FACULTAT
DE CIÈNCIES
I TECNOLOGIA
UVIC | UVIC·UCC

Degree Final Project

Applied Analytics for clinical decision support

Estela Cabezas Espanyol

Degree in Biotechnology

Advisor: Gaston Besanson

Endorser: Carlo Manzo

Vic, June 2019

Acknowledgements

I would like to thank the collaboration and patience of the Accenture Health Dream Team, specially to Ruben Sanchez to provide the base of the ORBDA dataset solution so I could develop the Case Study 1. Also, I would like to make special mention to Alejandro Suárez and Xabier Beraza for their will to make sure I was understanding all the theoretical background on R packages, their implementation and how the results were being analyzed.

To the AML/ALL Research Team for spending their weekends and nights helping me to develop the model used in the Case Study 2 and for answering my questions when possible. Also, to the Intel's Software Developer project for the events I was able to attend and by giving me voice on crazy ideas as presenting a computational model using a biological background.

Most appreciation on the tutoring of Gaston Besanson and Carlo Manzo, whom within their busy lives spent some time pushing and challenging me to achieve knowledge on a complete unknown area.

Finally, to my family and friends, specially to you Mom, whom never understood what I was doing nor studying but supported me in the most difficult times.

"My mother always taught us that if people don't agree with you, the important thing is to listen to them. But if you've listened to them carefully and you still think that you're right, then you must have the courage of your convictions." - Jane Goodall

Summary of the final degree project in Biotechnology

Title: Applied Analytics for clinical decision support

Key words: *Analytics, Machine Learning, Electronic Health Records, Convolutional Neural Networks*

Author: Estela Cabezas Espanyol

Advisors: Gaston Besanson and Carlo Manzo

Date: June 2019

System digitalization has allowed storage of massive data, Big Data, and its management through Data Science techniques. When information is available at this scale, it is possible to measure and, therefore, analyze it. Big Data is now becoming a wide used methodology for data storage in an organized manner and thus, facilitates the assessment of clinical decisions. Still, approximately the 95% of large datasets are comprised of un-structured data and noise.

This work provides an insight on how data quality is the major concert when applying processing techniques, such as intelligent analytics and Machine Learning algorithms, to improve an ongoing procedure.

The major focus of study is the optimization health care systems through the analysis of how the impact of a health-related advertising campaign produced changes in the hospitalization frequency of Brazilian Hospitals during 2010. The analytics-based analysis was performed using a predictive R package, which defined an increase of the 13% on the general frequency of hospitalizations based on Chemotherapy, Radiotherapy and Renal dialysis treatments.

A second methodology was applied to study the implementation of a Convolutional Neural Network for the automatic classification of images obtained from blood samples associated to Leukemia. A binary classification of two different augmented datasets was used to train the model, using the original images to predict its performance. The results obtained showed an accuracy of 66,3% and 44.2%, revealing that this model requires further modifications and a wider exploration of the hyperparameters' space.

Analytics provides a visual understanding of the data distribution along Electronic Health Records, and other datasets, being a good methodology to improve hospital facilities for a predictive disease outbreak or to provide personalized treatments. Convolutional Neural Networks, when applied correctly, provide a major optimization of sample classification due to its assembly to human brain pattern recognition procedure and are a good method to assist pathologists. Overall, these methodologies provide a novel approach to develop further health applications for clinical decision support.

Contents

1 Introduction

The transition from paper-based data to digitalized records has increased the amount of available treatable information known as *Big Data*. Big Data not only refers to quantity but to the complexity, significance and difficulty to read. This progress has led to a real necessity of a new processing system and technology, known as *Data Science*, to understand, manage, store and visualize the information. [1], [2]

Nowadays, there is no clear definition of Big Data, but there is the necessity to implement grounded technologies on data mining and statistical analysis. Therefore, Data Science procedures englobe many areas of data processing.

Software solutions as R or other programming interfaces have been developed and improved over the time for easy usage.

It is quite well - known that intelligent methodologies are taking over a wide range of sectors and could not be less when referring to healthcare systems.

Predictive algorithms to increase medical staff for a given disease outbreak, implementing efficient drug supply to needed countries or even a simulation of what type of harm would an environmental disaster produce, can help to avoid hospital crushes and decrease mortality rates.

It is even possible to optimize doctor's decision by implementing intelligent clinical support systems, as early detection of chronical diseases or to reduce medical errors by increasing healthcare quality and efficiency.

All these methodologies rely in a wide and new area of study known as *Machine Learning* (ML), where the main process is to train a model in order to obtain an optimized product. The training step produces variability over different ML algorithms and their performance, which can be classified in classical algorithms, Neural Networks - Deep learning, Reinforcement learning and ensemble methods.

Hence, this thesis intends to reveal some of the methodologies and techniques learnt during a six-month internship in Accenture Applied Intelligence where the aim was to give client access to non-traditional healthcare services by combining *Artificial Intelligence* (AI) with data, analytics and automation. Accenture is well known for its consulting services in many other areas and industries such as automotive, banking and capital markets, among others.

Understanding data and Analytics and are key points to develop a good structured procedure. These processes can be developed under visualization techniques, representing a story and through synthetic techniques to arrange, divide or calculate particular objects comprised in datasets.

The knowledge acquired though the internship is reflected along the first two thirds of the thesis and justified under Case Study 1. The study relies on categorical data from the Brazilian healthcare system during a time-period of four years, 2008-2012.

Not all Artificial Intelligence development is about computational processes; it is also to broadcast emerging techniques and ideas.

New ideas come from what society needs and why. The deployment of intelligent models keeps on frightening most of the population due to the lack of knowledge on how those work and producing a back propagation of innovative projects.

Independently of the work done during my internship, I have been volunteering in a Research Project based on image classification of blood samples for Leukemia early detection. This project is also shown in this thesis under Case Study 2. The images used for the model were previously validated and supported by professional oncologists.

This project encouraged me to become more active on opensource solutions by pushing me to develop better communication and knowledge-transfer skills. After few months working on the project, I became a member of the Intel Software Innovator program.

The program supports independent developers who aim to create and demonstrate new projects by providing speaking and demonstration opportunities in events. With that, I was able to attend the Embedded World, last February, where a demo was displayed about the classification model, and the Intel Developer Affinity Day, last March, where I explained to different members of the initiative which further steps had been developed on the project since Embedded World.

1.1 Data Cleaning

Statistical analysis can be performed to describe knowledge about a procedure or to make decisions about a method. To determine the feasibility of the data it is necessary to discriminate between sampling errors or extra sampling errors. Extra sampling errors are the ones produced when gathering or manipulating data and can be systematic or aleatory affecting the magnitude of the overall dataset [3].

When data is collected it is important to know what type of variables are in it and their distribution. If the distribution is bad, as having more variables of one class than the other, or if the dataset is too large, it requires a treatment defined as data cleaning to filter the information [4].

For that, we need to know the tags appearing in the dataset such as the coding used for different diseases or procedures. If relaying to a set of patients identified by numbers is important to ensure no patient is repeated, and if so, join the available information into one observation.

Large datasets tend to have missing values due to incompatibility when transferring into electronic form or just because data is physically unavailable (not available, NA). Thus, it is necessary to, again, understand the overall and make the best decision on whether discarding or fixing these NA values.

1.2 Electronic Health Records and Privacy

This section describes the different type of electronic records available and distinguishes between each other in structure and security level. The content in the documents is important thus gives a clue of possible coding systems. Privacy laws are becoming harsher towards civil protection when concerning sensitive data.

1.2.1 Electronic Health Records

Nowadays we are witnessing an increasing trend on adopting electronic sources able to summarize enormous quantities of medical data supported by businesses, in this case hospitals, as well as and medical decisions. The more data we gather, the more valuable this resource becomes. Pre-processing is needed before being able to apply any type of computational work due to human intervention and its disorganized nature, plus requires of a standardization process. This methodology is called ***Electronic Health Records*** (EHR) and can also be found as ***Electronic Medical Records*** (EMR) which are used in clinical settings and for decisions support but are only applicable in local clinics or hospitals.

EMR provide a more accurate privacy level per patient identification. If going deeper, one can even find ***Personal Health Records*** (PHR), where the user must manage and grant permission for access or for sharing with third-parties.

The usage of EHR reduces unnecessary costs in low patient flux periods and, on in the other hand, improve patient attention when the flux is high. Decision management can be improved for entities such as health care administration by process automation when transforming paper to electronic records, but the authentication process decreases in efficiency due to the large amount of data or when the user interface is not easy-to-use by the user. User interfaces are important in analytics since it is possible to create visual representations of the information or even generate automatic reports.

High dimensionality of datasets challenges the statistical significance and large computational time affecting its accuracy such as when performing disease detection, sequential prediction of clinical events, concept embedding or when trying to avoid de-identification (privacy).

Now that we have discussed the most typical issues with open data, let's go deeper into analytics. With the available technology and the will to improve comfort, it is possible to help professionals to make decisions within seconds. In 2013, a new *Clinical Decision Support System (CDSS)* was made available for diabetes cases by suggesting the next steps to treat the disease, alert providers of available information missed, catch potential problems with drug interactions, as well as helping with the diagnosis. After CDSS many other systems have since emerged [5].

Trying to arrange all the advantages mentioned previously into a system able to face different types of EHR, is an important consideration. Very known is the fact that hospital stuff have different script and usage of jargon but also different hospitals have different ways of recording medication, shifts or interventions. These kinds of syntactic and representative heterogeneities present an incompatibility issue.

Collecting data is not like recording a patient's ID when visiting, it is common to have a batch of the same sample. Timing differences between a doctor's decision and a laboratory result have a high complexity of measurement, high cost and quality issues. Other parameters involving data such as integration, quality or frequency rely on optimizing algorithms rather than into human intervention. The process leads into an analytic point of view. Furthermore, to choose the integration platform we must consider if we need real time analytics (Full-streaming or microbatch) or batch analytics.

To develop a system, and avoid these complications, standards have been set such as International *Organization for Standardization* (ISO/TC 215) and *Health Level 7* (HL7) (Health Level-7)[6]–[8].

1.2.2 Data Privacy

De-identified, encrypted or pseudonymized personal data that can be used for re-identification and relies on the standards of *General Data Protection Regulation* (GDPR), which regulate the processing by an individual, company or organization within the European Union.

To improve interoperability between clinical information models some standards, defined below, and committees such as the European Committee for Standardization (CEN/TC 251), the Clinical Information Modeling Initiative (CIMI) or the Conceptual Framework (CF) have been set [9], [10] .

1.2.2.1 Health Level Seven

Health Level Seven (HL7) refers to a set of international standards for transfer of clinical and administrative data between software applications used by various healthcare providers by relaying on how data is collected and structured.

The most widely used messaging standard is version 2 (HL7v2), which regulates the electronic data exchange. While it does not provide specific elements for security and privacy there are some elements that can be used to convey information to support GDPR as the CON Segment. This purpose is mainly for consent to medical treatment but can also be used to authorize the disclosure of protected health information.

The next version, HL7v3, was developed to improve HL7v2, but ended up not being as used.

HL7v3 is based on the implementation of ISO 22600 policy ontology to identify the information and system behavior to implement technological controls though specific labelling which assign security labels to healthcare information allowing tracking when information is transferred between systems.

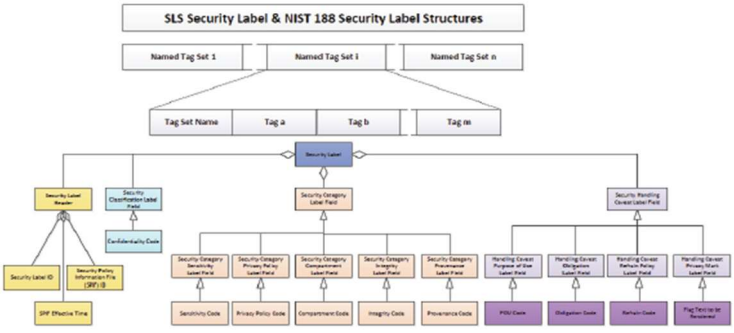


Figure 1. HL7 Security and Privacy labels. Caption from HL7 Standards and Components to support implementation of the GDPR.

The functional capabilities and infrastructures should contain a set of label considerations, fabric services, security and Privacy Ontology Based Terminology Services (POBTS) and Privacy Protective Services (PPS).

HL7 generally provides permission vocabulary as {operation, object} structure according to role-based access control.

1.2.2.2 Fast Healthcare Interoperability Resources

Security and privacy labels can be attached to a resource or bundled as metadata to provide specific security and privacy attributes and information such as context of use, data sensitivity and control of flow.

The compartment resource was designed to group resources which share common properties such as patient, reason of encounter, relatives, practitioner and device.

FHIR defines two resources: (i) Provenance Resource, which gathers information about an event which occurred during a specific time-period and if it was either created de novo or provided by another entity; and (ii) AuditEvent, which provides records of an event with the purpose of maintaining a security log and it is meant to be used only for Security purposes and personnel.

```
<Patient xmlns="http://hl7.org/fhir">
  <meta>
    <security>
      <system value="http://hl7.org/fhir/v3/Confidentiality"/>
      <code value="R"/>
      <display value="Restricted"/>
    </security>
  </meta>
...  [snip] ...
</Patient>
```

Figure 2. Example for security label as metadata for a FHIR resource. Caption from HL7 Standards and Components to support implementation of the GDPR.

1.3 Analytics

Analytics is nothing less than the application of Data Science to different fields of knowledge to improve an ongoing procedure. The era of Big Data has allowed the development of this technique in different aspects. As with any approach, it is necessary to take into consideration a real target. In healthcare systems, this target could be the discovery of patterns, trends and association in an inputted set of data to optimize the manufacture of a vaccine or to predict a disease outbreak. It therefore creates an insight advantage for better decision-making in the overall picture of society or companies.

Data access and its structure is essential, analytics is used when willing to optimize an existing procedure. If data is disorganized or has a lack of uniqueness, an analytic development requires more time and the initiative for a project will face big delays. These data characteristics must be considered when planning a project.

No part of this procedure englobing analytics would be of success without expertise minds working on it, for now the emerging background knowledge that raises as analyst are people with a formation on Data Science, model developers or process applicants. This limitation makes it even more interesting for up-to-date companies when willing to hire employees for their research and development departments[11]–[16].

1.4 Visualization

Visualization techniques as Graphs, Charts or Maps have been used in different study fields to represent a solution, result or requirement. Now, with Big Data, human comprehension triggers to understand the meaning of the gathered information or data's pattern.

To solve it, different softwares have been developed to ease users understanding of data. As examples, two different visualization techniques are defined, and their respective solutions are shown.

First, *Power BI* is an interactive tool which allows easy understanding on how data is related by providing a story along its variables. Information is shown as a Dashboard format providing correlation among different reports. *Power B*I has a set of metrics to show data (Average, Mean, Max, Min) and sometimes the desired information is seen triggered, for example when the desire is to represent an accumulative history or a single value within a time frame.

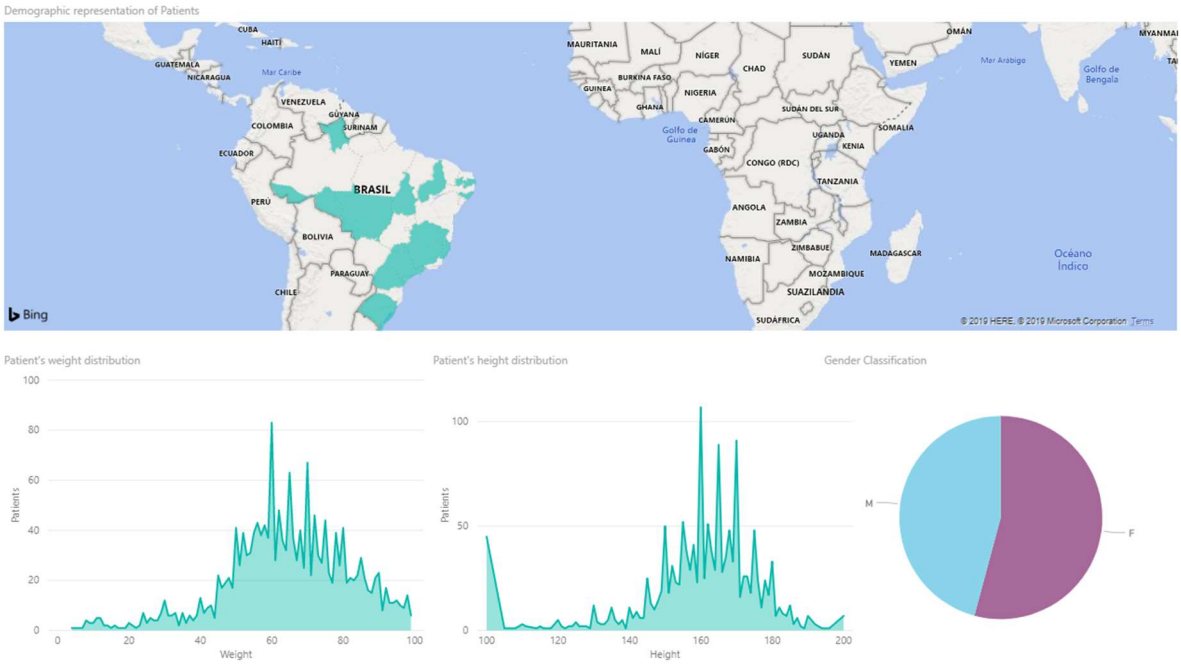


Figure 3. Through information selection as a specific region of Brazil, the graphs below display its respective patient record as Height, Weight and Gender. Power BI dashboard.

While PowerBI works locally, *Shiny* (an R package) is used to build interactive web applications.

Shiny allows the automatic actualization of data when inputs are uploaded, provides easy observation of large Datasets and is protected with security authentication for remote access. Shiny deployment does not require knowledge on JavaScript programming and has different default models for dashboard's visualization.

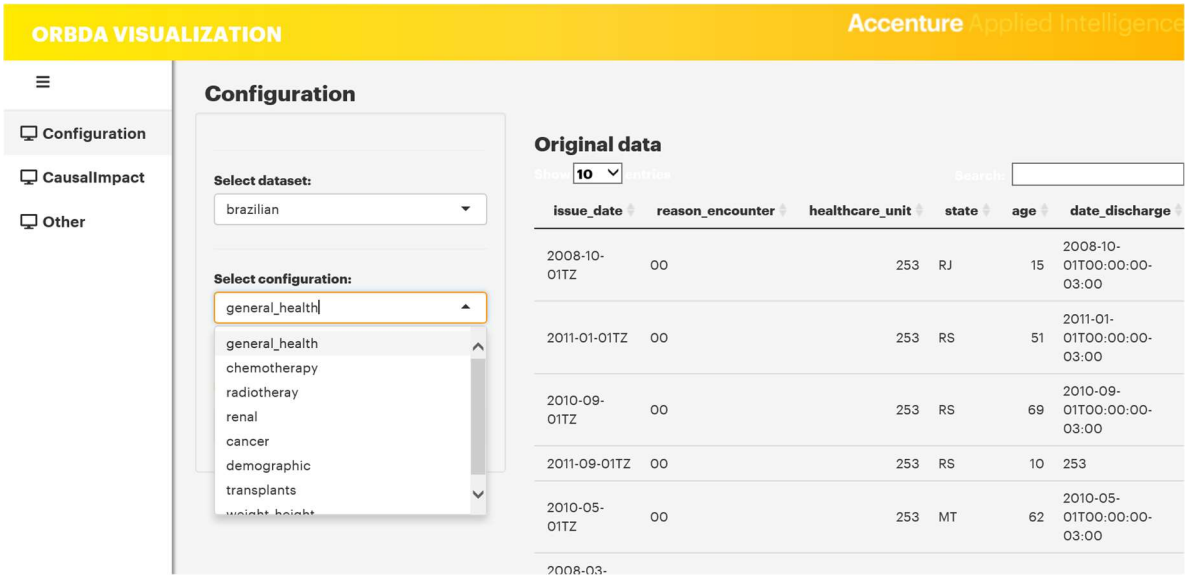


Figure 4. Through the selection of a Dataset and the features to display, is possible to visualize the original dataset. When browsing within Causal Impact or Other, different display of the data appear. Shiny dashboard.

1.5 Machine Learning

The ability of intelligent software to automate challenging human labor, *Artificial Intelligence* (AI), has been object of research since the early 20th century. This area faces complications when trying to perform unconscious human actions such as auditive and facial recognition. For this reason, it is necessary to expand the topic and find different-emerging ways of computational solutions. AI standards have been too high for the small steps taken forward but it is true to say that a decade ago no one would have thought an algorithm would make our lives easier by assessing on what we need or want [17], [18].

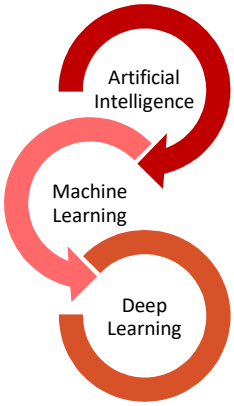


Figure 5. Concept Relationship.

AI has a wide range of fundamental knowledge representation techniques[19]. One of them is *Machine Learning* (ML), dated to the 1980's with the simple thought of a computer playing chess on its own. The first insight was the one already existing back then, using as input some data and the main rules of the game, and as output the correct answers for each movement. Instead, the new approach would give as input some data and the expected answers and as output the rules of how the game is performed. This last process allows the system to learn from the output and apply it to the new inputted data. The brain of the procedure is a sequence of instructions, gathered as an *algorithm,* performing the mentioned data transformation.

ML is also characterized by the ability to recognize patterns and extrapolate information to apply in different circumstances. The present era is the one comprising new technological advances providing tools to manage large amounts of data remotely [20].

ML algorithms are still being discovered and yet to be fully understood, for this reason there is no an exact classification along all available methods. The following two classes are defined according to how a system learns.

- ***Supervised learning*** models are fed with datasets containing defined features or defined labels. For each random observation made, the model associates it to a Y value or vector, previously defined as label.
- ***Unsupervised learning*** is a type of model which learns the probability distribution and generates a dataset. It is mostly used in problems where a relationship between objects exists and it is aimed to be understood and performed by clustering algorithms

Machine learning can be itemized into different types of *learning,* one of them is *Deep Learning* (DL), which arose in the early 2010s, and which is characterized by a set of successful layer representations. This subset of models differs from the other in the depth, number of layers, representing a *Neural Network (NN)* – the name was inspired by biological brain networks.

As mentioned before, the standards of these new approaches to improve different living areas are high, and with the new available techniques, businesses tend to go straight to DL algorithms, however this methodology is not always the most appropriate one since it relies on available data.

1.5.1 Neural Networks and Deep Learning

There is another large group of classifications concerning different algorithms and their architectures which include; Convolutional neural networks, Recurrent neural networks, Autoencoders and Generative Adversarial Networks [21] , [22].

- ***Convolutional Neural Networks*** (CNN), are mostly used in image recognition or computer vision. The algorithm takes an inputted image, assigns a set of learnable features to it and performs a differentiation task. This model works correctly for 2-D image since it has the ability to transform it to the correct first layer input through the activation functions. Unfortunately, it needs a large set of layers and training data. CNN is used in the case study of this thesis.

- ***Recurrent Neural Networks*** (RNN), are characterized by having a feedback connection allowing the activations to flow in a loop where the output is dependent on the previous computation, therefore having *memory*[1]. Within the RNN there are many models that have every neuron connected with each other or not, gradient descent procedures for backpropagation or non-linear mapping bases. This type of NN is mostly used to perform sequential computations of DNA.

- ***Autoencoder*** (AE) models have the same number of inputs and outputs but differ from the previous type since they are not used to classify but rather to recreate the inputted object. The algorithm is capable of rearranging the object and self-employ it to re-train the model. AE models can be used to predict de-identified gen functions within a large genomic dataset. [23]

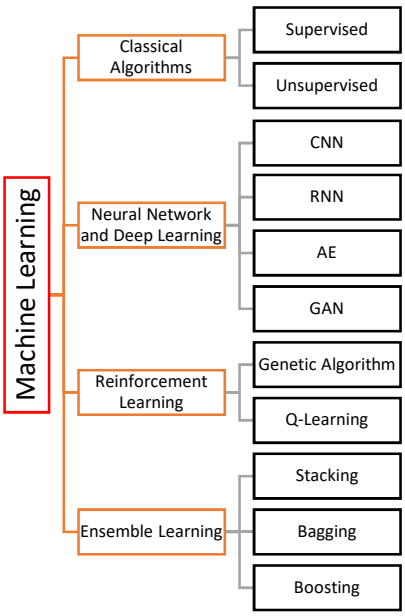


Figure 6. Machine Learning algorithm classification.

- ***Generative Adversarial Networks*** (GAN), have two differentiated components; a *generator*, which creates new synthetic data through the distribution of real samples and the *discriminator*, which decides whether each data belongs or not to the actual

[1] The calculated information is captured, stored and employed to determine the final output.

training set by looping through the generator's output. GAN has been accessible since 2014 and its classification is still unclear on whether it is a supervised, unsupervised or neither. This methodology can be applied to generate new molecules for drug design [24].

Deep learning models yield better performance in many tasks than traditional Machine learning methods and require less manual feature engineering.

1.5.2 Others

- ***Reinforcement learning*** is characterized by path performance under a reward hypothesis. This means that an agent performs the same action in different environments and when choosing the most optimized path it gets rewarded. This methodology can be implemented on evolutionary computation[25].

- ***Ensemble learning*** methodologies are based on the combination of different predictive algorithms in order to improve the performance of a specific model. Ensemble Learning can be used to study dependence or independence between base learners, for example to identify recombination target sites of DNA in a given recombination procedure [26].

1.6 Model Architecture

A *Convolutional Neural Network* (CNN) is an artificial intelligence model based under Deep learning characteristics. It is widely used for image processing techniques because, if the design is accurate enough, it can perform both generative and descriptive tasks, image recognition or natural language processing.

The layers of a CNN consist of an input layer, pooling layers, fully connected layers and normalization layers.

A CNN model requires a set of dimensions for the inputted image such as height, width and channel, which will be converted into a matrix. Channels, or depth, can be determined as 1 for gray scaling and 3 for RGB (red-green-blue). Image features are extracted through convolutional layers maintaining pixel relation by a learning procedure.

The channel or batch size indicates the number of samples getting into the NN at one time and it is defined in every layer. Once all the samples have passed through the NN, one epoch is completed.

Pixels shift over the input matrix following a set of strides, which defines the number of pixels moved at a time. If the filter does or does not fit the input image it is possible to tell the model whether to take it (as zero) or drop it through a padding parameter. Padding allows the pixels at the border of the image to be centered and interact with the filter.

For example, if the stride is defined as (1,1), the filter will move 1 pixel left for every horizontal movement, and 1 pixel down for every vertical movement, to create a feature map. Once the feature map has become smaller than the designated input dimensions, padding will solve it by adding zeros systematically. Note a filter of 3X3.

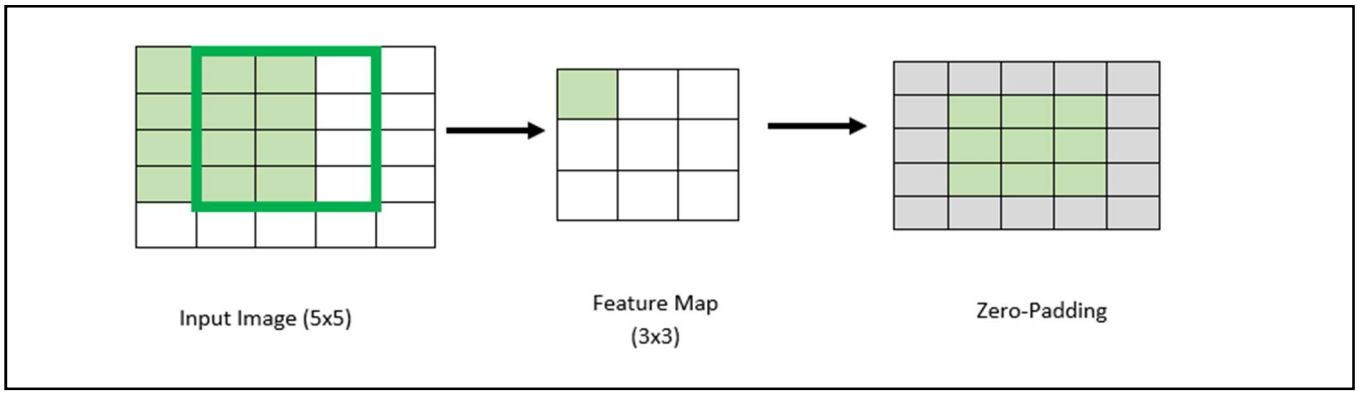


Figure 7. Visual representation of filtering parameters in a Convolutional Neural Network.

Another important parameter is the pooling Layer, which acts when the feature map is being processed. The pooling layer reduces the dimensionality of the map without discarding important information and operates in each map independently.

When a *max pooling* function is defined, the model will take the highest pixel value within a filter X and a stride Y.

To avoid the same mapping path between the input and response variables within a *neural network* (NN), activation functions are applied. These functions avoid a linear behavior in our NN and therefore, avoid its linearity. When having to choose which one to apply, we need to be aware of the input's behavior and numerical range , as for example ReLu, Softmax or Sigmoid which are the most frequently used ones.

The final layer of the NN is usually connected to the previous layer. Fully connected layers avoid feature assumptions by the network through outputting the real dimensions of the input.

2 Hypothesis

Machine Learning development on healthcare systems provide an optimization of procedures related to hospitalization rates, personalized vaccination and clinical decision support.

3 Objectives

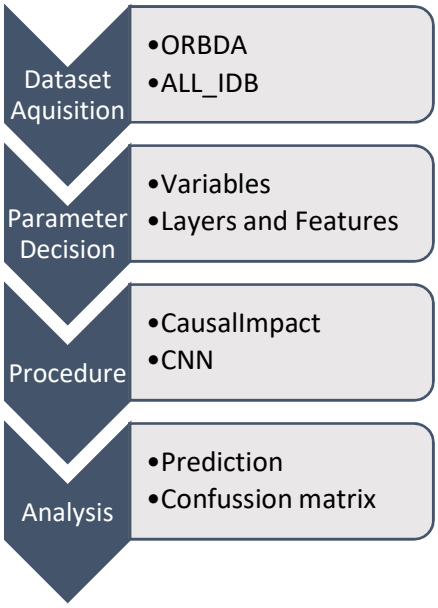
To reach the hypothesis, the two-fold goal of this thesis consisted in:

- The development of a complete Data Science framework, from data treatment until result visualization though a Brazilian Hospitalization dataset comprised of categorical data gathered along different time periods.
- The development of a Deep Learning architecture based on CNN the for automatic classification of microscopy images of blood samples associated to Leukemia.

The nature of the study required interdisciplinary competences to develop all the tasks set as objectives.

- Learn and achieve complete knowledge on artificial intelligence applications in real-life cases
- Bibliographical review, synthesis and analysis
- Reference research and review
- Comprehension of Data Science techniques and Analytic models
- Development of R and Python models
- Result metrics and visualization
- Communication skills to transmit ideas and points of view

3.1 Organization and structure



This thesis can be represented with a similar workflow in two different case studies. The first step consisted in obtaining data and testing its veracity. These were followed by understanding what the data was trying to describe and how it was possible to implement it in the set objectives. For each dataset, two models were implemented so knowledge on the methodology was achieved. Finally, the different representations for the obtained results were studied.

Figure 8. Process diagram.

This thesis englobes all the methodologies, insights, challenges and situations I have faced and is structured as follows:

a) A main block concerning Electronic Health Records which are based on empirical data obtained from an open dataset. This data provides information acquisition through an analytics point of view. It means that approaches as prediction or patient flow are taken as variables. Also, with this information, it is possible to obtain an optimization of medical resources.

b) A secondary insight based on a personal interest to learn other applications of deep learning techniques through a basic Image classification using neural networks.

4 Methodology

In this thesis two methodologies are used; the Causal Impact package, of R, for a categorical dataset and a Convolutional neural network, developed under Python programming language, for an image dataset.

4.1 Causal Impact R Package

Causal Impact package implements an approach to estimate the casual effect of a designed intervention on a time series. An e*vent* can be any death, disease incidence, recovery or a designated experience of interest affecting an individual but to apply this package only one can be considered (e.g., treatment). *Time series* can be described in days, weeks, months or years [27]. The model is used to try and predict the counterfactual by constructing a Bayesian structural time-series model, which focuses on how a trial can change our opinion about the effect of an event [28].

Given a response time series and a set of control time series, the function constructs a time-series model, performs posterior interference on the counterfactual and returns a Causal Impact object that can be expressed as a table, verbal description or plot.

It is important to assume that the time series was not affected by the intervention. If the time series was affected there is going to be a false under-estimation of the true effect.

The package constructs a synthetic baseline for the post-intervention period based on a Bayesian structural time series model that incorporates multiple matching control markets as predictors, as well as other features of the time series. The package's behavior can be differentiated into steps. (i) Pre-screening, is carried out to find the best control markets for each market in the dataset using dynamic time warping. The user defines how many matches should be retained. (ii) The interference, where the control markets, identified in the previous step as predictors, fits into Bayesian structural time series model.

When assigning the control group, one must consider some features as non-random assignment of treatment based on observable measures of skill and past performance and the spillover effects on the comparison and pre-existing differences between the treatment.

The package returns a set of three panels (a) panel 1 shows the data and a counterfactual prediction for the post-treatment period, (b)panel 2 shows the difference between the observed data and the counterfactual prediction and (c) panel 3 shows the addition of point-wise contributions from the previous panel in a cumulative plot to prove the effect of the intervention [29].

4.2 Convolutional Neural Network

Generally, the application of CNN requires of a large training and testing datasets. The number of images comprised in those may differ on how different are the classes trying to be separated or if a normalization step will be added to the model. Often, researchers lack of a complete large dataset and it is possible to apply a process known as *Augmentation*. This process requires complete understanding of the data to know how aggressively a dataset can be augmented to perform the training step and therefore, reduce the number of errors produced by the model.

- Augmentation

Seven parameters were used to transform the original dataset to obtain more images and perform the training of the neural network.

As mentioned previously, filters are applied to images in order to transform them with a less complex structure. The gray scaling technique is employed through a 3x2 pattern array consisting of image dimensions and a coding number for a specific color filter,1 for gray or 3 for RGB. This filter may affect the contrast of the original image, so histogram equalization balances these changes by frequently increasing the intensity of the colors and providing more identifiable objects in the image.

To change the distribution of an image over the space reflection and rotation effects are applied. These make the image to move along the horizontal or vertical axis and within an angle distribution of 180 to -180 degrees. A widely used technique for image treatment is Gaussian Blur, which reduces noise and details into an object, but unfortunately this technique can also lead to image overfitting.

Since different objects can be found throughout the image, translation was defined by determining a fixed gray background color, so the object of interest was moved along the X and Y axis. To avoid image deformation the shearing technique displays every point horizontally or vertically proportionally according to its coordinates.

- Specific analysis

The different frameworks used in this case study can be classified in Low Level API, such as *Caffe* whose requirements and dependencies are more detailed, and High Level API, such as *TensorFlow Keras* which do not require a full understanding in Security Assertion Markup Language (SAML) by providing easy coding techniques and less code-actualizations after an update on the application [30],[31].

Google Colaboratory (Google Colab), is based on a *Jupyter Notebook* environment and provides fully runtime containing the leading libraries available nowadays and robust *Graphic Processor Unit* (GPU).

Google Colaboratory provides a faster training performance than Linux servers, according to an analysis of performance [32], by accepting TensorFlow Keras environment. The environment is characterized with faster prototyping, advanced research and advanced deployment. Its interface is way optimized for use-cases. The models are built by connecting configurable blocks together and by allowing the possibility to apply them in an easier form; new layers, loss functions and new changes on the basic structure [33].

Once a decision on which environment and framework to use is made, an article analysis of different studies concerning image classification on CNN was performed. In Table 1 it is possible to observe the features chosen by different researchers under the concept of "Leukemia" and "Cell Image" classification. It is necessary to point out that in this case it was necessary to make a more refined approach on what was aimed to be classified.

	Reference [34]	**Reference** [35]	**Reference** [36]
Emulator	Matlab	Matlab	Matlab
Data	113 Images	108 Images	1188 Images
Accuracy	Attributed to Operator's capabilities	96,43	96.6%
Neural Network	Feed-Forward Neural Network (FFNN)	CNN	CNN
Layers	5 layers with 2 hidden layers	4 layers (3 detecting 2 classifying =1)	7 (5 Feature Extraction, 2 for classification)
Matching Augmentation Parameters	No	N/A	Yes

Table 1. Previous article analysis to determine which parameters to compare and with which methodology[37],[38].

Different processes are widely used for image classification depending on the object of study. In this case a basic classification model was designed since more complex techniques would have required more time to be implemented.

Automation of a fully intelligent method to detect abnormal cells is still an open research area. Cellular disruption can happen in many ways and not all of them are sufficiently enough to train a Neural Network.

Basically, classification is characterized by iteratively processing the images in two steps. Beginning with a *training*, where the algorithm selects the optimal model parameters through a large number of imputed images belonging to different target groups. The model minimizes the performance of criteria by considering the error between training classifications and the true labels. This first step is followed a *prediction* process where the model itself must classify the input data into the different defined labels [38]–[41].

5 Case Study 1: A Brazilian Dataset for hospitalization frequency prediction

In this case study an open dataset is used to perform the relevant analytics. ORBDA is provided by the Brazilian Public Healthcare System and contains pseudo-anonymized patients across the different geographical states in Brazilian Federation Units. According to records and data from ORBDA, Brazilian Hospitals had available EHR information since 1991, but it was not an official information system until 2003. The idea was to avoid sensitive data for public analysis. ORBDA is structured in an Archetype model, where each node contains encoded information related to hospitalizations and complex procedures stored in different categories [42].

Besides these variables it is possible to observe date of discharge, healthcare unit, issue date, reason of discharge, age, gender, nationality, state, procedure, main diagnosis and secondary diagnosis, defined under different coding systems (as Date, ICD10/Local/CNES/SIGTAP and Numeric) [43].

Due to its complicated original architecture a pre-treatment was performed to obtain nine individual csv documents comprising diverse clinical and technical information.

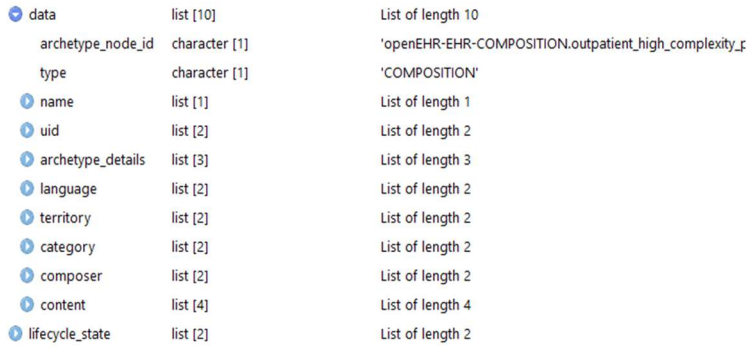


Figure 9. Overview of ORBDA dataset architecture before data treatment.

5.1 State of the Art

The Brazilian health system is based and organized by *Sistema Único de Saúde* (SUS), comprising a public model and universal coverage, which collaborates with *Agência Nacional de Vigilância Sanitária (*Anvisa). Anvisa provides control over all the products and services based on health such as drugs and nourishment.

In countries like Brazil treatment of long-standing illnesses requires high individual expenditure which contributes to population's decline in living standards reducing availability and affordability of basic needs such as food, housing or education.

The World Health Organization declared in 2010 that 100 million people are pushed towards poverty in countries where direct payment of health services is required.

During 2010, 1315 advertising campaigns towards pharma and food were distributed simultaneously.

In Brazil, arterial hypertension and diabetes occur to be the first cause of hospitalization in public health services and are related to other chronic diseases and complications. According to WHO, at least three interventions to prevent and deal with diabetes appear to reduce expenditure and health improvement.

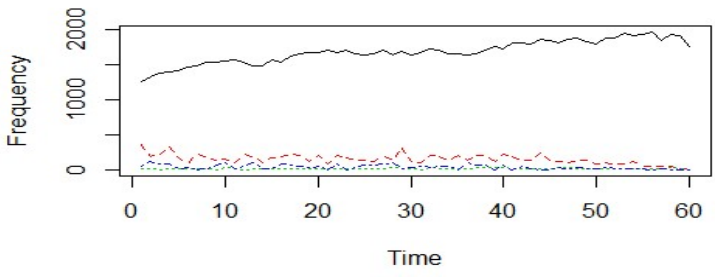
Diabetes, cancer or circulatory illnesses are burdensome for individuals, families and health services. With that, SUS and Anvisa organizations, together with the World Health Organization (WHO) sponsored in 2010 an initiative towards safer surgical interventions, which stood along with an advertising campaign towards healthier food supply [44].

For these reasons, in the following pages there are representations of how the hospitalization frequency was affected by scheduled Chemotherapy, Radiotherapy or dialysis treatments.

Causal Impact package requires a descriptive event to analyze how the response variable, general frequency of hospitalization, is seen affected through the advertising campaign. It also requires of a time series, which comprises a time-period of four years (from 2008 to 2012).

The time-series is represented by the model from 1, belonging to January of 2008, to 60, corresponding to December of 2012.

Before starting the model, a test to see if the variables were previously affected by the intervention was performed. (View Graph1)



Graph 1. Explicatory plot of all the predictive variables towards the response variables. X-axis as time-series (2008-01;2012-12) and Y-axis as frequency of hospitalizations. The upper black line describes total hospitalization frequency.

Data has been treated independently and incomplete time-series have been removed. With that, the amount of observations has decreased by a half to end up having useful information as shown in Table 2. The proportion of general hospitalizations and the predictive variables frequency is 1:3. With those two references is possible to proceed with the model.

Data	Renal dialysis	Radiotherapy	Chemotherapy	Cancer	General
Observations	71	60	94	117	60
Time-Period	2007-2012	2008-2012	2005-2012	2003-2012	2008-2012
Variables	Date of dialysis	Date of treatment initiation	Date of treatment initiation	Date of pathological identification	Date of hospitalization
Frequency	689	211	2415	1936	17436

Table 2. Observations after data cleaning.

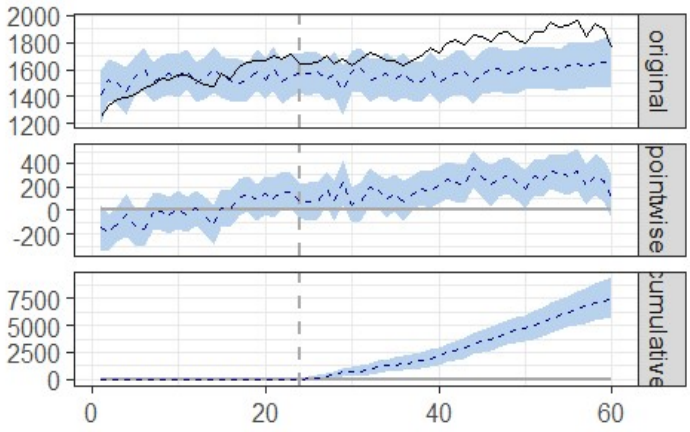
After analyzing the response under different prediction variables independently concluded on using them together to predict how the hospitalization frequency was affected under a pre-period of two years (2008-2009), period before the event whose effect is willed to be measured, and a post-period for testing of one year (2010).

Cancer descriptive variable was discarded due to its incomplete information during some months of 2012.

An average of 1781 real observations were performed, within a confidence interval of [1519,1924]. The effect augmented the value by 206 as average, producing a relative effect on the response variable of 13%. This means that the positive effect observed during the intervention was not due to random fluctuations and according to the confidence interval [+10%,+17%] the intervention period is statistically significant.

```
Posterior inference {CausalImpact}

                          Average        Cumulative
Actual                    1781           64116
Prediction (s.d.)         1575 (26)      56707 (946)
95% CI                    [1519, 1624]   [54696, 58475]

Absolute effect (s.d.)    206 (26)       7409 (946)
95% CI                    [157, 262]     [5641, 9420]

Relative effect (s.d.)    13% (1.7%)     13% (1.7%)
95% CI                    [9.9%, 17%]    [9.9%, 17%]

Posterior tail-area probability p:   0.00102
Posterior prob. of a causal effect:  99.89848%

For more details, type: summary(impact, "report")
```

Table 3. Summary of Causal Impact package. Average column reflects relevant statistics across time during the post intervention period, whereas the Cumulative column shows the total sum of those values.



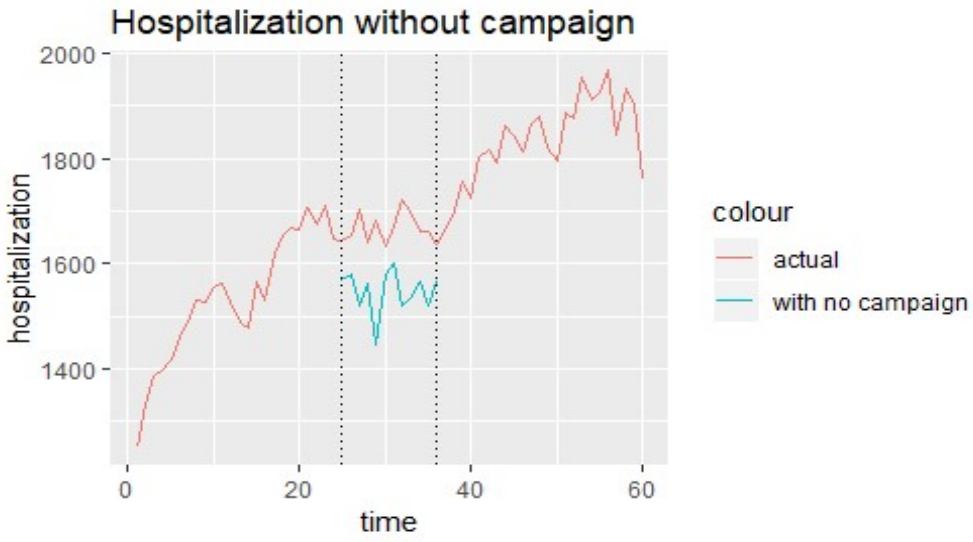
Graph 2. Descriptive plot of Causal Impact Package. X-axis represent time. (—) Actual values (----) Predicted values (▬) 95 % Confidence interval.

Graph 2 points a first panel where it is possible to observe how general frequency of hospitalizations would have behaved if the advertising campaign was not launched though a hypothetical estimation. In the second panel the differences between the observed data and the predicted values are plotted. Finally, the third panel displays the pointwise contributions of the previous panel from the time-period when the campaign was launched. This last panel provides visual information of an increase effect.

5.2 Results

If the effect on the prediction variable had no effect on the response variable, there would be only a 0,102% chance to see an effect as large as the one observed.

The probability of obtaining this effect without the advertaising campaign in 2010 is very small. Therefore, the causal effect can be considered statistically significant.



Graph 3. Explicative plot to understand package's results.

The difference on hospitalization frequencies of the actual values, present in the original data, towards the predicted values in 2010's period if the campaign was not released is shown in Graph 3. The boundaries within January of 2010 and December 2010 are delimitated by vertical doted lines to set a visual impact on the results obtained.

6 Case Study 2: An Image Dataset for automatic blood sample classification

This second case study aims to give an example of deep learning application through image classification. The AML/ALL Research Project was set by a group of volunteers who develop opensource solutions for clinical support.

ALL-IDB dataset was provided by the ALL-IDB initiative, which proposes a free dataset of microscopic images of blood samples where each image has been classified and processed by expert oncologists for Acute Lymphoblastic Leukemia. This dataset is only meant to be used for processing techniques and improvement of classification models. Related approaches can be found in its website [45].

It is possible to distinguish between two sets, ALL_IDB1 build for segmentation algorithms, classification systems and image processing methods. ALL_IDB1 is composed of 108 images separated in 59 negative diagnosed images and 49 images containing at least one blast cell. For the study, 10 random images of the negative folder were discarded for homogenization. This set also contains the coordinates of the centroids of the blast cell.

The second set, ALL_IDB2, designed to test the performance of classification system is composed of 260 homogenized images (130 positively classified images and 130 negative images containing at least one blast cell). This set is an adaptation of the previous one where the image dimensions have been reset to fit and center cells of interest.

6.1.1 Leukemia

Leukemia is a blood cell cancer characterized by the overproduction of immature white blood cells. According to its speed of progression and by the type of white blood cells affected, the following classification is applied: acute or chronic, and, lymphocytic and myelogenous, respectively.

The possible combinations of behavior and condition lead to the following nomenclature: acute lymphoblastic, acute myeloid, chronic lymphocytic and chronic myeloid.

Acute Lymphoblastic Leukemia (ALL) can also be sub-classified (according to the French-American-British classification systems of hematologic diseases) in; L1 – affecting 25-30%of adults and 85% of children, L2 – affecting 70% of adults and 14% of children and L3 – affecting 1 to 2% of the population. This inner classification is based on the morphology of the white blood cells and its original cellular precursor (B- precursor, T – precursor and B-cell).

6.2 State of the Art

After achieving knowledge on the architecture of classification models, Kaggle solved exercises were reproduced. Kaggle is an online platform where data scientists explore and learn different Machine learning applications through public datasets and active competitions to improve existing solutions. Two Kaggle examples were selected, both build on a TensorFlow Keras environment.

The first one being reproduced was a digit recognizer competition, where it was possible to deploy computational vision fundamentals from a csv format dataset containing pixel information.

Digit recognizer Kaggle was not completely useful since the dataset available from ALL-IDB was image based. For that, a second competition was reproduced, concerning images of cats and dogs, which provided the basic structure of the model used in this Case Study [46]–[48].

Dataset treatment was performed since the available samples to train the model were not enough. Data reached up to 1.567 images and 4.160 images for ALL_IDB1 and ALL_IDB2 respectively.

Three articles were analyzed to select the most appropriate parameters of the desired model (Recall Table 1).

```
from keras.models import Sequential
from keras.layers import Dense, Conv2D, MaxPooling2D, Flatten
from keras.models import Sequential
from keras.layers import Conv2D

# create model
model = Sequential()
model.add(Conv2D(1, (5,5), padding='valid', input_shape=(50, 50, 3)))
model.add(MaxPooling2D(pool_size=(1, 1), strides = (2,2)))
model.add(Conv2D(1, (5,5), padding='valid'))
model.add(MaxPooling2D(pool_size=(1, 1), strides = (2,2)))
model.add(Conv2D(1, (5,5), padding='valid'))
model.add(Flatten())
model.add(Dense(2, activation='softmax'))
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
# summarize model
model.summary()
```

Figure 10. CNN model architecture [35].

Sequential model allows to slack sequential layers in order through the network from input to output.

The first convolutional layer processes the two-dimensions input images, with 1 channel. The input will have a window moving along with a size of 5x5, and it is in here where the padding is defined with a zero-padding instruction. This means, the image will be filled with zeros to fit the dimensions of the input. The next line defines the addition of a two-dimensional max pooling layer, which will choose the highest value of pixels within the filter window. This window will be then moved one pixel along X axis and one pixel along Y axis.

The stride argument can either be defined in the first convolutional layer or in the max pooling layer. In this case the stride indicates to the model to keep two pixels between each sample.

This sequence is repeated to build a three layers model.

Finally, to fit and connect all layers in the model a Flatten function is defined so the output of one layer can be the input of the next one.

Now, each layer has to be connected with the previous one and, for that, a Dense function is used. This function is set with a parameter of 2, meaning that the two last layers will be fully connected. Each one will be activated through a Softmax function

The Softmax function transforms the input values into a vector and normalizes it following a probability distribution. The output will now range between [0,1] avoiding binary classification.

- ALL_IDB1

The same model was applied to the dataset in two different ways (i) using original data for training and predicting and (ii) using augmented data for training and original images for prediction, in both, data was split into 75% for training and 25% to validate the model.

- ALL_IDB2

In this case, image dimensions were doubled, producing a noticeable change on the model augmenting approximately three times the trainable parameters. Filter size, zero-padding, pool size and strides were kept untouched to avoid more changes on the model. The same to two tests, as in the other dataset, were developed but due to the overall size of the dataset, data was split into 40% for training and 30% for model validation.

Preprocessing parameters were applied in both original data, same parameters used to perform the augmentation, but those only altered the shape and filter, not the total amount, of images. This technique was implemented to increase the input parameters the CNN had to learn.

6.3 Results

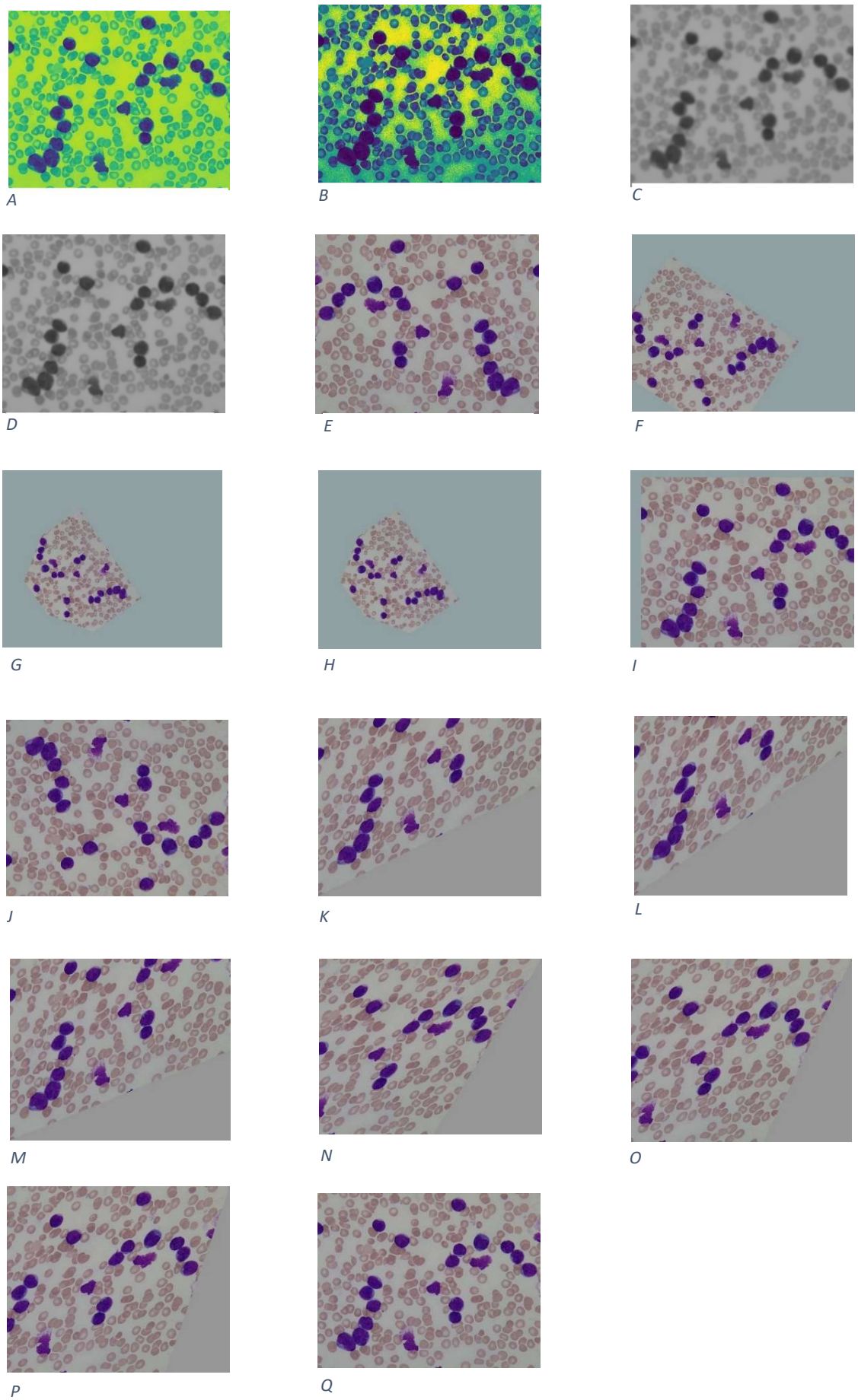


Figure 11. Augmented output of a Positive image from the ALL_IDB1 dataset. (A) Gray scaling to RGB (B) Histogram equalization (C) Gaussian Blurred (D) Gaussian (E) Reflection – as horizontal flip (F-H) Rotation (I) Translation (J) Reflection – as vertical flip (K-M) Shear towards the X-axis (N-P) Shear towards the Y-axis (Q) Original Image.

Figure 11 shows the output of the augmentation parameters of a single image to obtain a greater amount to images to train the model.

Accuracy and loss were used to represent model's process during the training step under 100 Epochs, or in other words, to represent the walk-through of the dataset by the neural network divided in 100 independent steps. The model takes one image at a time to train the NN.

When comparing training process' behavior between Original and Augmented data for ALL_IDB1 , is possible to discard overfitting, since this effect is observable when both training and validation start diverging. In this case, both match over iterations meaning the same pattern is found. Through Figure 13, Model Accuracy 's plot, is possible to define underfitting of validation progress while its Model Loss represents a typical behavior diverging after approximately 50 epochs. This divergence assesses when the model could have been stopped instead of letting it perform the hole iterations.

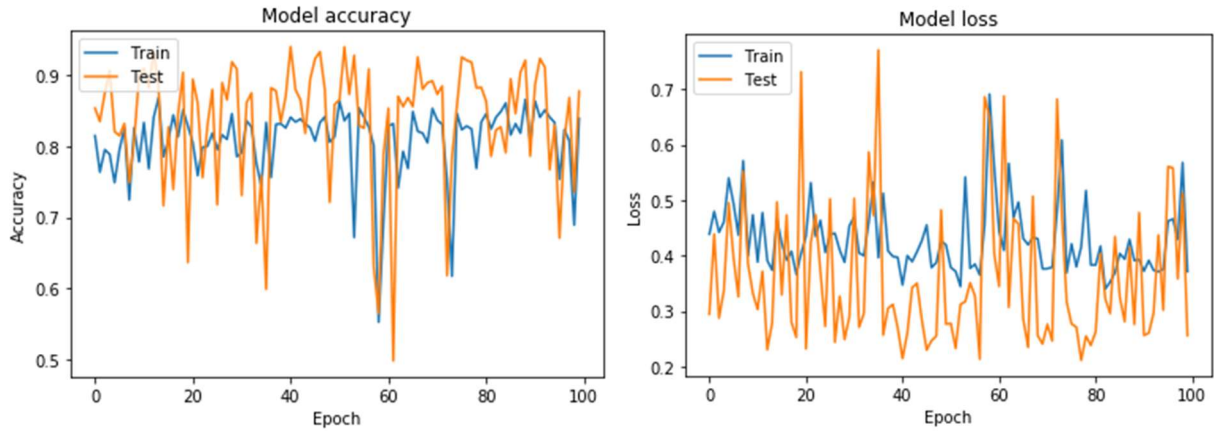


Figure 12.Behavior of trained model using ALL_IDB1 dataset- Original data.

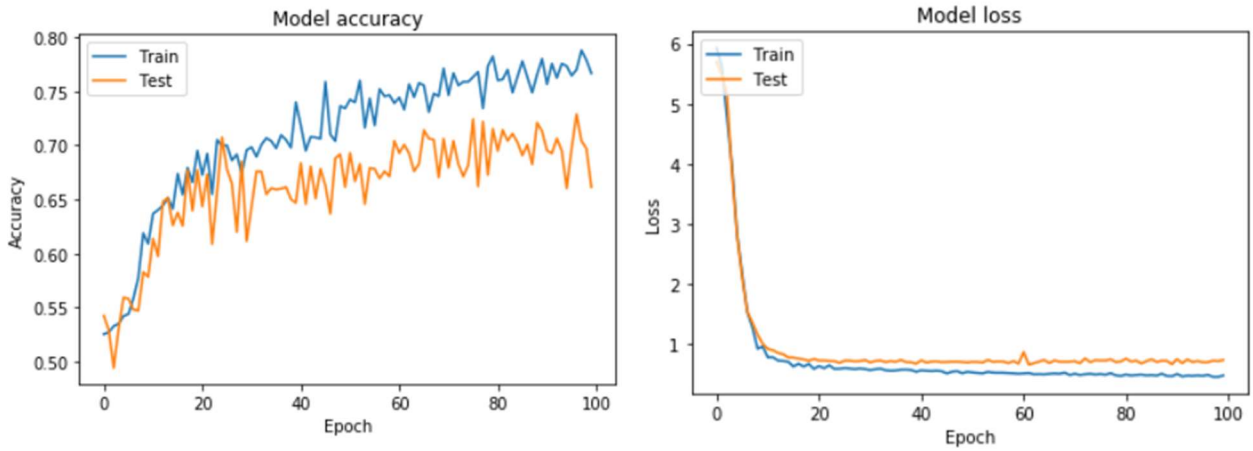


Figure 13. Behavior of trained model using ALL_IDB1 dataset- Augmented data.

When using the ALL_IDB2 dataset to perform the training step, different Epoch parameters were set for each Original data and Augmented data – 100 and 200, respectively. Therefore, it causes difficulties for visual comparison but is possible to easily determine that Augmented data might have a wrong parameter thus its model accuracy has a wide fluctuation rate and it is not a good performance indicative. Whereas the Original data, besides having some fluctuations, provides a clearer performance.

The analysis of performance when using Original data shows that the model has not over-learned the training set showing overlapping in most of the epochs and it is confirmed by the loss plot, where there are similar values between training loss and validation loss.

In the other hand, when using augmented data, accuracy plot bounces off training and validation sets but no clear conclusion on performance can be taken since the fluctuation per epoch suggests that longer steps are needed during the training process of the model. Again, model's loss confirms the necessity of more training data by the curves' behavior which starts with a perfect match between training loss and validation loss but differ too soon in time representing overfitting parameters.

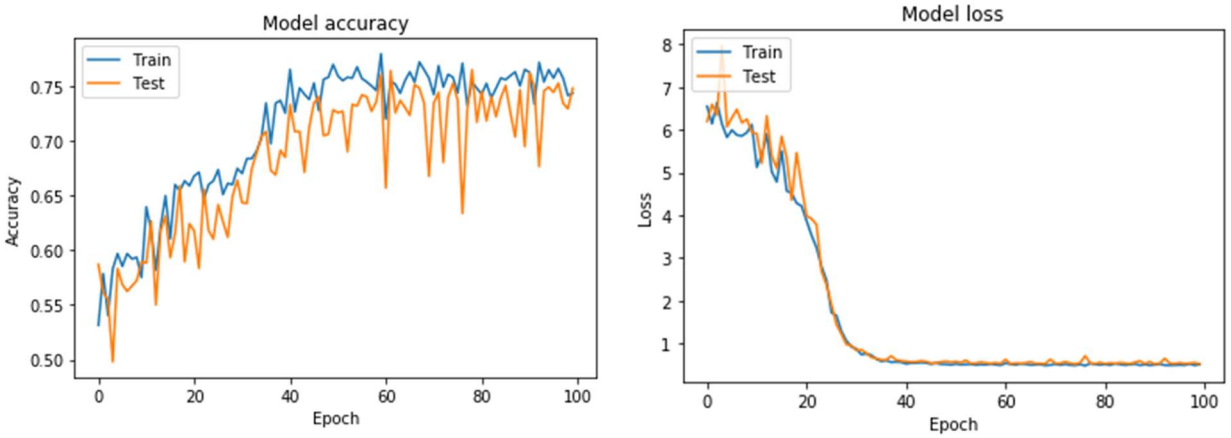


Figure 15. Behavior of trained model using ALL_IDB2 - Original Data.

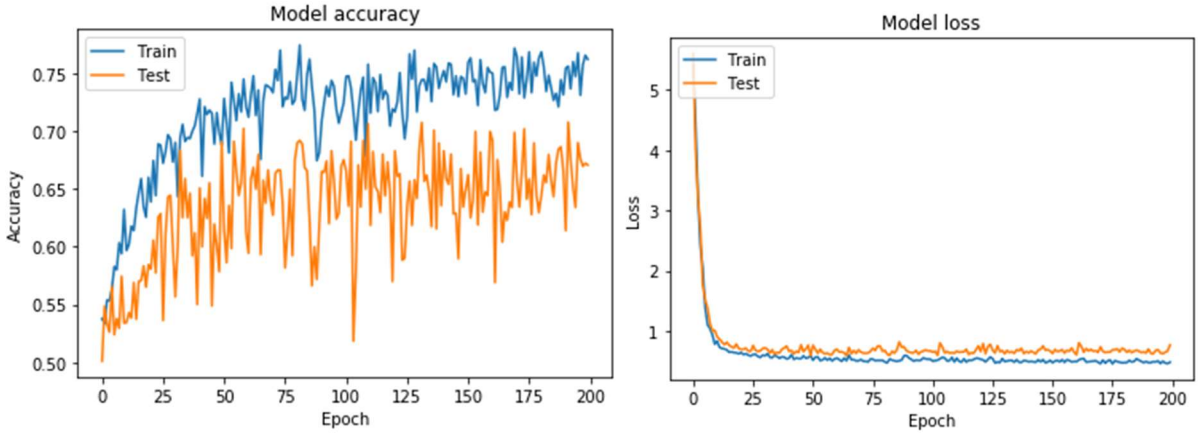


Figure 14. Behavior of trained model using ALL_IDB2 dataset - Augmented data.

Two different evaluating metrics for prediction performance where calculated per each case based on binary classification of the image as True Positives, False Positives, True Negatives and False Negatives.

Receiver operating characteristic curve (ROC) was used to represent the classification rate between True Positive and False Positive images, where the Area Under the Curve (AUC) determinates the friability of the predictions in a scale of 0 to 1.

For both parameters the based on Original images for training and prediction were discarded. Results show 0,865 AUC for ALL_IDB1 dataset and 0,295 AUC for ALL_IDB2.

To get more information on how images were classified along False Positives and False Negatives confusion matrixes were plotted and ensured, through a F1 score (perfect precision and recall), the AUC values to get a conclusion.

A

	Precision	*Recall*	*F1-score*	*Support*
1	0,83	0,39	0,53	49
0	0,60	0,92	0,73	49

C

	POSITIVE	**NEGATIVE**
POSITIVE	19	30
NEGATIVE	4	45

B

	Precision	*Recall*	*F1-score*	*Support*
1	0,47	0,81	0,59	130
0	0,29	0,08	0,12	130

D

	POSITIVE	**NEGATIVE**
POSITIVE	105	25
NEGATIVE	120	10

Table 4. (A) Shows the metrics of the classification matrix for ALL_IDB1 with an Accuracy of 0,663. (B) Shows the same parameters for ALL_IDB2 with an Accuracy of 0,442. (C-D) Represent the number of images classified, where columns are predicted values and rows actual values.

Neither of the sets for the model are useful to give a positive result affirmation, since the problem might not relay on the data, all indicators point to the model's structure. It could also be due to the lack of images to train the model. Usually, the number of un-processed images to perform this step is in the order of ten thousand.

Statistical tests to measure the classification represent opposite performance by having major Sensitivity over negative samples in the ALL_IDB1 and by contrast, in ALL_IDB2, the better Sensitivity is given in positive samples. When looking the relation between precision and sensitivity, F1 Score, the results are also opposed where ALL_IDB1 has a higher proportion of negative classification and ALL_IDB2 has it over positive classification.

The classification error per dataset is of 34% on the ALL_IDB1 and of 56% on the ALL_IDB2.

By these parameters and their difference within datasets, is possible to state that if given more training time the ALL_IDB2 would have had a better performance leading to a higher accuracy value and a decrease in the classification error.

7 Conclusions

This thesis has reveled the importance of pre-processing techniques before going deeper into the testing stage.

- Visualization techniques are required through all the steps of data treatment, allowing an easy understanding on how datasets narrow and change over the process.
- Opensource datasets, with interesting objects to perform analytics are difficult to find since most of them have missing data or have invalid inputs. This can be a consequence of the relatively new technique to store Big Data.
- The followed methodologies for clinical decision support have no intention to replace doctor's performance nor opinion, but effectiveness with different computer vision system can improve the time cost of data analysis and sample classification.

Electronic Health Records require an inner insight on how data is gathered for an easier treatment. Further steps in this area could be creating an even more standardized and universal model accessible for everyone following the user-requirements. Cloud based data is easy to access but difficult to avoid de-identification.

- The data cleaning step narrows down the overall number of objects and therefore synthetizes too much the dataset, avoiding the deployment of a wide range of tests.
- The Causal Impact package is a simple prediction algorithm to determinate the effect of an event over response variables. More accurate algorithms and packages exist to perform these kinds of predictions.
- The advertising campaign released on 2010 in Brazil regarding pharma and food has an impact over the general hospitalization frequencies, when taking as descriptive variables Chemotherapy hospitalization frequency, Radiotherapy hospitalization frequency and Renal dialysis hospitalization frequency.

Convolutional Neural Networks can capture relevant features from an image with a similar performance to a human brain.

- Further stages of CNN and other NN need to be taken as this test was performed with very low training data and the augmentation process is not completely reliable in real world, but the accuracy obtained with the measurements and different environments plus the capacity to replicate a paper, leads to a full understanding on how this procedure can be optimized.
- Higher Batch sizes can be implemented to increase the performance of ALL_IDB2, since is the one having a decent number of images to train the model, and therefore improve its Sensitivity and Precision over negative samples in the predictive stage.
- Further methodologies after fixing this model can be studied, as segmentation or feature extraction from the ALL_IDB dataset.

8 Personal valuation

The development of a final degree project has allowed me to get a new insight of science related areas. Consultancy for research and development projects is a competitive and challenging work by always willing to provide the best and newest approach towards an emerging idea.

This experience has been by far the most rewarding compared to other internships and project developments over the years spent in academic formation.

I have been able to open my boundaries and knowledge, which is in my opinion what a Biotechnologist should do over a professional career.

Regarding the project, I have revived the early years in the career by experiencing the learning performance curve. Where the beginning is clear and straight, but at some point, this exponential performance starts to become flatter over time. This moment is where personal growth is achieved by keeping the will to improve.

Now, I closure a huge personal period with high curiosity on the adaptation of technology towards biological systems, not only in pharmaceutical industries nor environmental areas but in how these techniques can improve society's difficultness and performances of all what health care systems englobe.

9 Bibliography

[1] J. S. Ward and A. Barker, “Undefined By Data: A Survey of Big Data Definitions,” 2013.

[2] E. X. In and S. E. D. Ata, “Big Data Analytics,” vol. 36, no. 4, pp. 1165–1188, 2018.

[3] E. U. Jiménez, *Análisis de datos: series temporales y análisis multivariante*. AC, 1995.

[4] M. J. Campbell and D. Machin, *Medical Statistics: A Commonsense Approach*, 3rd, illustr ed. Wiley, 1999, 1999.

[5] A. Makam, O. Nguyen, B. Moore, Y. Ma, and R. Amarasingham, “Identifying patients with diabetes and the earliest date of diagnosis,” *BMC Med. Inform. Decis. Mak.*, 2013.

[6] W. B. Lober *et al.*, “Barriers to the use of a personal health record by an elderly population.,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, pp. 514–8, 2006.

[7] F. Meng, K. L. Teow, C. K. Ooi, B. H. Heng, and S. Y. Tay, “Analysis of patient waiting time governed by a generic maximum waiting time policy with general phase-type approximations,” *Health Care Manag. Sci.*, vol. 18, no. 3, pp. 267–278, 2015.

[8] L. Zhao and B. Lie, “Modeling and Simulation of Patient Flow in Hospitals for Resource Utilization,” *SNE Simul. Notes Eur.*, vol. 20, no. 2, pp. 41–50, 2017.

[9] A. Roehrs, C. A. da Costa, and R. da Rosa Righi, “OmniPHR: A distributed architecture model to integrate personal health records,” *J. Biomed. Inform.*, vol. 71, pp. 70–81, 2017.

[10] A. Mense and B. Blobel, “HL7 Standards and Components to Support Implementation of the European General Data Protection Regulation (GDPR),” *Eur. J. Biomed. Informatics*, vol. 13, no. 1, pp. 27–33, 2019.

[11] K. E. Schrampfer and W. J. Whitsitt, “Establishing the Effectiveness ofProcedural Interventions The Limited Role ofRandomized Trials,” no. 289, pp. 5–6, 1988.

[12] F. Khennou, Y. I. Khamlichi, and N. E. H. Chaoui, “Improving the use of big data analytics within electronic health records: A case study baseD OpenEHR,” *Procedia Comput. Sci.*, vol. 127, pp. 60–68, 2018.

[13] L. Gligic, A. Kormilitzin, P. Goldberg, and A. Nevado-Holgado, “Named Entity Recognition in Electronic Health Records Using Transfer Learning Bootstrapped Neural Networks,” *ArXiv e-prints*, 2019.

[14] A. H. Nordo *et al.*, “Use of EHRs data for clinical research: Historical progress and current applications,” *Learn. Heal. Syst.*, vol. 3, no. 1, pp. 1–9, 2019.

[15] E. Ammenwerth, S. Lannig, A. Hörbst, G. Muller, and P. Schnell-Inderst, “Adult patient access to electronic health records,” *Cochrane Database Syst. Rev.*, vol. 2017, no. 6, 2017.

[16] P. Y. Wu, C. W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, “-Omic and Electronic Health Record Big Data Analytics for Precision Medicine,” *IEEE*

Trans. Biomed. Eng., vol. 64, no. 2, pp. 263–273, 2017.

[17] F. Chollet, *Deep Learning with Python*. 2018.

[18] I. Goodfellow, Y. Benigo, and Courville Aaron, *Deep Learning (Adaptive Computation and Machine Learning series): Ian Goodfellow, Yoshua Bengio, Aaron Courville: 9780262035613: Amazon.com: Books*. MIT Press, 2016.

[19] M. T. Jones, *Artificial Intelligence: A Systems Approach*. Jones & Bartlett Learning, 2015.

[20] E. Alpaydin, *Introduction to Machine Learning*, Second. Istambul: MIT Press, 2009.

[21] G. Zaccone, M. R. Karim, and A. Menshawy, *Deep Learning with TensorFlow*. Packt Publishing Ltd, 2017.

[22] D. Ravi *et al.*, "Deep Learning for Health Informatics," *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 1, pp. 4–21, Jan. 2017.

[23] D. Chicco, P. Sadowski, and P. Baldi, "Deep autoencoder neural networks for gene ontology annotation predictions," pp. 533–540, 2014.

[24] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F. Y. Wang, "Generative adversarial networks: Introduction and outlook," *IEEE/CAA J. Autom. Sin.*, vol. 4, no. 4, pp. 588–598, 2017.

[25] R. Sutton and A. Barto, *Reinforcement Learning: An introduction*. MIT Press, 2018.

[26] B. Liu, S. Wang, R. Long, and K. C. Chou, "IRSpot-EL: Identify recombination spots with an ensemble learning approach," *Bioinformatics*, vol. 33, no. 1, pp. 35–41, 2017.

[27] D. G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text*, Illustrate. Springer Science & Business Media, 2013.

[28] D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles, *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Illustrate. John Wiley & Sons, 2004.

[29] Brodersen et al, "CausalImpact 1.2.1, Annals of Applied Statistics." 2015.

[30] K. Chellapilla *et al.*, "High Performance Convolutional Neural Networks for Document Processing To cite this version : HAL Id : inria-00112631 High Performance Convolutional Neural Networks for Document Processing," 2006.

[31] Y. Jia, "Caffe," 2014. [Online]. Available: https://caffe.berkeleyvision.org/.

[32] T. Carneiro, R. V. M. Da Nobrega, T. Nepomuceno, G. Bin Bian, V. H. C. De Albuquerque, and P. P. R. Filho, "Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications," *IEEE Access*, vol. 6, pp. 61677–61685, 2018.

[33] the T. logo and any related TensorFlow and marks are trademarks of G. Inc., "Tensorflow." [Online]. Available: https://www.tensorflow.org/guide.

[34] V. Piuri and F. Scotti, "Morphological classification of blood leucocytes by microscope images," no. July, pp. 103–108, 2005.

[35] T. TTP, G. N. Pham, J.-H. Park, K.-S. Moon, S.-H. Lee, and K.-R. Kwon, “Acute Leukemia Classification Using Convolution Neural Network in Clinical Decision Support System,” pp. 49–53, 2017.

[36] T. T. P. Thanh, C. Vununu, S. Atoev, S.-H. Lee, and K.-R. Kwon, “Leukemia Blood Cell Image Classification Using Convolutional Neural Network,” *Int. J. Comput. Theory Eng.*, vol. 10, no. 2, pp. 54–58, 2019.

[37] R. Donida, V. Piuri, and F. Scotti, “ALL-IDB : The acute Lymphoblastic Leukemia image database for image processing,” *Ieee Int. Conf. Image Process.*, pp. 2089–2092, 2011.

[38] F. Scotti, “Robust segmentation and measurements techniques of white cells in blood microscope images,” *Conf. Rec. - IEEE Instrum. Meas. Technol. Conf.*, no. April, pp. 43–48, 2006.

[39] S. Alférez, A. Merino, A. Acevedo, L. Puigví, and J. Rodellar, “Color clustering segmentation framework for image analysis of malignant lymphoid cells in peripheral blood,” *Med. Biol. Eng. Comput.*, no. Cll, 2019.

[40] J. Rodellar, S. Alférez, A. Acevedo, A. Molina, and A. Merino, “Image processing and machine learning in the morphological analysis of blood cells,” *Int. J. Lab. Hematol.*, vol. 40, no. January, pp. 46–53, 2018.

[41] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*, Illustrate. Springer, 2008.

[42] J. Pires Machado, M. Martins, and I. da Costa Leite, “Qualidade das bases de dados hospitalares no Brasil: alguns elementos,” *Rev. Bras. Epidemiol.*, vol. 19, no. 3, pp. 567–581, 2016.

[43] D. Teodoro, E. Sundvall, M. J. Junior, P. Ruch, and S. M. Freire, “ORBDA: An openEHR benchmark dataset for performance assessment of electronic health record servers,” *PLoS One*, vol. 13, no. 1, pp. 1–22, 2018.

[44] J. F. Kell, *Plan de Acciones Estratégicas para el Enfrentamiento de las Enfermedades no Transmisibles (ENT) en Brasil*. 2011.

[45] F. Scotti, R. Donida, and P. Vincenzo, “Acute Lymphoblastic Leukemia Image Database for Image Processing,” 2010. [Online]. Available: https://homes.di.unimi.it/scotti/all/. [Accessed: 20-Sep-2001].

[46] “Stackoverflow,” 2008. [Online]. Available: https://stackoverflow.com/.

[47] Github inc, “Github,” 2007. [Online]. Available: https://github.com/EstelaCaEs.

[48] B. Hamner and G. Anthony, “Kaggle.” [Online]. Available: https://www.kaggle.com/.