Choosing a machine learning model for breast cancer detection in images

Ricardo Avila Hernandez
La Salle University in Mexico City
Mexico

Kevin Ricardo Rossell Mendoza
La Salle University in Mexico City
Mexico

Josue Alejandro Soto Mora
La Salle University in Mexico City
Mexico

Research article
Received: June 1, 2020
Accepted: September 28, 2020
Available online: December 21, 2020

Abstract

Machine Learning comprises a wide range of models aimed at solving real life problems using supervised and unsupervised algorithms capable of finding even the finest causalities and correlations between any given phenomena portrayed in data. Given the current extraordinary software capabilities, we can exploit this tool in practically any field – Oncology. For instance, a medical speciality which focuses on Cancer treatment can make use of these models to provide a more accurate diagnosis when it comes to Breast Cancer Detection. In this article we delve into a catalogue of Machine Learning models and discuss their effectiveness through specific criteria in order to choose the most suitable one for this problem. The Analytic Hierarchy Process displayed conclusive results assigning to the Random Forest the highest scores in each one of the analyses employed, over 10% better than the Logistic Regression, the second highest evaluated model in the overall analysis. The models we developed with data describing different features of different breast tumour nuclei, therefore, for another type of data results may differ.

Key words: Breast Cancer, Classification, Decision-Making Theory, Machine Learning, Supervised Learning.

Búsqueda del mejor modelo de aprendizaje de máquina para detección de cáncer de mama a partir de imágenes

Resumen

El Aprendizaje de Máquina comprende una amplia gama de modelos que pretenden resolver problemas mediante algoritmos Supervisados y No Supervisados, éstos son capaces de encontrar relaciones causales y correlaciones que pueden pasar desapercibidas por otros métodos. Dados los avances tecnológicos, en concreto software, se pueden utilizar estas herramientas a varias disciplinas, como lo es Oncología. Ésta es una especialidad médica que se enfoca en el Cáncer y puede ser beneficiada al utilizar estos modelos para detección de Cáncer de Mama. En el presente artículo, exploramos un catálogo de modelos de Aprendizaje de Máquina Supervisados y estudiamos su eficiencia mediante diferentes criterios, para encontrar el más adecuado para resolver este problema. El método Analytic Hierarchy Process brindó resultados claros, mediante el cual se asignó al Random Forest como el mejor modelo en los tres análisis que se llevaron a cabo; con una calificación más de

10% más alta que el segundo mejor modelo, la Regresión Logística. Estos modelos fueron entrenados con datos sobre diferentes células de tumores en mamas, por lo que, con diferentes datos, los resultados pueden variar.

Palabras claves: Cáncer de Mama, Clasificación, Teoría de las Decisiones, Aprendizaje de Máquina, Aprendizaje Supervisado.

1 Introduction

The word cancer today still evokes great fears about a silent killer that creeps towards us without being publicized. Cancer is a genetic disease caused by changes in genes that control the way our cells work, especially the way they grow and divide (American Cancer Society, 2016). They can originate in any part of the body. It all begins when cells grow uncontrollably beyond normal cells, making it difficult for the body to function optimally (American Cancer Society, 2016). Each year, cancer affects more than 10 million people worldwide and kills around 6 million people. Without an effective control of this disease, these figures will continue growing significantly, and the most marked increase will occur in developing countries (WHO, 2004). Breast Cancer (BC) is the second most common cancer in the world. United States Cancer Society report showed that roughly 1.3 million American women were diagnosed with BC, resulted in half a million deaths each year because of malignancies (American Cancer Society, 2012). Projections estimate that upwards of 20 million BC cases will be recorded worldwide in the year 2030. Figure 1 shows the global incidence of the breast cancer in thousands. As we can see, the most affected female populations are reported from the developed countries in North America, Europe and Australia. This can be due to developed health care system, and consequently higher usage of BC tests in those countries. Many types of cancer have a high chance of cure if diagnosed early and treated adequately (WHO, 2018a). Between 30-50% of cancers can currently be prevented by avoiding risk factors. Therefore, it is very important to have well developed the BC prevention system.

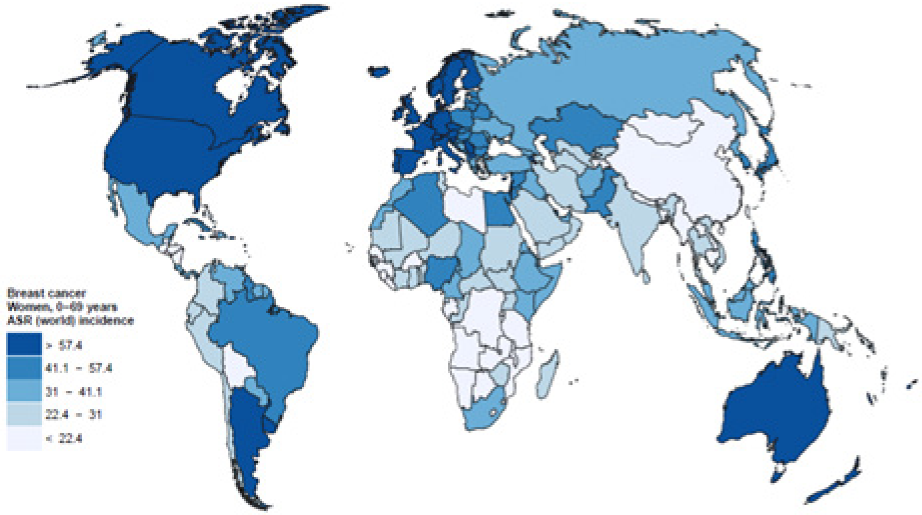

Figure 1: Breast Cancer Global Incidence in thousands. A darker shade of blue reflects greater manifestation of Breast Cancer in a given country (Anderson, 2014)

In Mexico, breast cancer has been an increasing cause of death in the recent years. From 2011 to 2016, the death toll of breast cancer rose from 13.92 to 16.12 deaths per 100 thousand women older than 20 (INEGI, 2018). Moreover, in most cases, breast cancer occurrences are chronic-degenerative conditions. Therefore, the breast cancer incidences and mortality rates tend to increase with age. It is estimated that by 2025 the cancer cases in Mexico will increase by 50%, increasing from 147 thousand to more than 220 thousand new cases (TAC, 2017). Figure 2 shows percentages of new cases by type of cancer in Mexico. The situation in Mexico is similar as in the World as the breast cancer was the most common type in 2018, regardless the gender and patient's age.

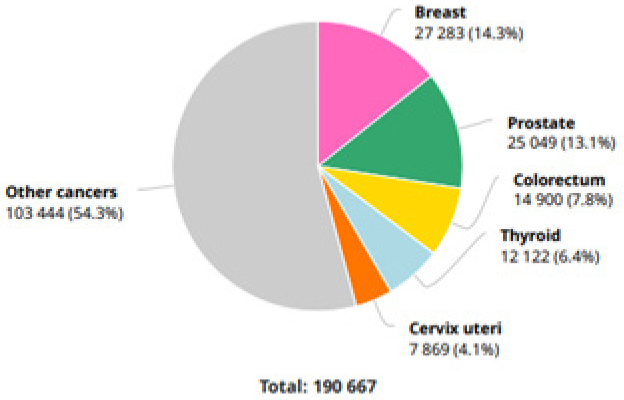

Figure 2: Number of new cases in Mexico in 2018, both sexes, all ages. Breast was the most frequent type of cancer, followed by Prostate Cancer. (WHO, 2018b)

Over the last few years, research related to diseases has been rapidly evolving, and the research in cancer has not been an exception. One of the key aspects of the breast cancer is the early detection, which can considerably improve the outcomes of breast cancer (Silverio, 2020). Women between the ages of 40 and 50 are often advised to go for a mammography screening. Although these measures contribute to early detection, there can still be cases of false negatives results due to the difficulty to correctly interpret the screening images (Silverio, 2020), and the performance of even the best clinicians leaves room for improvement (Elmore et al., 2009). Moreover, false positives can lead to patient anxiety (Tosteson et al., 2014), unnecessary follow-up and invasive diagnostic procedures. Cancers that are missed at screening may not be identified until they are more advanced and less amenable to treatment (Houssami & Hunter, 2017). Therefore, scientists have applied different methods either to find types of cancer before they cause symptoms or to distinguish malignant breast masses from benign ones.

The evolution of new technologies applied to medicine has collected large amounts of data on cancer, allowing accurate prediction of this disease (Kourou et al., 2015). In the face of this, Machine Learning (ML) has become a quite useful and significant tool to discover and identify patterns and relationships between variables. In case of complex data sets, machine learning is able to effectively predict the future outcomes of a type of cancer. The branch of ML that can tackle this problem is the Supervised Learning, where the presence of the outcome variable is to guide the learning process (Hastie, Tibshirani & Friedman, 2008). As a further matter, Supervised Learning encompasses Regression and Classification, where the former involves predicting a quantitative output and, the latter, predicting a qualitative output (Hastie, Tibshirani & Friedman, 2008). The outcome variable in the case of predicting the breast cancer is a categorical variable in a dataset indicating whether the patient has a malignant or benign breast mass. Specifically, Mangasarian, Street and Wolberg (1994) described the application of breast cancer diagnosis using the following steps: they took a minimally invasive sample of fluid from the patient's breast, from it 10 features where computed for each nucleus: area, radius, perimeter, symmetry, number and size of concavities, fractal dimension, compactness, smoothness and texture. They achieved great results distinguishing between benign lumps from malignant ones, also they gave an outlook for the patients who had the cancer surgically removed, predicting whether their tumour will recur or not. More recently, the attention in the BC diagnosis focused on application of Artificial Intelligence (McKinney et al., 2020). In this case, an Artificial Intelligence project funded by Google outperformed doctors in flagging breast cancer. They applied Computer-Vision to mammograms in this diagnosis. Computer-Vision is a trending branch of ML that focuses on describing the world that we see in one or more images and to reconstruct its properties, such as shape, illumination, and colour distributions (Szeliski, 2010). Although Computer-aided detection software for mammography was introduced in the 1990s, and in spite of early promise, that generation of software failed to improve the performance of readers in real-world settings as noted by Fenton et al. (2007), Lehman et al. (2015) and Kohli and Jha (2018), opposed to the Google project were they even reduced the workload of the human reader of mammographies by 88%.

Machine Learning has been applied to a wide variety of industries, some applications are predictive maintenance of equipment, recruiting employees, customer experience, finance, customer service and more (Taulli, 2019). ML pros and cons vary by a model, peculiarly the application of ML in medical diagnosis can struggle because people like to visit a doctor, people do not understand enough about how ML get things done or patients do not like to share their personal data. Conversely, ML could save time by avoiding a trip to the doctor and, what is more, AI powered doctor can be available whenever needed, ML can assess vast amounts of relevant information to discover important patterns (Accenture, 2018).

One problem regarding ML models is that the standard process consists in training several models, then evaluate these models and choose the most precise one. As Brian D. Ripley said, "Machine Learning is statistics minus any checking of models and assumptions" (R Project, 2016). To evaluate the models, some ready-made metrics are available. For the regression analysis, we can use the pseudo-R2, Median Absolute Percentage Error (MAPE), Root Mean Squared Log Error (RMSLE), among others, whereas for the classification part of the diagnosis can be used Accuracy, Precision, Negative Predictive Value, Sensitivity, Specificity, F1 Score, Area Under de Curve of Receiver Operating Characteristic (AUROC), among others (Lantz, 2015). Despite having lots of metrics to assess model performance, usually only one is considered for choosing the best model. However, different metrics help us to evaluate distinct characteristics of the compared models. Therefore, the objective of this article is to apply Decision-making Theory to choose a ML model for distinguishing between benign lumps from malignant ones, being this a more holistic approach where we consider more than one metric and hopefully achieve a better performance overall.

According to Khan (2016), The most popular ML models are the following:

- *Probabilistic outcomes*: Naive Bayes, Gaussian Mixture Models, Logistic regression
- *Linear Classification problems*: Decision Trees, KNN, SVM with Linear Kernel
- *Sequential Modelling / Time Series*: Hidden Markov Model, Recurrent Neural Nets

- *Feature Learning*: Auto Encoders, Convolution Neural Networks and other Deep Net Architectures
- *Non-Linear Classification Problems*: Random Forest, SVM with RBF/Polynomial Kernels, Neural Networks
- *Clustering*: K-Means and its variants, Hierarchical Clustering methods

2 Materials and Methods

Analytic Hierarchy Process (AHP)

Analytic Hierarchy Process was developed by Saaty (1977) and works with both qualitative and quantitative evaluation of preferences. To obtain criteria priorities, pairwise comparisons based on the fundamental verbal/numerical 1-9 scale is required (Table 11). The number of necessary comparisons for each comparison matrix is $n\left(n-1\right)/2$, where n is the number of criteria. Each criterion gains a geometric mean of its comparisons, which are then normalized.

An important requirement is to test consistency of our stated preferences, as human-made decisions can be mutually inconsistent because of the human nature. The most commonly used method for consistency check was developed by Saaty (1977) who proposed a consistency index (CI) related to eigenvalue method. CI is obtained as:

$$\mathrm{CI}=\frac{\lambda_{\max}-n}{n-1} \tag{1}$$

$\lambda_{\max}$ is the maximal eigenvalue of the pairwise comparison matrix. The consistency ratio (CR) is given by:

$$\mathrm{CR}=\frac{\mathrm{CI}}{\mathrm{RI}} \tag{2}$$

RI is the random index shown in Table 1.

n	3	4	5	6	7	8	9	10
RI	.58	.9	1.12	1.24	1.32	1.41	1.45	1.49

Table 1: AHP - Random indices (Saaty, 1977)

Data

This analysis considered a database donated by the University of Wisconsin to a public ML Repository (Mangasarian, Street & Wolberg, 1995). It gathers the information of 569 patients with a breast lump, with 30 features each. Features were computed from a digitalized image of a fine needle aspirate of the breast mass. They used a curve-fitting program to determine the exact boundaries of the nuclei. Figure 3 shows an example of the nuclei. Then, they computed 10 features for each nucleus: area, radius, perimeter, symmetry, number and size of concavities, fractal dimension (of the boundary), compactness, smoothness (local variation of radial segments), and texture (variance of grey levels inside the boundary). Thus, the mean value, extreme value (i.e., largest or worst value: biggest size, most irregular shape) and standard error of each of these cellular features were computed by them for each image, resulting in a total of 30 real-valued features.

This is the database with the 30 independent variables, and the response variable, i.e. Malignant or Benign Breast Lump, in which our ML models were trained and tested, leading to the results in Table 2.

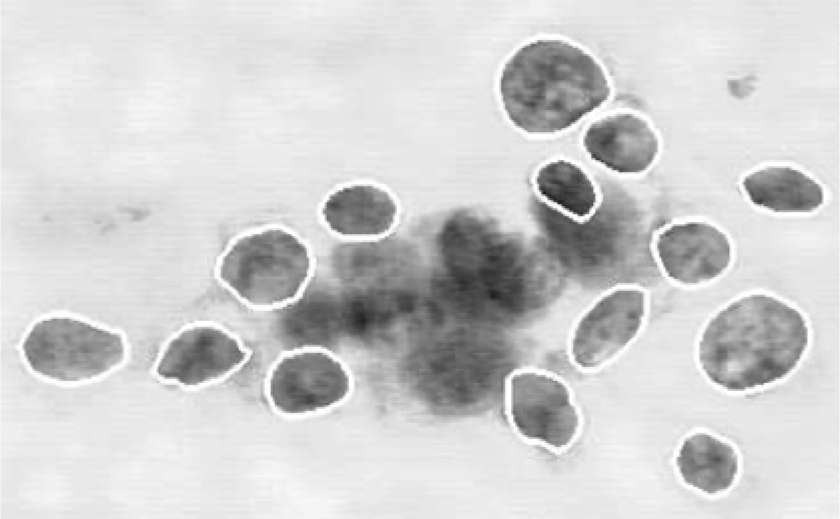

Figure 3: A magnified image of a malignant breast fine needle aspirate, outlined by a curve-fitting program. Model's features were extracted by using means, standard errors and extreme values (Mangasarian, Street & Wolberg, 1994).

For the purpose of the analysis, following ML models are alternatives of the AHP model: Decision Trees, K-Nearest Neighbours, Logistic Regression, Artificial Neural Network, Naïve-Bayes, Random Forest and Support Vector Machine. As we had a binary response variable, Malignant or Benign Breast Mass, we encoded the variable as 1 if the Breast Mass was Malignant, 0 if not. As a reminder, ML models for binary classification outputs a score between 0 and 1, then we may choose an adequate cut-off point for the score were values above the threshold are classified as 1, and the others as 0. Generally accepted, the cut-off point is 0.5, which was used in our analysis. Then, for each model, we measured two kinds of metrics: Metrics Dependent on the Score Cut-off and Metrics Independent on the Score Cut-off. The former metrics evaluate the results given in terms of the binary classification, i.e. benign and malign tumour prediction; the latter do so in terms of the estimated probability that a tumour is malign. These metrics were our main criteria of the AHP model. For the former criterion, following usual metrics are used as its sub-criteria:

- *Sensitivity*: measures how often the model correctly generates a positive result for people who have a malign tumour.
- *Specificity*: measures the model's ability to correctly generate a negative result for people who have a benign tumour.
- *Accuracy*: measures how often the model makes a correct diagnosis.
- For the latter, following usual metrics are used:
- *AUROC*: evaluates how well the model classifies positive and negative outcomes at all possible cut-offs.
- *Divergence*: measures the difference between the means of the malign and benign tumour standardized distributions using variances.
- *KS Statistic*: measures the maximum difference between the cumulative true positive and cumulative false positive rate.

Figure 4 is the schematic representation of the constructed AHP model. Finally, we used Python programming language to train every model on the patients' lumps data, and to get the predicted class, i.e. Malignant or Benign. Thus, we calculated each metric for each model (Table 2). To be specific, we trained the models on 70% of the data, and on the other 30% calculating the metrics. This is a generally accepted train/test split (Bronshtein, 2017).

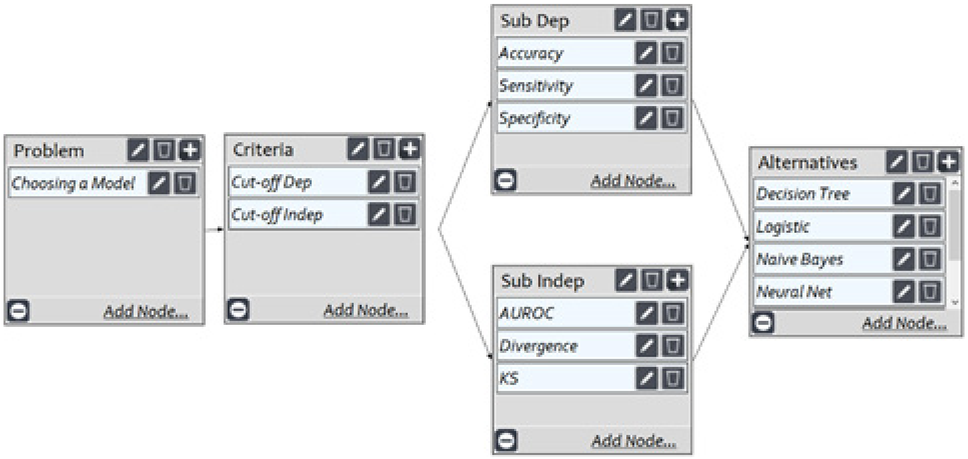

Figure 4: Schematic representation of the problem. Made using "Super Decisions" software. This representation is standard for Decision-Making Theory.

Model	Cut-off dependent			Cut-off independent		
	Specificity	Sensitivity	Accuracy	AUROC	KS	Divergence
Logistic	0.9683	0.9537	0.9591	0.9950	0.9312	29.6777
Decision Tree	0.9524	0.8981	0.9181	0.9253	0.8505	10.4526
Random Forest	0.9841	0.9722	0.9766	0.9971	0.9656	32.6920
SVM	0.9365	0.9815	0.9649	0.9949	0.9405	28.0637
KNN	0.9365	0.9815	0.9649	0.9751	0.9180	17.5027
Naïve Bayes	0.9048	0.9352	0.9240	0.9864	0.8611	9.4815
Artificial Neural Network	0.9048	0.9630	0.9415	0.9733	0.8862	13.3902

Table 2: Metrics by model (Own calculations).

Criteria and sub-criteria importance

To calculate the importance of the criteria and sub-criteria in the AHP model, we used experts' opinion. In total, we asked 10 experts to express how important is each criterion/sub-criterion for the cancer diagnosis. As the model includes two principal criteria, we divided the set of experts into the two groups regarding each criterion. In the first case, we used Data Scientists and Medics, whereas in the second case, only Data Scientists were used. The experts were asked to express the importance of each criterion on a scale from 1 to 10. 1 being "Zero importance" and 10 being "Very important". After, we used the AHP methodology to calculate the final importance of each criterion and sub-criterion. Table 3 presents the all the obtained importance.

Criteria	Sub-criteria			Inconsistency
Cut-off independent metrics	AUROC	Divergence	KS	0%
66.66%	60%	20%	20%	
Cut-off dependent metrics	Accuracy	Sensitivity	Specificity	0%
33.33%	42.85%	42.85%	14.28%	

Table 3: AHP Model Performance Metrics Importance

3 Results

In order to evaluate the models, we applied two different rating scales for the criteria. Since AUROC, KS, Accuracy, Sensitivity and Specificity require maximization and range from 0 to 1 (in this case from 0.85 to 1), we used the following rating scale:

Scale Item						Inconsistency
0.85-0.875	0.875-0.9	0.9-0.925	0.925-0.95	0-95-0.975	0.975-1	1.95%
4.28%	6.41%	10.1%	15.96%	25%	38.25%	

Table 4: AHP Rating Scale I

On the other hand, for Divergence, which also required maximization, but its range was different (from 9.48 to 32.69), we used the following rating scale:

Scale Item						Inconsistency
9.48-13.35	13.35-17.2	17.2-21	21-24.96	24.96-28.8	28.8-32.69	1.95%
4.28%	6.41%	10.1%	15.96%	25%	38.25%	

Table 5: AHP Rating Scale II

The objective of the study was to identify the most adequate model to undertake the task in hand: “Cancer Tumour Diagnosis” in order to choose a correct medical treatment to each clinic case. Given that the criterion used in the evaluation of these models fall into two categories: cut-off independent and cut-off dependent metrics, there was a perfect opportunity to make a more in-depth analysis into the models’ performance by splitting the inquiry into 3 different studies. The first one comparing the models considering only the cut-off independent metrics, a second analysis contemplating the cut-off independent metrics and an overall analysis.

The rationale behind this separation is the nature of the output: should the probability of having a malign tumour be required, the results of the first analysis are more accurate; whereas if the raw binary prediction is the predilected output, then the results of the second analysis are more precise. The overall analysis comprises both sets of metrics, whose schematic representation can be seen in Figure 4.

Cut-off independent metrics analysis

Having selected only this set of metrics, the “Super Decisions” software was run with the data given in Tables 2 and 3 using the Analytical Hierarchy Process (AHP) previously highlighted. This led to the following insightful results about the predictive power of each model in terms of the probability of a tumour of being malign.

Alternative	Ideal Results
Decision Tree	31.70%
KNN	75.76%
Logistic	94.90%
Naïve Bayes	69.26%
Neural Net	49.39%
Random Forest	100%
SVM	87.48%

Table 6: AHP Cut-off independent metrics results

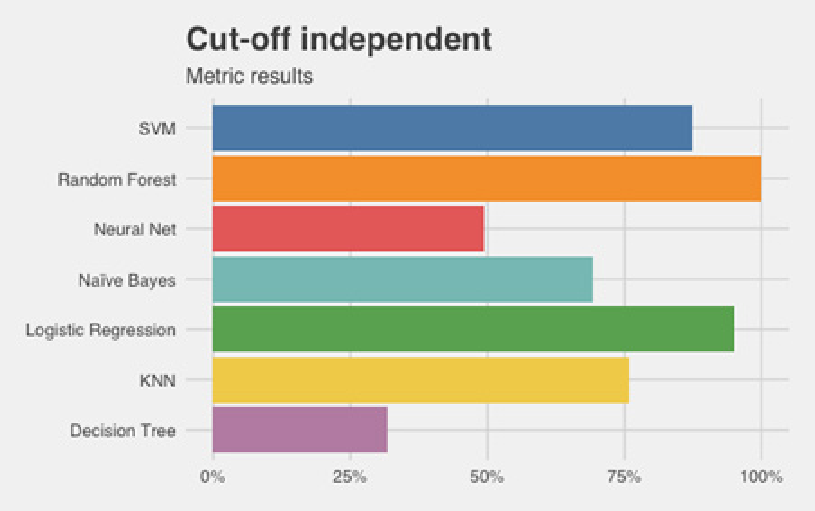

Figure 5: AHP Cut-off independent metrics results bar plot. Cut-off independent performance was spread between models.

As seen in Table 6 and Figure 5, the Random Forest resulted the best model with an ideal score of 100%, mainly because it was the highest evaluated in the three metrics: 0.9950, 0.9312, 29.6777 in AUROC, KS and Divergence respectively (Table 2). Nonetheless, there are two other models that also stand out, the Logistic Regression came a close second with a score of 94.90% followed by the Support Vector Machine (SVM) scoring 87.49% (Table 6). Therefore, these three models have the upper hand when it comes to probability as the output. On the other hand, the Decision Tree turned out to be the worst model (31.7%, Table 6) despite not being the worst evaluated in all three metrics: having a Divergence score of 10.4526, just above Naïve Bayes (Table 2). The Neural Network and Naïve Bayes also performed poorly in the test, with scores of 49.39% and 69.26% respectively (Table 6).

Cut-off dependent metrics analysis

Having selected the second set of metrics, the "Super Decisions" software was once again run with the data given in Tables 2 and 3 using the Analytical Hierarchy Process (AHP) previously highlighted. This led to the following interesting results about the predictive power of each model in terms of the binary classification:

Alternative	Ideal Results
Decision Tree	32.64%
KNN	90.23%
Logistic	76.84%
Naïve Bayes	38.63%
Neural Net	58.33%
Random Forest	100%
SVM	90.23%

Table 7: AHP Cut-off dependent metrics results

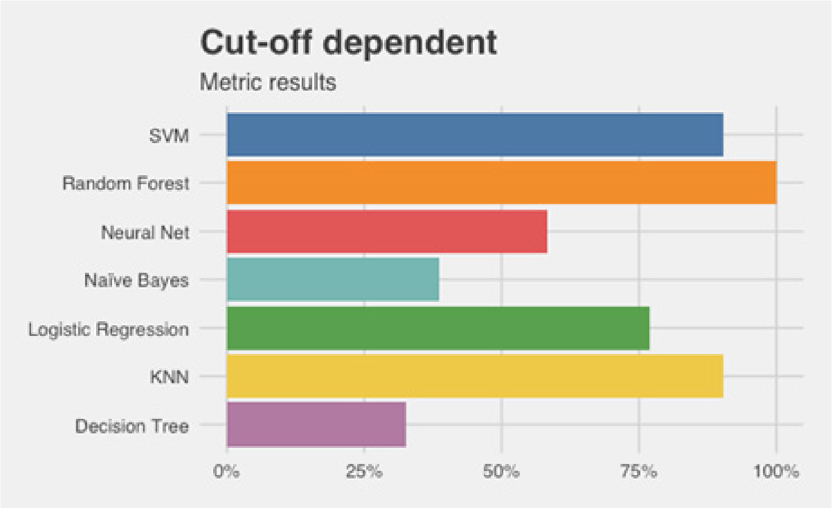

Figure 6: AHP Cut-off dependent metrics results Bar Plot. Colors are consistent with the last graph, for an easier comparison.

As seen in Table 7 and Figure 6, the Random Forest once again turned out to be the highest ranked model having and ideal score of 100% (Table 7), in this case not necessarily evaluated the best in all criterion: ranking third in the Sensitivity metric with a value of 0.9722 (Table 2). In contrast to the previous results, the Logistic Regression is no longer in the top 3, as the model decreased its ideal score from 94.90% to 76.84% (Tables 6 and 7), being outranked by the tie between K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) scoring 90.23%, almost 10% lower than de Random Forest (Table 7). This means, these three models are the most adequate when it comes to binary classification. Looking at the worst evaluated models, the Decision Tree was much penalized by its Sensitivity value (0.8981, Table 2) which makes it the worst evaluated model (32.64%, Table 7). Unlike the previous results, Naïve Bayes and Neural Net switched positions as the second and third worst models respectively (38.63% and 58.33%).

Alternative	Results Difference
Decision Tree	-0.94%
KNN	-14.47%
Logistic	18.06%
Naïve Bayes	30.63%
Neural Net	-8.94%
Random Forest	0%
SVM	-2.75%

Table 8: Results Comparison

Table 8 show the difference in ideal scores between the two previous analyses. The models with negative values (Decision Tree, KNN, Neural Net and SVM) are the ones which scored higher for the cut-off dependent metrics, being the KNN the one with the widest difference (-14.47%, Table 8). As for the models with positive differences (Logistic and Naïve Bayes), these models had a significantly higher evaluation in terms of cut-off independent metrics, 18.06% and 30.63% respectively (Table 8). It is worth noting that 30.63% means that the model is quite unstable in terms of its performance, as its results can vary considerably from one analysis to another.

Overall analysis

Since the experts assigned different weights to the two set of criteria, the overall analysis has the capability of displaying more robust results owing to its consideration of the importance shown in Table 3. Table 8 and Figure 7 show the results displayed by the "Super Decisions" software.

Alternative	Ideal Results
Decision Tree	31.99%
KNN	80.30%
Logistic	89.23%
Naïve Bayes	59.65%
Neural Net	52.20%
Random Forest	100%
SVM	88.34%

Table 9: Overall analysis results

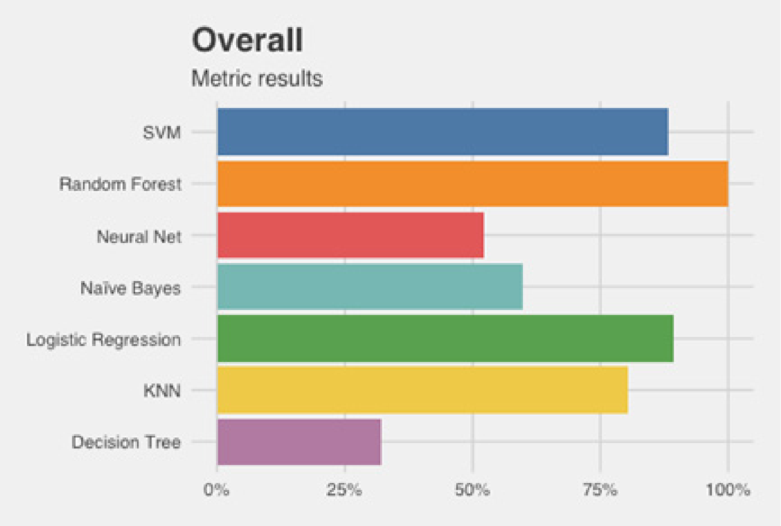

Figure 7: AHP Overall analysis results Bar Plot. The best model was assigned 100%, while the other's performance is rescaled by a factor of the Random Forest performance.

As seen in Table 8 and Figure 7, it comes to no surprise that the Random Forest came out on top in the overall analysis, since it got an ideal score of 100% in the both previous analyses (Tables 6 and 7). However, the most insightful results can be observed in the other positions. For instance, Logistic and SVM obtained similar results approximately 11% behind the Random Forest, which means that the first place is significantly better than the rest of the models (Table 9). As for the worst evaluated models overall, the Decision Tree is not only the worst, but also significantly low, approximately 20% worse than the Neural Network, the second lowest evaluated model, which scored 52.20% (Table 9). It is also worth noting that Naïve Bayes performed poorly as well with a score of just below 60% (Table 9). Finally, Figure 8 shows the comparison between the 3 analysis and how they vary from each model.

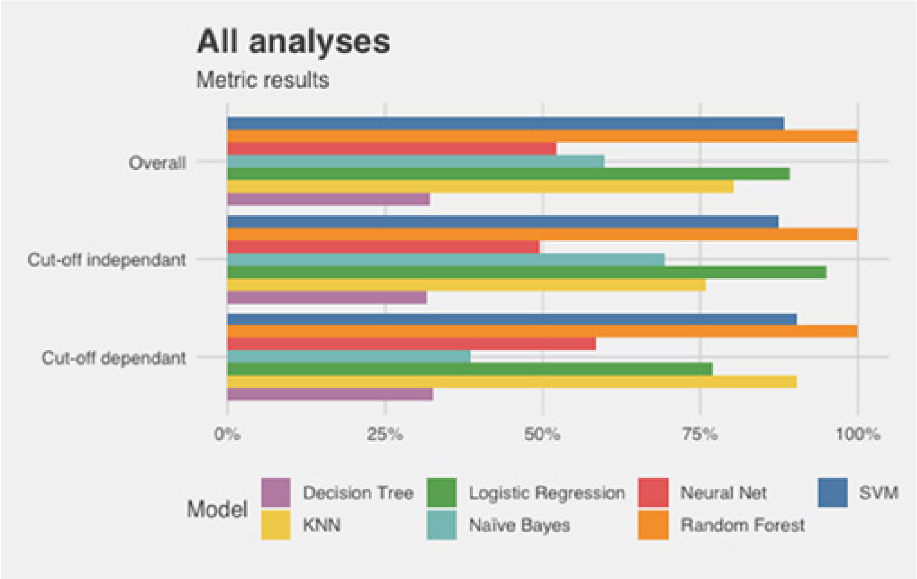

Figure 8: AHP Results Summary Bar Plot. All past graphs regarding performance are summarized here.

Figure 8 exhibits the Logistic Regression and Naïve Bayes as the most unstable models given that their scores vary significantly form one analysis to another (difference of 18.06% and 30.63% respectively, Table 8), whereas the rest of the models tend to have similar scores in every analysis, with difference in their results of below 15% (Table 8). To sum up, the Random Forest is clearly de best model for Breast Cancer Diagnosis by a long margin, unlike the Decision Tree, whose scores are significantly lower than the rest of the models (overall score of 59.65%, Table 9), which is clearly portrayed in figure 8. So, if a model were to be discarded, the Decision Tree would be removed.

4 Discussion

In all three analyses, out of the 7 models reviewed, the Random Forest turned out to be the best alternative for Malignant Breast Cancer diagnosis. It is fair to say that there is a wide variety of more complex ML models worth analysing. For instance, Reddy Vaka, Soni and Reddy (2020) introduced a Deep Neural Network with Support Value (DNNS) which outperformed models such as Naïve Bayes and SVM classifiers. Likewise, Chaurasia and Pal (2004) compared the performance of different supervised learning models, such as SVM-RBF kernel, RBF neural networks, Naïve Bayes and Decision Trees, showing that SVM-RBF kernel was the most accurate model. Finally, Asri et. al. (2016) in their analysis comparing SVM, NB and KNN empirical results demonstrate that SVM achieves the highest accuracy (97.13%) having the lowest error rate.

Delving more deeply into the winning model and comparing it to the rest of the models, there are a few prons and cons worth considering when it comes to their attributes:

Model	Pros	Cons
Logistic Regression	Simple and linear, reliable and with no parameters to tune	Cannot handle non-linearities in the data
Decision Tree	Decision rules are easily visualized and interpreted by people including users without machine learning expertise. Tend to work well with data sets that have a mixture of feature types	Despite the use of pruning they can still overfit all or parts of the data and may not achieve the best generalization performance compared to other methods.
KNN	Simple and easy to understand why a particular prediction is made. It can be a reasonable baseline against what you can compare more sophisticated methods.	When the training data has many instances, or each instance has lots of features, this can really slow down the performance of a KNN model.
Naïve Bayes	Fast to train and use for prediction and thus are well suitable to high dimensional data including text	Conditional independence assumption. When getting confidence or probability estimates associated with predictions, Naive Bayes classifiers produce unreliable estimates, typically.
SVM	Perform well on a range of datasets and have been successfully applied on data that ranges from text to images and many more types.	As the training set size increases, the run time, speed, and memory usage in the SVM training phase also increases. So, for a large dataset with hundreds of thousands, or millions of instances, an SVM may become less practical.
Neural Network	Neural networks form the basis of advanced learning architectures that capture complex features and give state-of-the-art performance on an increasingly wide variety of difficult learning tasks.	These larger and more complex models typically require significant volumes of data, computation, and training time to learn. Careful pre-processing of the input data is needed, to help ensure fast, stable, meaningful solutions to finding the optimal set of weights.
Random Forest	Reliable, powerful and ability to handle non-linearities in the data	It can be very difficult for people to interpret, making it difficult to see the predictive structure of the features or to know why a particular prediction was made.

Table 11: Machine Learning Models Comparison (Collins-Thompson, 2013; Ravindran, 2018)

When creating a predictive model, all the techniques should be tried, and the best performing method should be taken (Ravindran 2018).

For ML models, as well as other models, you are ought to evaluate the performance of your model, owing to the fact that there is no “Jack of all trades” model that can be applied for all range of problems. As implied by the works of Wolpert and Macready (1997), there is no algorithm that outperforms every other algorithm in every situation. Thus, given a set of ML trained models, one problem is choosing the best one. Nevertheless, this is no trivial problem, for there are many metrics from which we can assess the model performance. In fact, a ML model can outperform another ML model measured by one metric but underperform measured by other metric. E.g. in our analysis, Logistic Regression performed better than KNN in Specificity, but KNN performed better than Logistic Regression in Accuracy (Table 2). With this Decision-Making Theory approach, rather than selecting one metric for comparing model performance, we gathered the expertise of Data Scientist and Oncologists, so the selection of a model is made based upon more data, combining different information available that enables us for a more complete inference of which of the models will generalize to patients the models were not trained on.

An important suggestion to point out is the implementation and evaluation of these models on other variables that involve potential cancerous tumours, namely variables related to symptoms, psychological, physiological or another anatomic factors besides the ones on which the models analysed in this paper where trained and tested. For example, Gupta and Chawla (2020) analysed histopathological images concluding that a neural network called "ResNet50" outperformed the SVM in terms of accuracy. Consequently, for different types of analyses, scores and ranking may differ substantially after running an AHP method on their metrics: while for the breast mass analysis the Random Forest is the optimum, for a symptomatic analysis, for instance, the best model might change. The challenge would be to collect and compile data of such nature. Regarding psychological factors, Lötsch et. al. (2018) explain that prevention of persistent pain after breast cancer surgery, via early identification of patients at high risk, is a clinical need and psychological factors are among the most consistently proposed predictive parameters for the development of persistent pain.

Finally, the analysis is not just over once a malign tumour has been correctly diagnosed- further tests should be made in order to carry out the most suitable treatment based on factors such as its stage and grade, size, and whether the cancer cells are sensitive to hormones (Mayo Clinic, 2011). Djebbari et al. (2008) considered the effect of ensemble of machine learning techniques, like the Randoms Forests, to predict the survival time in breast cancer. Therefore, an inquiry into the process of following up the clinic cases would perfectly complement our analysis, the more resources and tools we get to tackle this appalling disease, the better. Additionally, Lopez Guerra et al. (2013) designed and created an advanced clinical decision support system (CDSS) based upon different artificial intelligence techniques such as data mining and the use of Bayesian Classification Trees to assist in the therapeutic process of breast cancer. In like manner, Al-Allak, Bertelli and Lewis (2016) showed that new ML algorithms can outperform the traditional statistical methods that have in the past been used to generate tools that predict survival.

5 Conclusions

Although we have demonstrated that the Random Forest is by far the most optimum model (Figure 8), the scientific knowledge of a medical specialist cannot be replaced by it. On the contrary, Doctors can make use of the model as a complement, rather than their substitution, as well as a welcome advancement in this medical field. The process of Breast Cancer Detection is a substantially delicate problem due to the health implications of the correct or incorrect diagnosis, being human lives at stake. For that reason, relying solely on Machine Learning algorithms to decide the proper treatment has many risks. Fortunately, however, using these models as a tool to aid the Oncologist in each clinical case is a much better option. Furthermore, the development of technologies for medical diagnosis, do not imply the need for a layoff of Oncologists whose part of their job is to read mammographies. But imply the opportunity of a symbiotic relationship, where the medic benefits from reducing his workload as shown by the work of McKinney et al. (2020).

Health and overall well-being are valuable assets so much more important than many people are aware, our resolution is that availing ourselves with the implementation of modern Machine Learning and

Decision-Making algorithms into the medical field can transform the lives of many people in the medium and short term. Not only due to the software capabilities of today, but also due to their high performance. The main advantage of this kind of models is that they have no limitations when it comes to their implementation and Oncology is no exception. As Al-Allak, Bertelli and Lewis (2016) stated, these methods can offer a viable alternative to generate more accurate predictive models with the potential of improving patient outcomes.

For the near future, we envisage that the integration of such algorithms into this field will bring about sweeping changes into the way many diseases are diagnosed and treated. As many more complex models are likely to be developed, the way experts study medicine will change as we know it: the fusion of modern computer science with the medical field will require a new generation of doctors who will not only have need of expertise in their specialities, but will also be obliged to acquire skills in reading and interpreting advanced models. All in favour of millions of patients eager to hear of new alternatives to undergo their treatments in a swifter and smoother manner.

6 References

Accenture (2018). *Consumer Survey on Digital Health*. [Online], Available: https://www.accenture.com/acnmedia/PDF-71/Accenture-Health-Meet-Todays-Healthcare-Team-Patients-Doctors-Machines.pdf#zoom=50 [10 Mar 2020].

Al-Allak, A., Bertelli, G. & Lewis, P.D. (2013). Random forests: The new generation of machine learning algorithms to predict survival in breast cancer. *International Journal of Surgery*, 11(8), 607. https://dx.doi.org/10.1016/j.ijsu.2013.06.112

American Cancer Society (2012). *Cancer Facts & Figures*. American Cancer Society (ACS), Atlanta.

American Cancer Society (2016). *What it is cancer?* ACS. Retrieved from: https://www.cancer.org/es/cancer/aspectos-basicos-sobre-el-cancer/que-es-el-cancer.html [10 Mar 2020].

Anderson, B.O. (2014). *UICC World Cancer Congress 2014: Global Breast Cancer Trends*. Washington. [Online], Available: www.worldcancercongress.org/sites/congress/files/atoms/files/UICC41_Anderson-Benjamin-O.pdf [12 Mar 2020].

Bronshtein, A. (2017). *Train/Test Split and Cross Validation in Python*. Towards Data Science. [Online], Available: https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6 [27 Abr 2020].

Chaurasia, V. & Pal, S. (2004). Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability. *International Journal of Computer Science and Mobile Computing IJCSMC*, 3(1), 10–22.

Collins-Thompson, K. (2013) *Applied Machine Learning in Python*. University of Michigan. Coursera, Available: https://www.coursera.org/learn/python-machine-learning [29 Aug 2020].

Djebbari, A., Liu, Z., Phan, S. & Famili, F. (2008). An ensemble machine learning approach to predict survival in breast cancer. *International Journal of Computational Biology and Drug Design*, 1(3), 275-294. https://dx.doi.org/10.1504/ijcbdd.2008.021422

Hastie, T., Tibshirani, R. & Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd Edition. Springer.

Asri, H., Mousannif, H., Al Moatassime, H. & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, 83, 1064-1069. https://doi.org/10.1016/j.procs.2016.04.224

Houssami, N. & Hunter, K. (2017). The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *NPJ Breast Cancer*, 3(12), 1-13. https://dx.doi.org/10.1038/s41523-017-0014-x

Elmore, J.G., Jackson, S.L., Abraham, L., Miglioretti, D.L., Carney, P.A., Geller, B.M., Yankaskas, B.C., Kerlikowske, K., Onega, T., Rosenberg, R.D., Sickles, E.A. & Buist, D.S.M. (2009). Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology*, 253(3), 641–651. https://dx.doi.org/10.1148/radiol.2533082308

Fenton, J.J., Taplin, S.H., Carney, P.A., Abraham, L., Sickles, E.A., Berns, E.A., Cutter, G., Hendrick, R.E., Barlow, W.E. & Elmore, J.G. (2007). Influence of computer-aided detection on performance of screening mammography. *The New England Journal of Medicine*, 356(14), 1399–1409. https://doi.org/10.1056/NEJMoa066099

Gupta, K. & Chawla, N. (2020). Analysis of Histopathological Images for Prediction of Breast Cancer Using Traditional Classifiers with Pre-Trained CNN. *Procedia Computer Science*, 167, 878-889. https://doi.org/10.1016/j.procs.2020.03.427

Kahn, S. (2016). *Scientist, Toronto Rehabilitation Institute, Canada. In response to the question "What are the most important machine learning algorithms?".* [Online], Available: https://www.quora.com/What-are-the-most-important-machine-learning-algorithms-What-are-the-most-commonly-applied-algorithms-when-attacking-a-problem [29 Aug 2020].

Kohli, A. & Jha, S. (2018). Why CAD failed in mammography. *Journal of the American College of Radiology*, 15(3), 535–537. https://doi.org/10.1016/j.jacr.2017.12.029

Kourou, K., Exarchos, T.P., Exarchos, K., Karanouzis, M.V. & Fotiadis, D. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8-17. https://doi.org/10.1016/j.csbj.2014.11.005

Lantz, B. (2015). *Machine Learning with R*. 2nd Edition. Packt publishing.

Lehman, C.D., Wellman, R.D., Buist, D.S.M., Kerlikowske, K., Tosteson, A.N.A., Miglioretti, D.L. & Breast Cancer Surveillance Consortium (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Internal Medicine, 175(11), 1828–1837. https://dx.doi.org/10.1001/jamainternmed.2015.5231

Lopez Guerra, J., Moreno, A., Parra, C., Gonzalez, R., Martinez, A., de Leon, J., Vieites, R., Ruiz, M., Lopez, M., Nieto, J., Fernandez, M., Rodriguez, E., Quintana, B. & Ortiz, M. (2013). Machine learning techniques to improve therapeutic decision-making in breast cancer. *Reports of Practical Oncology and Radiotherapy*, 18, Supplement 1. http://dx.doi.org/10.1016/j.rpor.2013.03.668

Lötsch, J., Sipilä, R., Dimova, V. & Kalso, E. (2018). Machine-learned selection of psychological questionnaire items relevant to the development of persistent pain after breast cancer surgery. *British Journal of Anaesthesia*, 121(5), 1123-1132. https://doi.org/10.1016/j.bja.2018.06.007

Mayo Clinic (2011). *Breast cancer. Patient Care & Health Information: Diseases & Conditions*. [Online], Available: https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475 [24 May 2020].

Mangasarian, O., Street, W. & Wolberg, W. (1994). Breast Cancer Diagnosis and Prognosis via Linear Programming. *Operations Research*, 43(4), 1-9. https://doi.org/10.1287/opre.43.4.570

Mangasarian, O., Street, W. & Wolberg, W. (1995). B*reast Cancer Wisconsin (Diagnostic) Data Set. Machine Learning Repository.* UCI Center for Machine Learning and Intelligent Systems. [Online], Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 [30 Mar 2020].

McKinney, S.M., Sieniek, M., Godbole, V. & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577, 89-94. https://doi.org/10.1038/s41586-019-1799-6

INEGI (2018). *STATISTICS ON WORLD CANCER DAY (4 FEBRUARY).* Instituto Nacional de Estadística y Geografía, [Online], Available: https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2018/cancer2018_nal.pdf [11 Mar 2020].

Reddy Vaka, A., Soni, B. & Reddy, K.S. (2020). Breast cancer detection by leveraging Machine Learning. *ICT Express*, in press. https://doi.org/10.1016/j.icte.2020.04.009

Ravindran, P. (2018). *Data Science and ML enthusiast, PhD, IIT Delhi. In response to the question "What are pros and cos of logistic regression and random forest?".* [Online], Available: https://www.quora.com/What-are-pros-and-cos-of-logistic-regression-and-random-forest [29 Aug 2020].

R Project (2016). *R Fortunes: Collected Wisdom*. [Online], Available: https://cran.r-project.org/web/packages/fortunes/vignettes/fortunes.pdf [10 Mar 2020].

Saaty, T.L. (1977). A Scaling Method for Priorities in Hierarchical Structures. *Journal of Mathematical Psychology*, 15(3), 234-281. http://dx.doi.org/10.1016/0022-2496(77)90033-5

Saaty, T.L. (1980). *The Analytic Hierarchy Process*. McGraw-Hill, New York.

Saaty, R.W. (1987). The Analytic Hierarchy Process - What it is and how it is used. *Mathematical Modelling*, 9(3-5), 161-176. http://dx.doi.org/10.1016/0270-0255(87)90473-8

Silverio, M. (2020). *Google AI for breast cancer detection beats doctors*. Towards Data Science. Retrieved from: https://towardsdatascience.com/

google-ai-for-breast-cancer-detection-beats-doctors-65b8983352e0 [12 Mar 2020].

Szeliski, R. (2010). *Computer Vision: Algorithms and Applications*. 1st Edition. Springer Science & Business Media.

Taulli, T. (2019) *Artificial intelligence basics: a non-technical introduction*. 1st Edition. Apress. https://doi.org/10.1007/978-1-4842-5028-0

TAC (2017). *Cancer panorama in Mexico*. Together Against Cancer, [Online], Available: https://juntoscontraelcancer.mx/panorama-del-cancer-en-mexico/ [11 Mar 2020].

Tosteson, A.N.A., Fryback, D.G., Hammond, C.S., Hanna, L.G, Grove, M.R., Brown, M., Wang, Q., Lindfors, K. & Pisano, E.D. (2014). Consequences of false-positive screening mammograms. *JAMA Internal Medicine*, 174(6), 954–961. https://dx.doi.org/10.1001/jamainternmed.2014.981

Wolpert, D. & Macready, W. (1997). No Free Lunch Theorems for Optimization. *IEEE Transactions on evolutionary computation*, 1(1), 67-82. https://dx.doi.org/10.1109/4235.585893

WHO (2004). *National cancer control programs*. World Health Organization, Washington DC. [Online], Available: https://www.paho.org/hq/dmdocuments/2012/OPS-Programas-Nacionales-Cancer-2004-Esp.pdf [12 Mar 2020].

WHO (2018a). *Fact sheets "Cancer"*. World Health Organization, [Online], Available: http://www.who.int/en/news-room/fact-sheets/detail/cancer [11 Mar 2020].

WHO (2018b). *International Agency for Research on Cancer: Mexico*. World Health Organization, [Online], Available: https://gco.iarc.fr/today/data/factsheets/populations/484-mexico-fact-sheets.pdf [11 Mar 2020].

7 Appendix

Intensity of importance on an absolute scale	Definition	Explanation
1	Equal importance	Two activities contribute equally to the objective
3	Moderate importance of one over another	Experience and judgement strongly favor one activity over another
5	Essential or strong importance	Experience and judgement strongly favor one activity over another
7	Very strong importance	An activity is strongly favored and its dominance demonstrated in practice
9	Extreme importance	The evidence favoring one activity over another is of the highest possible order of affirmation
2, 4, 6, 8	Intermediate values between the two adjacent judgements	When compromise is needed
Reciprocals	If activity *i* has one of the above numbers assigned to it when compared with activity *j*, then *j* has the reciprocal value when compared with *i*	
Rationales	Ratios arising from the scale	If consistency were to be forced by obtaining *n* numerical values to span the matrix

Table 11: AHP – fundamental scale (Saaty, 1987: 165)