB)	25%
Leong *et al.*, 2015 (NPY)	GSE68833	AML	GPL570: 148 BM	80.87%	GPL570: 35 BM	19.13%
Cao *et al.*, 2016	GSE69565	AML	GPL570: 12 PB	100%	-	0%
Meng *et al.*, 2015 (NPY)	GSE71226	Healthy	GPL570: 3 PB	100%	-	0%
Bohl *et al.*, 2016 (NPY)	GSE84334	AML	GPL570: 42 (23 BM & 19 PB)	93.33%	GPL570: 3 (2 BM & 1 PB)	6.67%
Tasaki *et al.,* 2017	GSE84844	Healthy	GPL570: 26 PB	86.67%	GPL570: 4 PB	13.33%
Tasaki *et al.,* 2018	GSE93272	Healthy	GPL570: 25 PB	71.43%	GPL570: 10 PB	28.57%
Leday et al., 2018	GSE98793	Healthy	GPL570: 49 PB	76.56%	GPL570: 15 PB	23.44%
Shamir *et al.,* 2017	GSE99039	Healthy	GPL570: 99 PB	81.82%	GPL570: 22 PB	18.18%
Green *et al.,* 2009 (NPY)	GSE14845	Healthy	-	0%	GPL570: 1 PB	100%
Leong *et al.*, 2015 (NPY)	GSE68833	AML	GPL570: 148 BM	80.87%	GPL570: 35 BM	19.13%

Cao *et al.*, 2016	GSE69565	AML	GPL570: 12 PB	100%	-	0%

*NPY; not published yet

Table 4. GEO accession, origin of study, AML subtypes and overall survival.

GEO accession	City, Country, Organization	AML subtypes (FAB or WHO classification)	Overall survival (days)
GSE10258	Vienna, Austria, Medical University of Vienna	M1, M5	-
GSE10358	St Louis, USA, Washington University School of Medicine	M0, M1, M2, M3, M4, M5, M6, M7	-
GSE11375	Boston, USA, Massachusetts General Hospital	-	-
GSE12417	Munich, Germany, University of Munich	M0, M1, M2, M4, M5, M6	Mean: 614,76 Std: 503,59
GSE14468	Houston, USA, MD Anderson Cancer Center	M0, M1, M2, M3, M4, M4 eos, M5, M6	-
GSE14479	Rotterdam, Netherlands, Erasmus University Medical Center	M0, M1	-
GSE15434	New York, USA, Columbia University Medical Center	-	-
GSE15932	Hangzhou, China, Second Affiliated Hospital, School of Medicine, Zhejiang University	-	-
GSE16028	Basel, Switzerland, F.Hoffmann-La Roche AG	-	-
GSE17114	Lisbon, Portugal, Instituto de Medicina Molecular	-	-
GSE18123	Boston, USA, Boston Children's Hospital	-	-
GSE18781	Portland, USA, Oregon Health & Science University	-	-

GSE19743	Palo Alto, USA, Stanford Genome Technology Center	-	-
GSE23025	Duarte, USA, City of Hope Beckman Research Institute	-	-
GSE25414	Barcelona, Spain, Institut de Recerca Hospital Vall d'Hebron	-	-
GSE2842	Bolzano, Italy, EURAC	-	-
GSE29883	Berlin, Germany, Charité	t(8;21), t(16;16)	-
GSE36809	Boston, USA, Massachusetts General Hospital	-	-
GSE37642	Munich, Germany, University Hospital Grosshadern, Ludwig-Maximilians-University (LMU)	M0, M1, M2, M3, M4, M5, M6, M7	Mean: 962,32 Std: 1106,70
GSE39088	Brussels, Belgium, Université catholique de Louvain	-	-
GSE39363	Berlin, Germany, Charité	t(3;3)	-
GSE46449	New York, USA, Columbia University Medical Center	-	-
GSE46819	Berlin, Germany, Charité	t(16;16)	-
GSE68833	Rockville, USA, NCI	M0, M1, M2, M3, M4, M5, M6, M7	-
GSE69565	Singapore, Singapore, Cancer Science Institute of Singapore	-	-
GSE71226	Changchun, China, the Department of Cardiology, China–Japan Union Hospital, Jilin University	-	-
GSE84334	Ulm, Germany, University Hospital of Ulm	-	-
GSE84844	Fujisawa, Japan, Takeda Pharmaceutical Company Limited	-	-

GSE93272	Fujisawa, Japan, Takeda Pharmaceutical Company Limited	-	-
GSE98793	Cambridge, United Kingdom, University of Cambridge	-	-
GSE99039	Tel Aviv, Israel, Tel Aviv University	-	-
GSE14845	Southport, Australia, Griffith Insitute for Health & Medical Research	-	-

Table 5. Summary statistics of the total dataset of 2177 individuals and of the training and test set, including disease state, sex, number of patients per age group, mean and standard deviation of age.

Summary statistics		Training set	Percentage	Test set	Percentage
Total dataset					
Disease state					
AML	Healthy				
1629	548	1303 AML & 438 Healthy	79.97%	326 AML & 110 Healthy	20.03%
Percentage					
74,83% AML	25,17% Healthy				
Sex					
Female	Male				
1013	943	807 Female & 755 Male	-	206 Female & 188 Male	-
Age					
Age group: Number of patients	Percentage	Age group: Number of patients		Age group: Number of patients	
0 to 19: 99	4.55%	0 to 19: 75	4.31%	0 to 19: 24	5.5%
20 to 29: 217	9.97%	20 to 29: 180	10.34%	20 to 29: 37	8.49%
30 to 39: 340	15.62%	30 to 39: 272	15.62%	30 to 39: 68	15.6%
40 to 49: 393	18.05%	40 to 49: 313	17.98%	40 to 49: 80	18.35%

50 to 59: 487	22.37%	50 to 59: 378	21.71%	50 to 59: 109	25%
60 to 69: 390	17.91%	60 to 69: 319	18.32%	60 to 69: 71	16.28%
70 to 79: 212	9.74%	70 to 79: 171	9.82%	70 to 79: 41	9.4%
80 to 89: 39	1.79%	80 to 100: 33	1.9%	80 to 100: 6	1.38%
Mean of age: 48.87		Mean of age: 48.98		Mean of age: 48.46	
Std of age: 17.01		Std of age: 17.06		Std of age: 16.79	

*Std; standard deviation

6. Methods

Firstly, the dataset of 3374 individuals and 44754 probe sets is randomly split into two sets, the training and the validation set, which consist of the 80% and the 20% of the total dataset, respectively.

Dimensionality reduction CatBoost model has 200 iterators, depth 6 and learning rate 0.1 is applied on the dataset mentioned above in order to obtain the set of 100 most important features regarding the PVC of CatBoost and the set of 100 most important attributes regarding its LFC. The above two sets may differ. The intersection of these two sets consists of 34 probe sets. Ten fold cross validation (10CV) is used in order to tune the dimensionality reduction CatBoost model on the training set and then validate it on the test set.

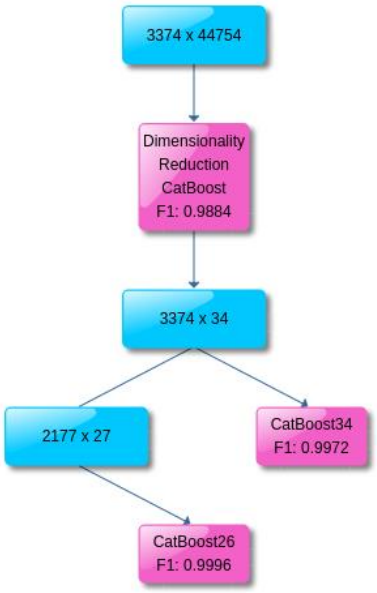
Later, a CatBoost model (CatBoost34) of 200 iterators, depth 5, learning rate 0.1 and 10CV is tuned on the dataset of 3374 data instances and 34 probe sets, which come from the intersection mentioned before.

The 8 probe sets of the 34 correspond to genes which are correlated from bibliographic references to AML. These 8 genes are excluded. In addition, individuals with no age filled-in are dropped-out from the dataset of 3374. A CatBoost model (CatBoost26) of 100 iterators, depth 11, learning rate 0.1 and 10CV is implemented on the final dataset, which consists of 2177 individuals, the 26 probe sets and the age as features.

In all three models above, weight balance parameters are used, and all the other parameters have their default values.

All the above approaches were implemented in Python 3.8 language, using Jupyter Notebooks, NumPy, Pandas, Scikit-learn and CatBoost 0.24.1 libraries.

Figure 5. Representation of the methodology, including datasets and CatBoost models as well as the F1 score. The first integer corresponds to the number of the data instances and the second corresponds to the number of attributes.



7. Results

Figure 6. Features' importance of the PVC of CatBoost model.

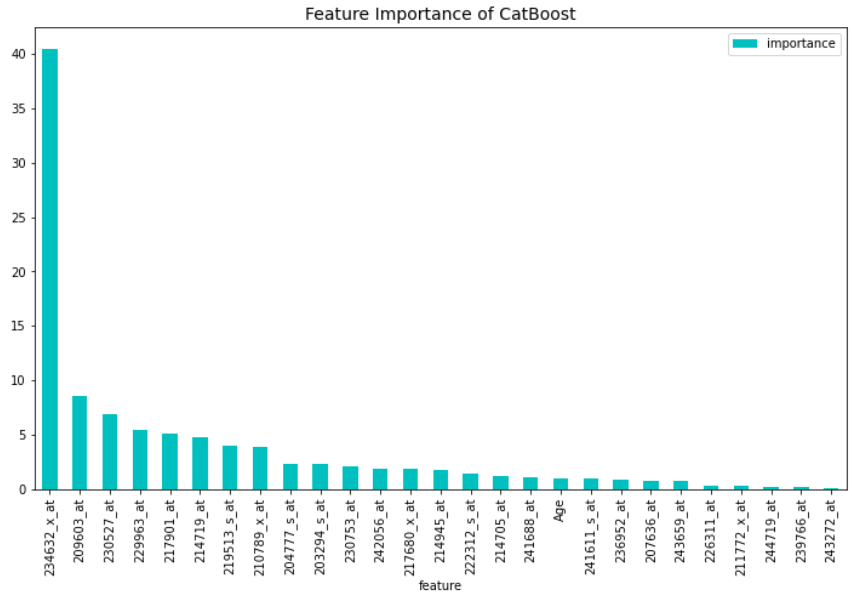
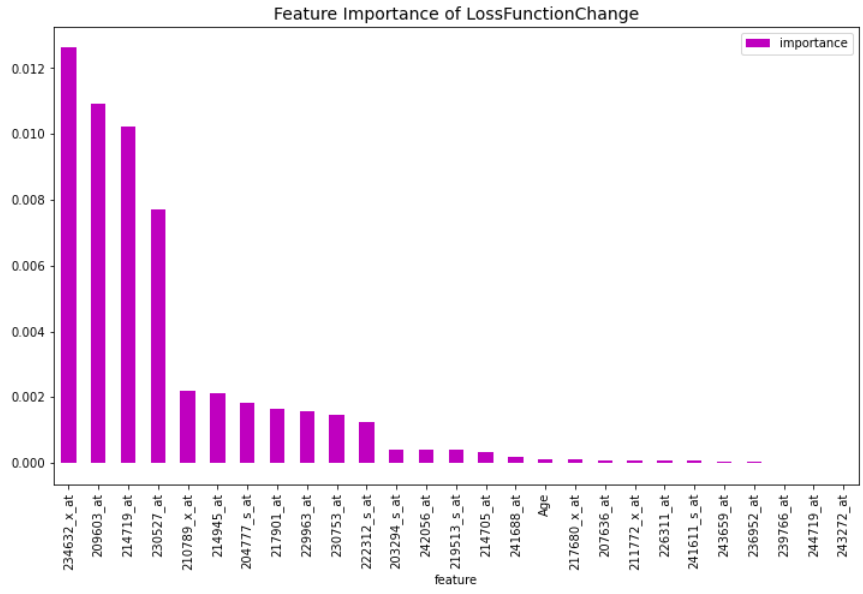


Figure 7. Features' importance of the LFC of CatBoost model.



Figures 6,7 show the 27 features (26 probe sets and the age) ranked by their importance regarding CatBoost26 PVC and LFC, respectively. The 26 probe sets correspond to 19 named genes, 4 uncharacterized genes, 6 EST's and 1 cDNA clone. It is obvious that probe set *234632_x_at*, which is a cDNA clone, with no annotations available, is the most important feature as regards both, the predictability of the CatBoost26 and its LFC.

From Figure 6, *234632_x_at* has more than 4 times higher feature importance regarding the PVC when compared to the other features. The next 7 features with high importance correspond to 6 named genes {GATA3, BEX5, DSG2, SLC46A3, SH2D3A, CEACAM3} and 1 uncharacterized gene {LOC101926907}.

In Figure 7, it is obvious that *234632_x_at* has at least 4 times higher feature importance than the 210789_x_at, while *230527_at* is approximately 3 times more important feature than *210789_x_at*. The 11 following most important features in Figure 7, correspond to 10 named genes {GATA3, BEX5, DSG2, SLC46A3, FAM153A, FAM153B, FAM153C, PATL2, CEACAM3, MAL}, 3 uncharacterized genes {LOC101926907, LOC100507387, LOC105377751} and 1 EST.

Table 6. The 26 probes in descending order according to their feature importance regarding the predictability of Catboost26. Information about probe set's identifier, corresponding gene symbols or NCBI accession numbers, blood malignancies and/or other types of cancer they are associated with, are presented here.

Probe set ID	Gene symbol/NCBI accession number	Blood malignancies	Other types of cancer
234632_x_at	AK026267*	-	-
209603_at	GATA3	acute lymphoblastic leukaemia (ALL) (Hou et al., 2017)	breast cancer (Mehra et al., 2005), bladder cancer (Li et al., 2014)
230527_at	LOC101926907	-	-
229963_at	BEX5	-	-
217901_at	DSG2	-	cervical cancer (Qin et al., 2020), epithelial-derived carcinomas (Brennan & Mahoney, 2009), pancreatic cancer (Hütz et al., 2017), breast cancer (Davies et al., 1997), colon cancer (T. Yang et al., 2021), lung cancer (Cai et al., 2017; Saaber et al., 2015), gastric cancer (Biedermann et al., 2005; Yashiro et al., 2006), ovarian cancer (Kim et al., 2020), laryngeal cancer (Cury et al., 2020), liver cancer (Han et al., 2018)
214719_at	SLC46A3	-	liver cancer (Zhao et al., 2019)
219513_s_at	SH2D3A	-	-
210789_x_at	CEACAM3	-	-
204777_s_at	MAL	-	gastric cancer (Buffart et al., 2008), breast cancer (Horne et al., 2009),

			ovarian cancer (P. S. Lee et al., 2010), colorectal cancer (Kalmár et al., 2015)
203294_s_at	LMAN1	-	-
230753_at	PATL2	-	-
242056_at	TRIM45	-	lung cancer (Peng et al., 2019), glioma (J. Zhang et al., 2017)
217680_x_at	RPL10	T-cell acute lymphoblastic leukaemia (T-ALL) (De Keersmaecker et al., 2013; Raiser et al., 2014)	ovarian cancer (Shi et al., 2018), pancreatic cancer (J. Yang et al., 2018)
214945_at	FAM153A & FAM153B & FAM153C & LOC100507387 & LOC105377751	-	-
222312_s_at	AW969803*	-	-
214705_at	PATJ	-	-
241688_at	AA677700*	-	-
241611_s_at	FNDC3A	multiple myeloma (Manfrini et al., 2020)	-
236952_at	AI309861*	-	-
207636_at	SERPINI2	chronic lymphocytic leukaemia (CLL) (Farfsing et al., 2009)	pancreatic cancer (Ozaki et al., 1998)
243659_at	N63876*	-	-
226311_at	ADAMTS2	mixed phenotype acute leukaemias (MPAL) (Tota et al., 2014)	gastric cancer (C. Jiang et al., 2019), kidney cancer (Roemer et al., 2004)
211772_x_at	CHRNA3	T-cell acute lymphoblastic leukaemia (T-ALL) (Laukkanen et al., 2015)	lung cancer (Wassenaar et al., 2011)
244719_at	AA766704*	-	-
239766_at	BF507518*	-	-
243272_at	LOC101593348	-	-

Table 6 provides information about the 26 most important probe sets. GATA3 has been implicated to acute lymphoblastic leukaemia and other types of cancer as well as breast cancer and bladder cancer. The gene DSG2 plays a role in various kinds of cancer including

cervical cancer, epithelial-derived carcinomas, pancreatic cancer, breast cancer, colon cancer, lung cancer, gastric cancer, ovarian cancer, laryngeal cancer and liver cancer. In addition, SLC46A3 is correlated to liver cancer, while the gene MAL has been correlated to gastric cancer, breast cancer, ovarian cancer and colorectal cancer. ADAMTS2 is implicated in MPAL as well as in gastric cancer and kidney cancer. Additionally, CHRNA3 plays a role in T-ALL and in lung cancer. The genes PATL2, FAM153A, FAM153B, FAM153C, BEX5, SH2D3A, CEACAM3 have not been related to any type of cancer yet.

Table 7. Confusion matrix of CatBoost26 on the training set.

$$\begin{bmatrix} 1302 & 1 \\ 0 & 437 \end{bmatrix}$$

Table 7 shows the confusion matrix of CatBoost26 model. The true positive predicts are up-left, the true negatives down-right, the false positives down-left and the false negative ones are up-right. A positive data instance corresponds to an AML patient and a negative one corresponds to a healthy individual.

Table 8. Performances of dimensionality reduction CatBoost model of the 10CV on 3374 and 44754 probe sets, CatBoost34 model of the 10CV on 3374 and 34 probe sets and CatBoost26 model of the 10CV on 2177 and 26 probe sets and the age.

Metrics	Validation set	10CV
CatBoost		
Specificity	0.9929	0.9805
Sensitivity	1.0000	0.9991
AUC	0.9965	0.9898
F1-score	0.9964	0.9884
CatBoost34		
Specificity	1.0000	0.9929
Sensitivity	1.0000	0.9926
AUC	1.0000	0.9920
F1-score	1.0000	0.9972

CatBoost26		
Specificity	1.0000	1.0000
Sensitivity	1.0000	0.9992
AUC	1.0000	0.9988
F1-score	1.0000	0.9996

From Table 8 is evident that the AML diagnosis model CatBoost26 performs very well. The AUC from the 10CV is 0.9988 with standard deviation 0.0023 and 95% confidence interval [0.9994, 1.000]. In addition, the mean accuracy is 0.994 with standard deviation 0.0011.

8. Discussion

In this study, CatBoost was not only used for the prediction of AML but also for feature selection.

CatBoost34 model is microarray platform agnostic. It has not been used the information if a data instance comes from Affymetrix Human Genome U133 Plus 2.0, Affymetrix Human Genome U133A or Affymetrix Human Genome U133B microarray platform. This enhances the robustness of the model and makes it generally applicable.

From bibliographic references, the 26 probe sets that CatBoost26 uses as features, are not correlated to AML yet. The age was used for two reasons. Firstly, its prognostic value is high as regards the overall survival of AML patients (Mosquera Orgueira et al., 2021). Secondly, studies in the field of deep learning with small datasets of 100 data instances derived from ultrasound indicate that a CatBoost model which uses age and features that come from a variety of sources can achieve high performance in binary classification tasks (Angelakis et al., 2018; Angelakis et al., 2021; Angelakis, 2021).

The accuracy of the CatBoost26 model is 99.94% and the F1-score is 0.9996. The performance of the model is the best one in the literature as regards the diagnosis of AML using transcriptomic data.

Similar study was carried out using as data the 3374 individuals and the 44754 probe sets as features (Roushangar & Mias, 2019). Statistical techniques were used to reduce the dimensionality of the dataset to 984 probe sets. In order to compare the results of this study to the current work, the same 80% training set that was used to train the dimensionality reduction CatBoost, was also used to train the k-NN model of the study of Roushangar & Mias, 2019. Dimensionality reduction CatBoost, CatBoost34 and CatBoost26 outperform *k*-NN as Table 9 shows.

Table 9. Performance of the *k*-NN of Roushangar & Mias, 2019 of the 10CV on 80% training set and on the 20% validation set of 3374 individuals and 984 probe sets.

Metrics	Validation set	10CV
Specificity	0.9716	0.9546
Sensitivity	0.9925	0.9920
AUC	0.9821	0.9788
F1-score	0.9925	0.9899

In another study, with data that originated from similar platforms (Affymetrix Human Genome U133A, Affymetrix Human Genome U133 plus 2.0 and Illumina RNA-seq) and different machine learning and statistical methods (kNN, LASSO, linear discriminant analysis, random forest, linear SVM, polynomial SVM, radial SVM, sigmoid SVM) have been used to solve the problem of AML diagnosis (Warnat-Herresthal et al., 2020). In this study, the individuals were separated into three main classes: AML, leukaemia and healthy or other diseases. The healthy or other diseases group contains healthy individuals and individuals that may have other diseases but not any type of leukaemia. The best results of the study of Warnat-Herresthal et al. regarding the accuracy are 97.6% when LASSO has been trained on Affymetrix Human Genome U133A dataset of 1049 AML and 1451 non AML, 98.0% when has been trained on Affymetrix Human Genome U133 plus 2.0, in which 2588 were labeled as AML and 5760 as non AML individuals and 99.1% when Illumina RNA-seq dataset of 508 AML and 673 non AML ones was used to train LASSO algorithm.

The last work which tried to solve the problem of AML diagnosis using machine learning and microarrays data of AML patients and healthy subjects, uses a DNN. The dataset consisted of 36 data instances and the accuracy of the model was 96.67% (Nazari et al., 2020).

9. Conclusion

This research aimed to assist in developing a diagnostic tool for identifying either if an individual has AML or is healthy. For the first time in the literature, a gradient boosted tree algorithm CatBoost, which uses probe sets and the age as features achieves the highest performance as regards the diagnosis of AML. CatBoost34 and CatBoost26 outperform other machine learning methods which use similar or different datasets.

In addition, it would be crucial the scientific community to further investigate the 26 probe sets shown in Table 6. Firstly, to identify their relation to known genes, as well as their role not only in AML but also in other types of cancer.

AML can appear suddenly without early detectable symptoms (Mottal et al., 2020). A screening tool where its performance, as regards the sensitivity and specificity is close to 1.00, and the sample could be PB is of high importance because it can contribute to prevent human errors during PB and BM examinations and can facilitate the quick diagnosis of AML so the patients could undergo treatment as soon as possible. In addition, the option of getting screened often will help not only in the diagnosis but also in its prevention since the treatment is more effective when started early.

Bibliography

American Cancer Society (2021). Tests for Acute Myeloid Leukemia (AML). Retrieved 10/09/2021, from http://www.cancer.org

Agaian, S., Madhukar, M., & Chronopoulos, A. T. (2014). Automated screening system for acute myelogenous leukemia detection in blood microscopic images. *IEEE Systems Journal*, *8*(3), 995–1004. https://doi.org/10.1109/JSYST.2014.2308452

Al Daoud, E. (2019). Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset. *International Journal of Computer and Information Engineering*, *13*(1), 6–10.

Angelakis, A., & Soulioti, I. (2021). Diagnosis of Acute Myeloid Leukaemia Using Machine Learning. *Myeloid Leukemia - Clinical Diagnosis and Treatment*. https://doi.org/10.5772/26177

Angelakis, A., Gatos, I., Theotokas, I. et al. (2018). A deep-learning approach to the significant liver fibrosis binary classification problem using gender, morphologic and haemodynamic measurements derived from B-mode ultrasound images, Insights Imaging. p.S279,9(Suppl 1):1.doi:10.1007/s13244-018- 0603-8

Angelakis, A., Gatos, I., Theotokas, I. et al. (2018). Binary Classification of Chronic Liver Disease Patients Using Deep Learning on Morphologic B-Mode and Demographic Data, Journal of Ultrasound in Medicine. AIUM 2018 Annual Convention,S3,doi:10.1002/jum.14750

Angelakis, A. (2021) Cats On The Classification Of Benign And Malignant Breast Lesions Using Ultrasound Shear Wave Elastography Features And BI-RADS Score, Journal of Ultrasound in Medicine. AIUM 2021 Annual Convention,doi:10.1002/jum.15752

Anghel, A., Papandreou, N., Parnell, T., De Palma, A., & Pozidis, H. (2018). *Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms*. http://arxiv.org/abs/1809.04559

Bahakeem, E., & Qadah, T. (2020). *Current Diagnostic Methods for Hematological Malignancies: A Mini-Review - Pharmacophore*. *11*(3), 63–68. https://pharmacophorejournal.com/en/article/current-diagnostic-methods-for-hematological-malignancies-a-mini-review

Baumann, T., Delgado, J., & Montserrat, E. (2020). CLL and COVID-19 at the Hospital Clinic of Barcelona: an interim report. *Leukemia*, *34*(7), 1954–1956. https://doi.org/10.1038/s41375-020-0870-5

Bengio, Y. (2009). Learning deep architectures for AI. In *Foundations and Trends in Machine Learning* (Vol. 2, Issue 1). https://doi.org/10.1561/2200000006

Biedermann, K., Vogelsang, H., Becker, I., Plaschke, S., Siewert, J. R., Höfler, H., & Keller, G. (2005). Desmoglein 2 is expressed abnormally rather than mutated in familial and sporadic gastric cancer. *Journal of Pathology*, *207*(2), 199–206. https://doi.org/10.1002/path.1821

Brennan, D., & Mahoney, M. G. (2009). Increased expression of Dsg2 in malignant skin carcinomas: A tissue-microarray based study. *Cell Adhesion and Migration*, *3*(2), 148–154. https://doi.org/10.4161/cam.3.2.7539

Buffart, T. E., Overmeer, R. M., Steenbergen, R. D. M., Tijssen, M., Van Grieken, N. C. T., Snijders, P. J. F., Grabsch, H. I., Van De Velde, C. J. H., Carvalho, B., & Meijer, G. A. (2008). MAL promoter hypermethylation as a novel prognostic marker in gastric

cancer. *British Journal of Cancer*, *99*(11), 1802–1807. https://doi.org/10.1038/sj.bjc.6604777

Burke, V. P., & Startzell, J. M. (2008). The Leukemias. *Oral and Maxillofacial Surgery Clinics of North America*, *20*(4), 597–608. https://doi.org/10.1016/j.coms.2008.06.011

Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, *81*(3–4), 429–450. https://doi.org/10.1007/s10472-017-9564-8

Cai, F., Zhu, Q., Miao, Y., Shen, S., Su, X., & Shi, Y. (2017). Desmoglein-2 is overexpressed in non-small cell lung cancer tissues and its knockdown suppresses NSCLC growth by regulation of p27 and CDK2. *Journal of Cancer Research and Clinical Oncology*, *143*(1), 59–69. https://doi.org/10.1007/s00432-016-2250-0

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cui, B., Zhang, N., Xu, J., Lian, X., Liu, D., & Hutchison, D. (2016). *Web-Age Information Management*.

Cury, S. S., Lapa, R. M. L., de Mello, J. B. H., Marchi, F. A., Domingues, M. A. C., Pinto, C. A. L., Carvalho, R. F., de Carvalho, G. B., Kowalski, L. P., & Rogatto, S. R. (2020). Increased DSG2 plasmatic levels identified by transcriptomic-based secretome analysis is a potential prognostic biomarker in laryngeal carcinoma. *Oral Oncology*, *103*(January), 104592. https://doi.org/10.1016/j.oraloncology.2020.104592

Davies, E. L., Cochrane, R. A., Hiscox, S., Jiang, W. G., Sweetland, H. M., & Mansel, R. E. (1997). The role of desmoglein 2 and E-cadherin in the invasion and motility of human breast cancer cells. *International Journal of Oncology*, *11*(2), 415–419. https://doi.org/10.3892/ijo.11.2.415

De Keersmaecker, K., Atak, Z. K., Li, N., Vicente, C., Patchett, S., Girardi, T., Gianfelici, V., Geerdens, E., Clappier, E., Porcu, M., Lahortiga, I., Lucà, R., Yan, J., Hulselmans, G., Vranckx, H., Vandepoel, R., Sweron, B., Jacobs, K., Mentens, N., … Cools, J. (2013). Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nature Genetics*, *45*(2), 186–190. https://doi.org/10.1038/ng.2508

De Kouchkovsky, I., & Abdul-Hay, M. (2016). 'Acute myeloid leukemia: A comprehensive review and 2016 update.' *Blood Cancer Journal*, *6*(7). https://doi.org/10.1038/bcj.2016.50

Dhananjay, B., & Sivaraman, J. (2021). Analysis and classification of heart rate using CatBoost feature ranking model. *Biomedical Signal Processing and Control*, *68*(December 2020), 102610. https://doi.org/10.1016/j.bspc.2021.102610

Döhner, H., Weisdorf, D. J., & Bloomfield, C. D. (2015). Acute Myeloid Leukemia. *New England Journal of Medicine*, *373*(12), 1136–1152. https://doi.org/10.1056/NEJMra1406184

Downey, S. (2020). Gazing into the Future: Using Ensemble Techniques to Forecast Company Fundamentals. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3580018

Farfsing, A., Engel, F., Seiffert, M., Hartmann, E., Ott, G., Rosenwald, A., Stilgenbauer, S., Döhner, H., Boutros, M., Lichter, P., & Pscherer, A. (2009). Gene knockdown studies revealed CCDC50 as a candidate gene in mantle cell lymphoma and chronic lymphocytic leukemia. *Leukemia*, *23*(11), 2018–2026. https://doi.org/10.1038/leu.2009.144

Gal, O., Auslander, N., Fan, Y., & Meerzaman, D. (2019). Predicting Complete Remission of Acute Myeloid Leukemia: Machine Learning Applied to Gene Expression. *Cancer Informatics*, *18*. https://doi.org/10.1177/1176935119835544

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gassenbeck, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., & others. (1999). Molecular Classification of Cancer: Class Discovery. *Science*, *286*(October), 531–537.

Goos, G., Hartmanis, J., & Leeuwen, J. (1999). Lecture Notes in Computer Science. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 1716).

Goutam, D., & Sailaja, S. (2015). Classification of acute myelogenous leukemia in blood microscopic images using supervised classifier. *ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology*, *March*, 3–7. https://doi.org/10.1109/ICETECH.2015.7275021

Han, C. P., Yu, Y. H., Wang, A. G., Tian, Y., Zhang, H. T., Zheng, Z. M., & Liu, Y. S. (2018). Desmoglein-2 overexpression predicts poor prognosis in hepatocellular carcinoma patients. *European Review for Medical and Pharmacological Sciences*, *22*(17), 5481–5489. https://doi.org/10.26355/eurrev_201809_15808

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00369-8

Hormann, A. M. (1964). Programs for machine learning. Part II. *Information and Control*, *7*(1), 55–77. https://doi.org/10.1016/S0019-9958(64)90259-1

Horne, H. N., Lee, P. S., Murphy, S. K., Alonso, M., Olson, J. A., & Marks, J. R. (2009). Inactivation of the MAL gene in breast cancer is a common event that predicts benefit from adjuvant chemotherapy. *Molecular Cancer Research*, *7*(2), 199–209. https://doi.org/10.1158/1541-7786.MCR-08-0314

Hou, Q., Liao, F., Zhang, S., Zhang, D., Zhang, Y., Zhou, X., Xia, X., Ye, Y., Yang, H., Li, Z., Wang, L., Wang, X., Ma, Z., Zhu, Y., Ouyang, L., Wang, Y., Zhang, H., Yang, L., Xu, H., & Shu, Y. (2017). Regulatory network of GATA3 in pediatric acute lymphoblastic leukemia. *Oncotarget*, *8*(22), 36040–36053. https://doi.org/10.18632/oncotarget.16424

Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., Zeng, W., & Zhou, H. (2019). Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *Journal of Hydrology*, *574*(December 2018), 1029–1041. https://doi.org/10.1016/j.jhydrol.2019.04.085

Huang, S., Nianguang, C. A. I., Penzuti Pacheco, P., Narandes, S., Wang, Y., & Wayne, X. U. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics and Proteomics*, *15*(1), 41–51. https://doi.org/10.21873/cgp.20063

Hütz, K., Zeiler, J., Sachs, L., Ormanns, S., & Spindler, V. (2017). Loss of desmoglein 2 promotes tumorigenic behavior in pancreatic cancer cells. *Molecular Carcinogenesis*, *56*(8), 1884–1895. https://doi.org/10.1002/mc.22644

Hwang, S. M. (2020). Classification of acute myeloid leukemia. *Blood Research*, *55*(S1), 1–4. https://doi.org/10.5045/br.2020.S001

Jakkula, V. (2011). Tutorial on Support Vector Machine (SVM). *School of EECS, Washington State University*, 1–13. http://www.ccs.neu.edu/course/cs5100f11/resources/jakkula.pdf

Jiang, C., Zhou, Y., Huang, Y., Wang, Y., Wang, W., & Kuai, X. (2019). Overexpression of ADAMTS-2 in tumor cells and stroma is predictive of poor clinical prognosis in gastric cancer. *Human Pathology*, *84*, 44–51. https://doi.org/10.1016/j.humpath.2018.08.030

Jiang, N., Leach, L. J., Hu, X., Potokina, E., Jia, T., Druka, A., Waugh, R., Kearsey, M. J., & Luo, Z. W. (2007). Methods for evaluating gene expression from Affymetrix microarray datasets. *BMC Bioinformatics*, *9*, 1–10. https://doi.org/10.1186/1471-2105-9-284

Johansson, B., & Harrison, C. J. (2021). Acute Myeloid Leukemia. In C. Röllig & G. J. Ossenkoppele (Eds.), *Cancer Cytogenetics*. Springer International Publishing. https://doi.org/10.1007/978-3-030-72676-8

Jordan, M. I., & Mitchell, T. M. (2015). *Machine learning: Trends, perspectives, and prospects*. *349*(6245).

Kalmár, A., Péterfia, B., Hollósi, P., Galamb, O., Spisák, S., Wichmann, B., Bodor, A., Tóth, K., Patai, Á. V., Valcz, G., Nagy, Z. B., Kubák, V., Tulassay, Z., Kovalszky, I., & Molnár, B. (2015). DNA hypermethylation and decreased mRNA expression of MAL, PRIMA1, PTGDR and SFRP1 in colorectal adenoma and cancer. *BMC Cancer*, *15*(1), 1–14. https://doi.org/10.1186/s12885-015-1687-x

Kazemi, F., Najafabadi, T., & Araabi, B. (2016). Automatic recognition of acute myelogenous leukemia in blood microscopic images using K-means clustering and support vector machine. *Journal of Medical Signals and Sensors*. https://doi.org/10.4103/2228-7477.186885

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 3147–3155.

Kim, J., Beidler, P., Wang, H., Li, C., Quassab, A., Coles, C., Drescher, C., Carter, D., & Lieber, A. (2020). Desmoglein-2 as a prognostic and biomarker in ovarian cancer. *Cancer Biology and Therapy*, *21*(12), 1154–1162. https://doi.org/10.1080/15384047.2020.1843323

Kumar, C. C. (2011). Genetic abnormalities and challenges in the treatment of acute myeloid Leukemia. *Genes and Cancer*, *2*(2), 95–107. https://doi.org/10.1177/1947601911408076

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, *36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11

Laukkanen, S., Liuksiala, T., Nykter, M., Heinäniemi, M., & Lohi, O. (2015). Identification of Novel Drug Targets in T-Cell Acute Lymphoblastic Leukemia. *Blood*, *126*(23), 3646–3646. https://doi.org/10.1182/blood.v126.23.3646.3646

Lee, P. S., Teaberry, V. S., Bland, A. E., Huang, Z., Whitaker, R. S., Baba, T., Fujii, S., Secord, A. A., Berchuck, A., & Murphy, S. K. (2010). Elevated MAL expression is accompanied by promoter hypomethylation and platinum resistance in epithelial ovarian cancer. *International Journal of Cancer*, *126*(6), 1378–1389. https://doi.org/10.1002/ijc.24797

Lee, S., Vo, T. P., Thai, H. T., Lee, J., & Patel, V. (2021). Strength prediction of concrete-filled steel tubular columns using Categorical Gradient Boosting algorithm. *Engineering Structures*, *238*(November 2020), 112109. https://doi.org/10.1016/j.engstruct.2021.112109

Li, Y., Ishiguro, H., Kawahara, T., Kashiwagi, E., Izumi, K., & Miyamoto, H. (2014). Loss of GATA3 in bladder cancer promotes cell migration and invasion. *Cancer Biology and Therapy*, *15*(4), 428–435. https://doi.org/10.4161/cbt.27631

Liu, H., Bebu, I., & Li, X. (2010). Microarray probes and probe sets. *Frontiers in Bioscience (Elite Edition)*, *2*, 325–338. https://doi.org/10.2741/e93

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, *13*(5), 1–23.

https://doi.org/10.1371/journal.pcbi.1005457
Manfrini, N., Mancino, M., Miluzio, A., Oliveto, S., Balestra, M., Calamita, P., Alfieri, R., Rossi, R. L., Sassoe-Pognetto, M., Salio, C., Cuomo, A., Bonaldi, T., Manfredi, M., Marengo, E., Ranzato, E., Martinotti, S., Cittaro, D., Tonon, G., & Biffo, S. (2020). FAM46C and FNDC3A are multiple myeloma tumor suppressors that act in concert to impair clearing of protein aggregates and autophagy. *Cancer Research*, *80*(21), 4693–4706. https://doi.org/10.1158/0008-5472.CAN-20-1357
Mehra, R., Varambally, S., Ding, L., Shen, R., Sabel, M. S., Ghosh, D., Chinnaiyan, A. M., & Kleer, C. G. (2005). Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. *Cancer Research*, *65*(24), 11259–11264. https://doi.org/10.1158/0008-5472.CAN-05-2495
MoradiAmin, M., Memari, A., Samadzadehaghdam, N., Kermani, S., & Talebi, A. (2016). Computer aided detection and classification of acute lymphoblastic leukemia cell subtypes based on microscopic image analysis. *Microscopy Research and Technique*, *79*(10), 908–916. https://doi.org/10.1002/jemt.22718
Mosquera Orgueira, A., Peleteiro Raíndo, A., Cid López, M., Díaz Arias, J. Á., González Pérez, M. S., Antelo Rodríguez, B., Alonso Vence, N., Bao Pérez, L., Ferreiro Ferro, R., Albors Ferreiro, M., Abuín Blanco, A., Fontanes Trabazo, E., Cerchione, C., Martinnelli, G., Montesinos Fernández, P., Mateo Pérez Encinas, M., & Luis Bello López, J. (2021). Personalized Survival Prediction of Patients With Acute Myeloblastic Leukemia Using Gene Expression Profiling. *Frontiers in Oncology*. https://doi.org/10.3389/fonc.2021.657191
Mottal, N., Issa, N., Dumas, P.-Y., Camou, F., Sauvezie, M., Gros, F.-X., Cazaubiel, T., Mourissoux, G., Leroy, H., Pigneux, A., Guisset, O., & Leguay, T. (2020). Reduce Mortality and Morbidity in Acute Myeloid Leukemia With Hyperleukocytosis With Early Admission in Intensive Care Unit: A Retrospective Analysis. *Journal of Hematology*. https://doi.org/10.14740/jh691
Muhsen, I. N., Shyr, D., Sung, A. D., & Hashmi, S. K. (2020). Machine Learning Applications in the Diagnosis of Benign and Malignant Hematological Diseases. *Clinical Hematology International*, *3*(1), 13. https://doi.org/10.2991/chi.k.201130.001
Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, *2*(4), 345–389. https://doi.org/10.1023/A:1009744630224
Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics*, *18*(6), 275–285. https://doi.org/10.1002/cem.873
Naqa, I. El, & Murphy, M. J. (2015). Machine Learning in Radiation Oncology. In I. El Naqa, R. Li, & M. J. Murphy (Eds.), *Machine Learning in Radiation Oncology*. Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3
Nazari, E., Farzin, A. H., Aghemiri, M., Avan, A., Tara, M., & Tabesh, H. (2020). Deep Learning for Acute Myeloid Leukemia Diagnosis. *Journal of Medicine and Life*. https://doi.org/10.25122/jml-2019-0090
Norton, S. W. (1989). Generating better decision trees. *Proceedings of the Eleventh International Joint …*, 800–805. http://www.ijcai.org/Past Proceedings/IJCAI-89-VOL1/PDF/128.pdf
Ozaki, K., Nagata, M., Suzuki, M., Fujiwara, T., Miyoshi, Y., Ishikawa, O., Ohigashi, H., Imaoka, S., Takahashi, E. I., & Nakamura, Y. (1998). Isolation and characterization of a novel

human pancreas-specific gene, pancpin, that is down-regulated in pancreatic cancer cells. *Genes Chromosomes and Cancer*, *22*(3), 179–185. https://doi.org/10.1002/(SICI)1098-2264(199807)22:3<179::AID-GCC3>3.0.CO;2-T

Parkinson, J., & Blaxter, M. (2009). Expressed Sequence Tags: An Overview. In J. Parkinson (Ed.), *Family medicine* (Vol. 533, Issue 8). Humana Press. https://doi.org/10.1007/978-1-60327-136-3

Pejovic, T., Schwartz, P. E., & Oncology, G. (2002). *Review of Leukemias*. *45*(3), 866–878. https://doi.org/10.1097/01.grf.0000022391.03824.2d

Peng, X., Wen, Y., Zha, L., Zhuang, J., Lin, L., Li, X., Chen, Y., Liu, Z., Zhu, S., Liang, J., Zhou, Z., Yuan, W., Li, Y., Wang, Y., Jiang, Z., Mo, X., Wan, Y., Shi, Y., Zhu, P., … Fan, X. (2019). TRIM45 Suppresses the Development of Non-small Cell Lung Cancer. *Current Molecular Medicine*. https://doi.org/10.2174/1566524019666191017143833

Percival, M. E., Lai, C., Estey, E., & Hourigan, C. S. (2017). Bone marrow evaluation for diagnosis and monitoring of acute myeloid leukemia. *Blood Reviews*, *31*(4), 185–192. https://doi.org/10.1016/j.blre.2017.01.003

Perner, P. (2012). *LNAI 7376 - Machine Learning and Data Mining in Pattern Recognition*. https://link.springer.com/content/pdf/10.1007%2F978-3-642-31537-4.pdf

Qin, S., Liao, Y., Du, Q., Wang, W., Huang, J., Liu, P., Shang, C., Liu, T., Xia, M., & Yao, S. (2020). DSG2 expression is correlated with poor prognosis and promotes early-stage cervical cancer. *Cancer Cell International*, *20*(1), 1–13. https://doi.org/10.1186/s12935-020-01292-x

Radakovich, N., Nagy, M., & Nazha, A. (2020). Machine learning in haematological malignancies. *The Lancet Haematology*, *7*(7), e541–e550. https://doi.org/10.1016/S2352-3026(20)30121-6

Raiser, D. M., Narla, A., & Ebert, B. L. (2014). The emerging importance of ribosomal dysfunction in the pathogenesis of hematologic disorders. *Leukemia and Lymphoma*, *55*(3), 491–500. https://doi.org/10.3109/10428194.2013.812786

Ranstam, J., & Cook, J. A. (2018). LASSO regression. *British Journal of Surgery*, *105*(10), 1348. https://doi.org/10.1002/bjs.10895

Rao, B. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179. https://doi.org/10.21275/ART20203995

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., … Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, *7*(11), 909–912. https://doi.org/10.1038/nmeth.1517

Rodriguez-Abreu, D., Bordoni, A., & Zucca, E. (2007). Epidemiology of hematological malignancies. *Annals of Oncology*, *18*(SUPPL. 1). https://doi.org/10.1093/annonc/mdl443

Roemer, A., Schwettmann, L., Jung, M., Stephan, C., Roigas, J., Kristiansen, G., Loening, S. A., Lichtinghagen, R., & Jung, K. (2004). The membrane proteases ADAMs and hepsin are differentially expressed in renal cell carcinoma. Are they potential tumor markers? *Journal of Urology*, *172*(6 I), 2162–2166. https://doi.org/10.1097/01.ju.0000144602.01322.49

Roushangar, R., & Mias, G. I. (2019). Multi-study reanalysis of 2,213 acute myeloid leukemia patients reveals age- and sex-dependent gene expression signatures. *Scientific Reports*,

9(1), 1–17. https://doi.org/10.1038/s41598-019-48872-0
Saaber, F., Chen, Y., Cui, T., Yang, L., Mireskandari, M., & Petersen, I. (2015). Expression of desmogleins 1-3 and their clinical impacts on human lung cancer. *Pathology Research and Practice*, *211*(3), 208–213. https://doi.org/10.1016/j.prp.2014.10.008
Sachdeva, S., & Kumar, B. (2021). Comparison of gradient boosted decision trees and random forest for groundwater potential mapping in Dholpur (Rajasthan), India. *Stochastic Environmental Research and Risk Assessment*, *35*(2), 287–306. https://doi.org/10.1007/s00477-020-01891-0
Salcedo-Sanz, S., Rojo-Álvarez, J. L., Martínez-Ramón, M., & Camps-Valls, G. (2014). Support vector machines in engineering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(3), 234–267. https://doi.org/10.1002/widm.1125
Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K. R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(11), 2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820
Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review. *Frontiers in Aging Neuroscience*, *9*(OCT), 1–12. https://doi.org/10.3389/fnagi.2017.00329
Sawyers, C. L. (1999). Chronic Myeloid Leukemia. *The New England Journal of Medicine*, *340*, 1330–1340. https://doi.org/10.1056/nejm19990429340170
Sekeres, M. A., Stone, R. M., Zahrieh, D., Neuberg, D., Morrison, V., De Angelo, D. J., Galinsky, I., & Lee, S. J. (2004). Decision-making and quality of life in older adults with acute myeloid leukemia or advanced myelodysplastic syndrome. *Leukemia*, *18*(4), 809–816. https://doi.org/10.1038/sj.leu.2403289
Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., & Gifford, E. M. (2016). Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. *Journal of Chemical Information and Modeling*, *56*(12), 2353–2360. https://doi.org/10.1021/acs.jcim.6b00591
Shi, J., Zhang, L., Zhou, D., Zhang, J., Lin, Q., Guan, W., Zhang, J., Ren, W., & Xu, G. (2018). Biological function of ribosomal protein L10 on cell behavior in human epithelial ovarian cancer. *Journal of Cancer*, *9*(4), 745–756. https://doi.org/10.7150/jca.21614
Short, N. J., Rytting, M. E., & Cortes, J. E. (2018). Acute myeloid leukaemia. *The Lancet*, *392*(10147), 593–606. https://doi.org/10.1016/S0140-6736(18)31041-9
Simeone, O. (2018). A Very Brief Introduction to Machine Learning with Applications to Communication Systems. *IEEE Transactions on Cognitive Communications and Networking*, *4*(4), 648–664. https://doi.org/10.1109/TCCN.2018.2881442
Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044
Stalteri, M. A., & Harrison, A. P. (2007). Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics*, *8*, 1–15. https://doi.org/10.1186/1471-2105-8-13
Thermo Fisher Scientific Inc. (2017). Retrieved 10/09/2021, from http://www.affymetrix.com
Tota, G., Coccaro, N., Zagaria, A., Anelli, L., Casieri, P., Cellamare, A., Minervini, A., Minervini, C. F., Brunetti, C., Impera, L., Carluccio, P., Cumbo, C., Specchia, G., & Albano, F. (2014). ADAMTS2 gene dysregulation in T/myeloid mixed phenotype acute leukemia. *BMC Cancer*, *14*(1), 1–6. https://doi.org/10.1186/1471-2407-14-963

Vakili, M., Ghamsari, M., & Rezaei, M. (2020). *Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification*. http://arxiv.org/abs/2001.09636

Walter, R. B., Othus, M., Burnett, A. K., Löwenberg, B., Kantarjian, H. M., Ossenkoppele, G. J., Hills, R. K., Van Montfort, K. G. M., Ravandi, F., Evans, A., Pierce, S. R., Appelbaum, F. R., & Estey, E. H. (2013). Significance of FAB subclassification of "acute myeloid leukemia, NOS" in the 2008 WHO classification: Analysis of 5848 newly diagnosed patients. *Blood*, *121*(13), 2424–2431. https://doi.org/10.1182/blood-2012-10-462440

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics in Western Equatoria State. *Nature Reviews Genetics*, *10*(1), 57.

Warnat-Herresthal, S., Perrakis, K., Taschler, B., Becker, M., Baßler, K., Beyer, M., Günther, P., Schulte-Schrepping, J., Seep, L., Klee, K., Ulas, T., Haferlach, T., Mukherjee, S., & Schultze, J. L. (2020). Scalable Prediction of Acute Myeloid Leukemia Using High-Dimensional Machine Learning and Blood Transcriptomics. *IScience*, *23*(1). https://doi.org/10.1016/j.isci.2019.100780

Wassenaar, C. A., Dong, Q., Wei, Q., Amos, C. I., Spitz, M. R., & Tyndale, R. F. (2011). Relationship between CYP2A6 and CHRNA5-CHRNA3-CHRNB4 variation and smoking behaviors and lung cancer risk. *Journal of the National Cancer Institute*, *103*(17), 1342–1346. https://doi.org/10.1093/jnci/djr237

Welch, J. S., Ley, T. J., Link, D. C., Miller, C. A., Larson, D. E., Koboldt, D. C., Wartman, L. D., Lamprecht, T. L., Liu, F., Xia, J., Kandoth, C., Fulton, R. S., McLellan, M. D., Dooling, D. J., Wallis, J. W., Chen, K., Harris, C. C., Schmidt, H. K., Kalicki-Veizer, J. M., … Wilson, R. K. (2012). The origin and evolution of mutations in acute myeloid leukemia. *Cell*, *150*(2), 264–278. https://doi.org/10.1016/j.cell.2012.06.023

WHO. (2014). ACUTE MYELOGENOUS LEUKEMIA (Including Acute Promyelocytic Leukemia). *2014 Review of Cancer Medicines*, *List of Essential Medicines*, 1–14.

Xie, C., Lu, J., & Parkany, E. (2003). Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks. *Transportation Research Record*, *1854*, 50–61. https://doi.org/10.3141/1854-06

Yang, J., Chen, Z., Liu, N., & Chen, Y. (2018). Ribosomal protein L10 in mitochondria serves as a regulator for ROS level in pancreatic cancer cells. *Redox Biology*, *19*(May), 158–165. https://doi.org/10.1016/j.redox.2018.08.016

Yang, T., Gu, X., Jia, L., Guo, J., Tang, Q., Zhu, J., Zhao, W., & Feng, Z. (2021). DSG2 expression is low in colon cancer and correlates with poor survival. *BMC Gastroenterology*, *21*(1), 1–10. https://doi.org/10.1186/s12876-020-01588-2

Yashiro, M., Nishioka, N., & Hirakawa, K. (2006). Decreased expression of the adhesion molecule desmoglein-2 is associated with diffuse-type gastric carcinoma. *European Journal of Cancer*, *42*(14), 2397–2403. https://doi.org/10.1016/j.ejca.2006.03.024

Zhang, J., Zhang, C., Cui, J., Ou, J., Han, J., Qin, Y., Zhi, F., & Wang, R. F. (2017). Trim45 functions as a tumor suppressor in the brain via its e3 ligase activity by stabilizing p53 through k63-linked ubiquitination. *Cell Death and Disease*, *8*(5), 1–11. https://doi.org/10.1038/cddis.2017.149

Zhang, Y., Zhao, Z., & Zheng, J. (2020). CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. *Journal of Hydrology*, *588*(April), 125087. https://doi.org/10.1016/j.jhydrol.2020.125087

Zhao, Q., Zheng, B., Meng, S., Xu, Y., Guo, J., Chen, L. jie, Xiao, J., Zhang, W., Tan, Z. rong,

Tang, J., Chen, L., & Chen, Y. (2019). Increased expression of SLC46A3 to oppose the progression of hepatocellular carcinoma and its effect on sorafenib therapy. *Biomedicine and Pharmacotherapy*, *114*(March), 108864. https://doi.org/10.1016/j.biopha.2019.108864