Artificial Intelligence for Breast Cancer Diagnosis: Future Trends in Histopathological Analysis

María Cancho Paniagua[1,2], María Durán Bonilla[1,2]
Paula Montes Jiménez[1,2], Helena Sofía Yaben Lopezosa [1,2]
Sergio Muñóz Romero[1,2,3], José Luis Rojo [1,2,3]

[1] Biomedical Engineering Degree, Rey Juan Carlos University, Alcorcón, Spain.
[2] Telecommunications Engineering Superior Technical School, Rey Juan Carlos University, Madrid, Spain. [3] Department of Signal Theory and Communications, Madrid, Spain.

Abstract

Breast cancer is the most common cancer diagnosed in women across the world. It accounts for 1 in 4 cancer cases for female, being the leading cause of death from cancer in women. Currently, biopsy of breast constitutes the most effective technique to distinguish between malignant and benignant breast masses.

Hystopathological analysis of the nuclei characteristics can determine the nature of the tissue since malignant nuclei typically present distinctive variations in size, shape, or texture when compared to non-malignant nuclei. Accurate diagnostic detection of the cancerous cells in a patient is critical and may alter the subsequent treatment and increase the chances of survival rate. Interpretation of histology slides is usually a time-consuming and complex task, where observer variability can affect classification outcome.

Through a comparative analysis, the aim of this article is to find the machine learning algorithm that provides the best performance metrics for the problem exposed, as well as to identify the nuclei characteristics that contribute the most to the classification of a sample. Results show that Neural Network classifier provides the best performance metrics (validation accuracy $\approx$ 98%), and that worst and mean values for concave points, concavity, radius, area, and perimeter are considered among the most relevant features for class prediction.

1 Introduction

Breast cancer is the most frequently diagnosed cancer in the female population. According to the Globocan 2020 data, this disease accounts for 1 in 4 cancer cases and 1 in 6 cancer deaths among women[1]. Last year, breast cancer surpassed lung cancer as the leading cause of global cancer incidence, with a total amount of 2.3 million diagnosis and 685 000 deaths reported by the World Health Organization.

Breast cancer originates from an out-of-control cell growth in breast tissue, most commonly in the inner lining of milk ducts or the lobules that supply the ducts with milk. This cancer occurs almost entirely in women but it can be given in men as well, and its prevalence becomes stronger with age. [2] There is much evidence showing that factors such as age, genetics, ethnicity, obesity and a sedentary lifestyle influence the development of this disease. The detection of this disease in the early stages can help to avoid the rising number of deaths.

Nowadays, there is no current treatment to prevent mammary cancer, for this reason, its early detection plays a key role in the effective management of the disease. The prominent gold standard for breast cancer diagnosis includes a triple assessment consisting of (i) clinical breast examination, carried out by a trained health professional, and breast awareness; (ii) screening mammography and/ or ultrasonography; and (iii) fine needle aspirate cytology. As the mammography is an expensive and complicated technique that requires the intervention of specialist clinicians, the best approach for the diagnosis of breast cancer is the analysis of histopathological samples obtained through cytologies or biopsies.

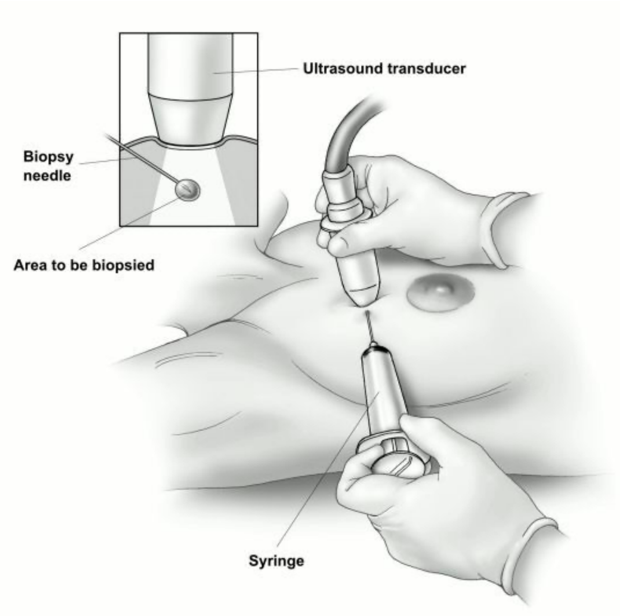


Figure 1. *Overview of fine needle aspiration procedure using ultrasound for the detection of breast cancer.*

Biopsy of breast tissue constitutes the most effective way to distinguish between malignant and benignant tissue. In concrete, the hystopathological analysis of the cell nuclei can be deterministic in the diagnosis since, typically, the nucleus of a cancer cell shows characteristic features related to size, shape and texture. Usually, nuclei become enlarged and darker and presents an irregular outline.

The complex nature of breast cancer requires careful stratification of patients in order to provide a tailored and efficient therapy, a process where histopathologic analysis is crutial. The interpretation of pathology slides often requires sophisticated, bulky and expensive technologies, as well as highly trained professionals. The digitalization of pathology slides and the increasing use of AI-based algorithms on hystopathological data have allowed to improve accuracy of current diagnosis techniques, as well as to reduce costs and interpretation time.

2 AI & Breast Cancer Diagnosis

The healthcare environment is one of the most appropriate fields for data science applications to break through, as it deals with the management of large volumes of data and complex information flows. Computer science-based systems are increasingly being recognised by many healthcare organisations and research professionals as it serves as a support strategy in medical decision-making.

Over the last years, artificial intelligence (AI) techniques have emerged as an important venue for health care since they have proven to be a powerful tool for detecting or predicting different medical conditions. AI-based algorithms offer a number of advantages over conventional predictive analytics in the medical field. They have resulted to be more precise and accurate as they interact with training data, allowing healthcare specialists to gain unprecedented insights into diagnostics, care processes, treatment variability, and patient outcomes.

In the context of breast cancer, a considerable amount of literature has been published on BC prediction using machine learning (ML) models. Some examples of AI methods that have been already integrated into clinical guidelines are the Breast Cancer Risk Assessment Tool (BCRAT) and the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA). These two models have been designed to identify women with high risk of suffering from breast cancer according to different risk factors. However, these models showed considerable limitations, which means that better ML algorithms should be incorporated in clinical practice [3].

A field that is currently undergoing a transformation towards a fully digital workflow is histopathology. This modality is distinctly analog[4], which supposes a higher barrier for the integration of AI methods. However, digitalization of histopathological practice would make easier and less time-consuming the analysis process of samples. In addition, feature extraction would be improved and the final analysis outcome would not be influenced by the level of experience of the pathologists involved. All in all, an automated diagnosis of breast cancer by analyzing histopathological data is essential for patients and their prognosis.

In the last decade, a significant effort has been put forth for breast cancer recognition from histological samples. During this period, the main focus has been placed on the classification of the two fundamental types of breast cancer, benign and malignant, using Computer Aided Diagnosis (CAD) in combination with AI. Machine learning approaches including the Support Vector Machine (SVM), Principle Component Analysis (PCA), and Random Forest (RF) have been used to study data whose features were extracted with Scale Invariant Feature Extraction (SIFT), Local Binary Patterning (LBP) and the Gray-Level Cooccurrence Matrix (GLCM) among others [5].

The major contribution that this paper offers is to inquire an adequate model basing on different predictive factors of early-stage breast cancer patients and evaluate its robustness through accuracy measures. Data visualization and machine learning applications will be used to provide a comprehensive comparison of the models from a broader perspective. Finally, the most effective features for predicting breast cancer will be established to understand general trends.

3 Materials and Methods

A series of 569 patients (212 cancer, 357 benign) provided the data used in this work, publically available at the UCI Machine Learning Respository [6]. This dataset was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. It was donated by Olvi Mangasarian on July 15th, 1992 [7].

To create the dataset, Dr. Wolberg used fluid samples aspirated from breast masses by fine-needle aspiration (FNA). The aspirates were placed on slides and later digitalised. Then, a program called Xcyt was used to locate a representative sample of the nuclei present in the aspirate. Xcyt uses a curve-fitting algorithm to determine the exact boundaries of the nuclei as showed in the figure 4. These final boundaries are known in the literature as a ”snakes” [8]. Finally, 10 features were calculated based on the boundaries obtained.

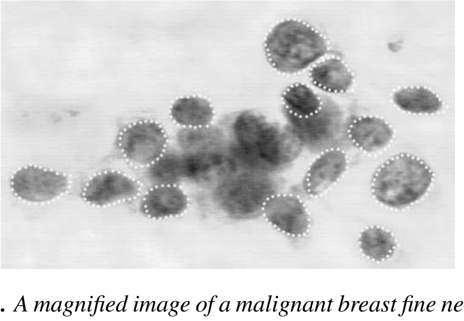


Figure 2. *A magnified image of a malignant breast fine needle aspirate. Visible cell nuclei are outlined by a curve-fitting program. The points are the snake points that Xcyt calculates.*

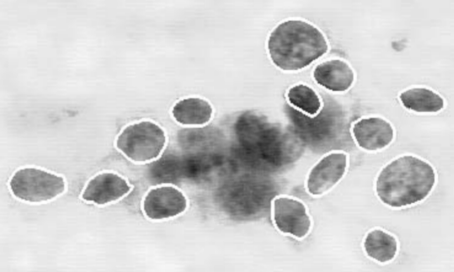


Figure 3. *Snakes are further converged to finally represent contours of the nuclei.*

The features were modeled such that higher values are typically associated with malignancy and that characterize nuclear size, shape, and texture:

1. Radius: the radius of an individual nucleus is measured by averaging the length of the radial line segments defined by the centroid of the snake and the individual snake points.
2. Perimeter: total distance between the snake points constitutes the nuclear perimeter.
3. Area: nuclear area is measured simply by counting the number of pixels on the interior of the snake and adding one-half of pixels in the perimeter.
4. Compactness: perimeter and area are combined to give a measure of the compactness of the cell nuclei using the formula $perimeter^2/area$.
5. Smoothness: the smoothness of a nuclear contour is quantified by measuring the difference between the length of a radial line and the mean length of the surrounding it.
6. Concavity: number and severity of concavities or indentations in a cell nucleus. Chords are drawn between non-adjacent snake points and the extent to which the actual boundary of the nucleus lies on the inside of each cord is measured. This parameter is greatly affected by the length of these chords, as smaller chords better capture small concavities. For this dataset, smaller chords were chosen to emphasize small indentations, since larger shape irregularities were captured by other features.
7. Concave points: number, rather than the magnitude, of contour concavities.
8. Symmetry: the major axis, or longest chord through the center, is found. Then, the length difference between lines perpendicular to the major axis to the cell boundary is measured.
9. Fractal Dimension: calculated using the ”coastline approximation”.
10. Texture: nuclear texture is measured by finding the variance of the gray scale intensities in the component pixels.

The mean value, worst (mean of three largest values), and standard error of each feature were computed for each image, resulting in a total of 30 features [7].

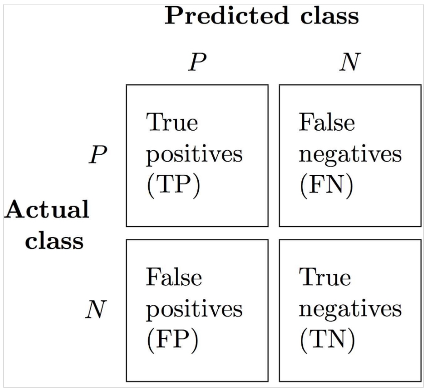
Different machine learning techniques are compared in this paper in order to build a classifier that correctly distinguishes between malignant and non-malignant samples of breast masses. All samples provided in the dataset are labelled, therefore constituting a supervised classification problem. In order to perform the analysis, *Python* was used as the programming language for coding.

In an effort to reduce computation times, especially during hyperparameter optimisation, cloud computing rather than local computing was preferred. In this sense, *Google Colab* was the platform selected on the basis of providing a free Jupyter notebook environment that runs on Google's cloud servers, enabling collaborators to eliminate the need to store data and install the environment locally, as well as to reduce processing times. Moreover, this way of computing allows synchronous contributions to the code, which makes it easier for the collaborators to maintain the project workflow.

Since this is a binary supervised classification problem, confusion matrix metrics were used for model performance comparison. Confusion matrix provides four numbers:

1. True Positives (TP): malignant samples classified as malignant.
2. True Negatives (TN): non-malignant samples classified as non-malignant.
3. False Positives (FP): non-malignant samples classified as malignant.
4. False Negatives (FN): malignant samples classified as non-malignant.

Scikit-learn calculates the confusion matrix by default on the basis of considering the value ”0” equivalent to negative, and the value ”1” equivalent to positive.



		Predicted class	
		P	*N*
Actual class	*P*	True positives (TP)	False negatives (FN)
	N	False positives (FP)	True negatives (TN)

Figure 4. *Confusion matrix where ”negative” represents non-malignancy, whereas ”positive” represents malignancy.*

In this manner, the confusion matrix provides a source for different performance measure calculation:

1. Accuracy: ratio of correctly classified samples to the total number of samples.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

2. Precision: also called positive predictive value, it is defined as the ratio of correctly positive classified samples to the total number of positives. It attempts

to determine the proportion of positive identifications that was actually correct.

$$Precision = \frac{TP}{TP+FP}$$

3. Recall: also called negative sensitivity, it is defined as the ratio of correctly positive classified samples to the total number of samples that should have been classified as positives. It attempts to determine the proportion of positives that was correctly identified.

$$Recall = \frac{TP}{TP+FN}$$

4. F1-Score: combination of precision and recall that can be interpreted as the weighted average of precision and recall.

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall}$$

4 Experiments and Results

4.1 Exploratory Data Analysis (EDA)

Once data is acquired, after ensuring its quality is optimal for data analysis, a deep exploration before pre-processing must be performed in order to better understand the data and to be able to perform a proper pre-processing and modelling. This is what is called *Exploratory Data Analysis (EDA)*. Understanding the dataset implies several steps:

1. Get information about data types, shape of the dataset and descriptive metrics.
2. Extract information about the relevancy of some features over others.
3. Identify outliers, missing values or human error.
4. Understand the relationship, or lack of, between variables.
5. Maximize the insight into the dataset and minimize the potential error that may occur during the next steps in the analysis porocess.

Hence, the main purpose of EDA is to explore the structure of the data and find patterns of behavior and distribution of the data.

4.1.1 Identification of Variables and Data Types

The dataset counts on 569 instances with 30 features each, these being the mean, standard error, and worst values of the nuclei characteristics of each sample. Moreover, the dataset also contains a column where diagnosis of each sample, namely malignant or non-malignant, is stored. As a result, each record has 31 features where the first thirty correspond to the predictors of the classifier and the last one corresponds to the target variable. Finally, every instance is identified by an ID. Including ID column in the analysis would have introduced bias and worsen the model performance. Therefore, this column was discarded in later steps.

Data types of each feature were also obtained. Artificial intelligence algorithms, as mathematical expressions, do only accept mathematical inputs. Therefore, if a feature is nominal, it should be transformed into numeric in order to be used in the analysis. In this dataset, labels were stored as letters, thus making diagnosis nominal. It was transformed into numeric in a way that class *B* was substituted by a 0, and class *M* was substituted by a 1. As a result, 0 indicates non-malignancy, whereas 1 indicates malignancy.

4.1.2 Descriptive statistics of the dataset

Once all the relevant variables were transformed into a suitable numerical form, descriptive statistical parameters count, mean, standard deviation, maximum, minimum, and quartiles were obtained for each of the features. Once statistical characteristics were displayed, we could deduce several conclusions such as if the minimum and maximum values were rational with respect to the variables themselves. This, is, we can see if there could be evident outliers. Initially, no abnormal values were identified by taking into account the nature of each feature.

Descriptive statistics allowed us to study null values as well. Two types of null or missing values can be understood: explicit and implicit missing values. Explicit missing values refer to those cells that were actually empty. This is, they make reference to the presence of an absence. On the other hand, implicit missing values refer to the absence of a presence. As an example, implicit missing values would be found in the case in which a record had all-zero features, where 0 does not fit the domain of the features. These values could mean that data was incorrectly obtained or stored. For this reason, mean maximum, and minimum of features must also be studied to identify non-explicit missing values.

No explicit missing values were found. Those records that had zero values for more than one feature were all associated to the non-malignant class, which led us to conclude that they did not represent implicit missing values.

Finally, class distribution was studied. The number of benignant samples is quite higher than the number of malignant samples, reaching almost 60/40 percentage difference. This is, our dataset is imbalanced, being *malignant* the majority class.

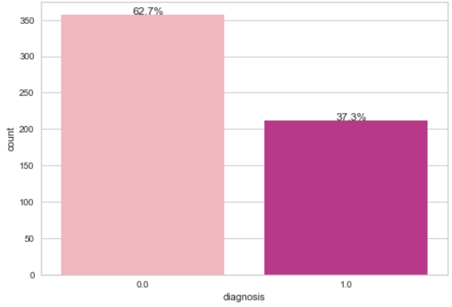


Figure 5. *Class distribution among non-malignant and malignant samples.*

Data imbalance may cause problems when predicting classes. If there is a high proportion of samples in one class when compared to that in the other class (e.g. 100:1), the model is more likely to predict most instances as belonging to the majority class since, due to the lack of data, it fails to correctly predict the minority class. The ideal case would be that in which we have 1:1.

Outliers, missing values, and class imbalance, will be addressed during pre-processing.

4.1.3 *Univariate Analysis: Graphical*

Distribution of feature values when comparing benign and malignant cases was analysed to determine if there was a significant difference between them. Histogram plotting was used for this purpose. Only mean values for each of the initial 10 nuclei features was plotted so that a representative picture of the feature distribution could be obtained.

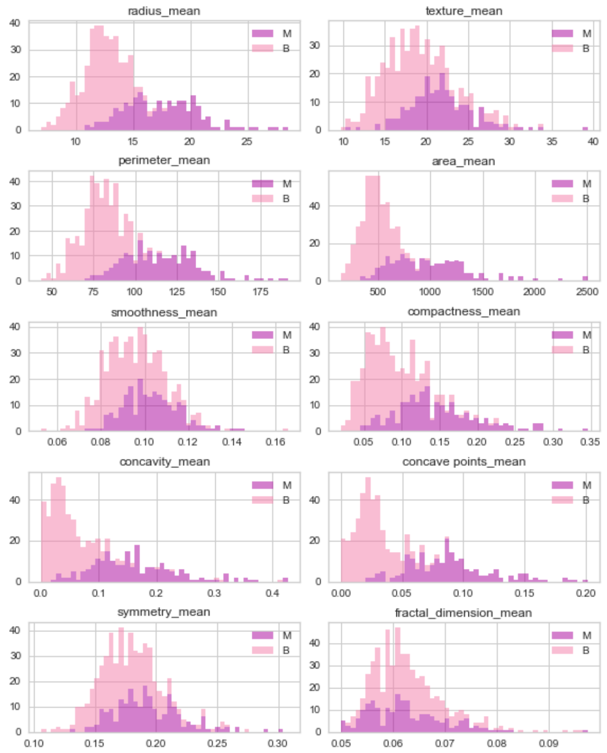


Figure 6. *Histograms of features, malignant vs benignant values.*

After analysing the histograms, some conclusions emerged:

1. Mean values for radius, texture, perimeter, area, compactness, concavity and concave points seemed to be larger in malignant tissue. Therefore, these variables might be useful for the analysis.
2. Mean values for smoothness, symmetry and fractal dimension did not seem to differ significantly between benign and malignant tumors. Therefore, these variables might not be relevant for the analysis.

4.1.4 *Multivariate Analysis: Non-Graphical*

In order to analyze the relationships between pairs of variables, correlation was studied. Correlation can provide information about: analyze the relationships between pairs of variables. Analyzing the correlation between variables can give us information about:

1. How to handle missing data: if two variables are highly correlated, they can be used to predict missing values among themselves.
2. Colinearity between variables: linear models rely on the independency of the variables. The interpretation of a regression coefficient is that it represents the mean change in the dependent variable for each 1 unit change in an independent variable when you hold all of the other independent variables constant. This way, if there exists a significant relationship between predictors, coefficient estimations and predictions can be less precise and less reliable. It also causes overfitting in linear regression analysis models. In this case, we are dealing with a classification problem so non-parametric models were mainly used. Therefore, no significant effect was caused in the analysis.
3. Redundant variables: features that show a very high correlation may suppose adding redundant data to our analysis, since two predictors may be providing the same information about the response variable.
4. Find relevant relationships for interpretation of the analysis.

Following the conclusions of histogram plots, correlation matrix showed that there was a strong positive linear relationship between the target variable, diagnosis, and the mean and worst values for radius, area, perimeter, concavity and concave points (Pearson's coefficient > 0.7), being the coefficient of concave points worst the highest. Therefore, it was deduced that these variables were important for predicting the target variable.

The features that showed the lowest (absolute) correlation coefficients with the variable diagnosis are the mean and worst values of fractal dimension, smoothness, symmetry and texture, being the mean value for fractal dimension the lowest. Therefore, these features were considered as eligible for dropping for a linear model, such as logistic regression. We could not confirm there are no relationships between these features and the target variable, we could only say that there are was no linear relationship.

In general, standard deviation values showed a low correlation with target variable, except for that of radius, perimeter and area.

We could also detect some relationships between predictors:

1. Radius, Perimeter and Area mean and worst values showed a correlation coefficient near to 1, which indicates that these variables had a positive relationship almost perfectly linear. Specially for linear models, this may be problematic because of the effects of colinearity on the estimations of the coefficients. It may also lead to overfitting and feature redundancy.
2. Concave points, concavity and compactness also showed a high correlation among them.

4.1.5 *Multivariate Analysis: Graphical*

2D Scatter plots allowed to characterize the distribution of outcomes with respect to each of the variables. Mean val-

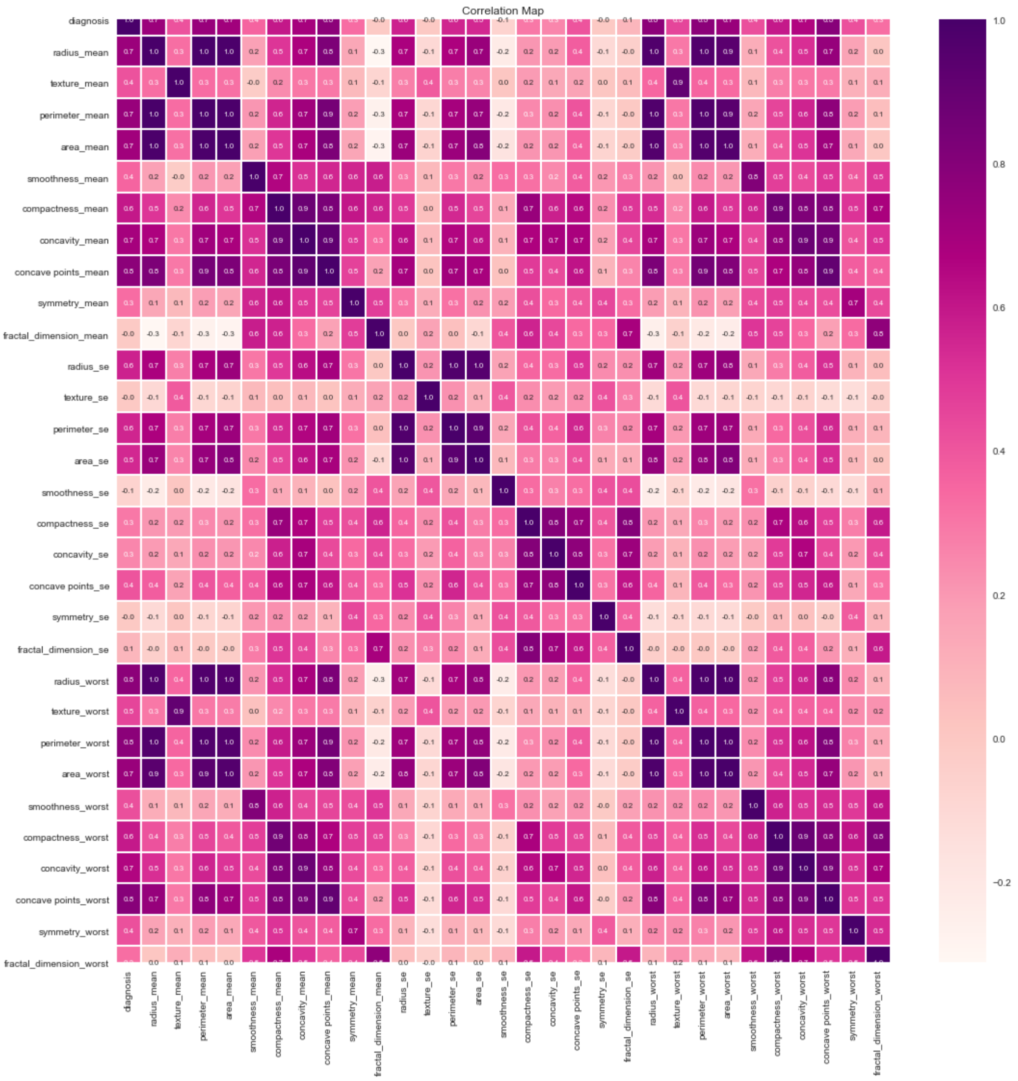


Figure 7. *Correlation matrix with Pearson's coefficients.*

ues for radius, perimeter, area, compactness, concavity and concave points seemed to be suitable for univariate logistic regression, since we clearly saw that samples were distributed in a way that the sigmoid function could be able to finely predict the desirable output.

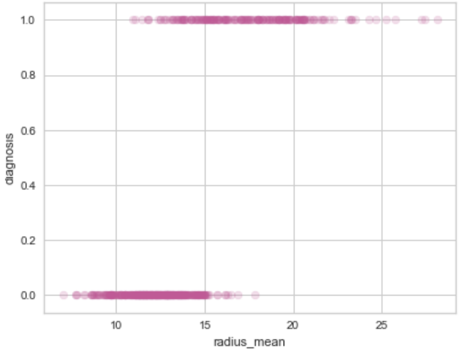


Figure 8. *Scatter plot of diagnosis vs radius mean.*

When analysing scatter plots between pairs of features, it was noticed that most of these distributions showed two clear differentiated clusters corresponding to malignant and non-malignant samples. Therefore, this led us to think that only two features could provide accurate classification performance, and that using all the features for training could result in overfitting due to redundant data.

4.2 Pre-processing

At this point, a sufficient initial knowledge about data distribution had been acquired in order to model a correct pre-processing.

4.2.1 Missing Values

As it has been mentioned during EDA, neither explicit nor implicit missing values are present in the dataset. Therefore, no further processing was needed in this sense.

4.2.2 Outliers

Two steps were followed: 1. Find the proportion of outliers for each variable. 2. Handle outliers: once determined the proportion of outliers for each feature, the methodology used for handling these values is defined.

Assuming a gaussian distribution of the data, we detected outliers by computing the z-score for each feature. Then, we considered as outliers those values that lay further than

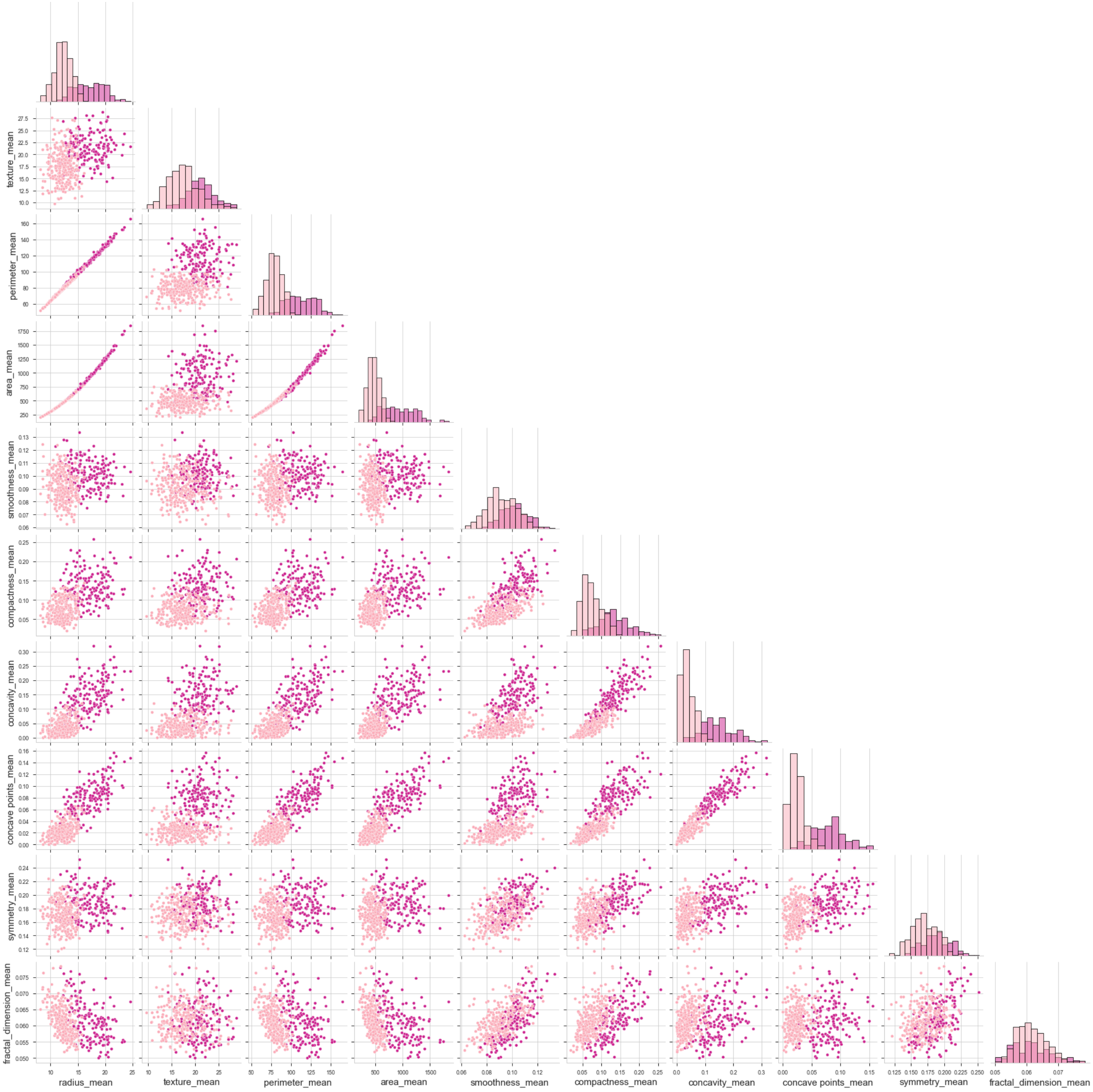


Figure 9. *2D-Scatter plots of features, where diagnosis = B is represented by light pink and diagnosis = M is represented by dark pink.*

2.5 standard deviations from the mean. Since variables showed shifted distributions between target values (B/M), we decided to identify outliers independently for benignant and malignant outcomes. This way, we wouldn't loose data that could have been important for the analysis.

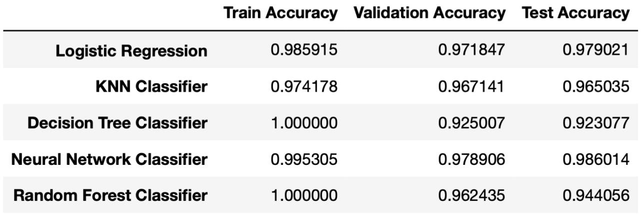
Outliers may not be incorrect values but values that lie far from a population. They may be exceptional values. It is not acceptable to drop an observation just because it is an outlier. Thse values can be legitimate observations that can provide interesting information about the data. For this reason, it's important to further analyse the nature of the outlier to handle it.

Results showed that the proportion of outliers for each variable was very little. If the proportion of outliers had been higher, we would have considered to drop some features but, in this case, the percentage was minimal (0.3-2%). Confidence intervals were swifted to the right for malignant records when compared to non-malignant ones. This fact reinforced the decision to manage outliers separately for both classes. The low number of outliers, together with the absence of outlier management in previous literature for this dataset, led us to think that these values could be actually correct values which lay exceptionally far from the mean. If this is the case, we shouldn't then eliminate them since they could bring some important information about the data distribution and we could make wrong assumptions and obtain bad results in their absence.

Proportion of rows with at least one outlier was 20.5%, which represented a huge part of the data. When managing outliers, in order to implement methods such as mean, median or random imputation, then features are assumed to be independent. As it was explained during EDA, this assumption does not hold for this data, since features are highly correlated. Then, by using these methods we could have introduced bias in our analysis.

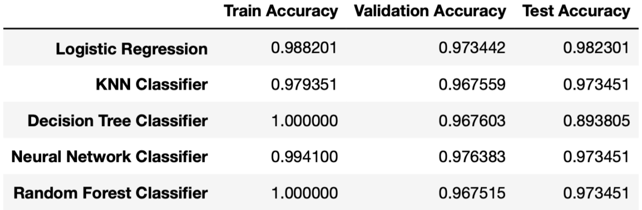
We should then need to handle outliers in a tailored way. Since features were highly correlated, then some features could predict others. However, this method was computationally too expensive and maybe, unnecessary since outliers could be just correct values that were not as frequent as others more close to the mean. Then, the only two options considered were either to eliminate the affected records or to maintain them.

In order to see if the presence or absence of outliers affected model performance, we compared the classifiers that were going to be used during modelling. Results showed that dropping records with outliers slightly improved model performance overall. However, removing such a high percentage of the data is not recommended. Moreover, in this case, outliers did not seem to represent incorrect observations since there was not a significant effect on performance.



	Train Accuracy	Validation Accuracy	Test Accuracy
Logistic Regression	0.985915	0.971847	0.979021
KNN Classifier	0.974178	0.967141	0.965035
Decision Tree Classifier	1.000000	0.925007	0.923077
Neural Network Classifier	0.995305	0.978906	0.986014
Random Forest Classifier	1.000000	0.962435	0.944056

Figure 10. *Performance on original data.*

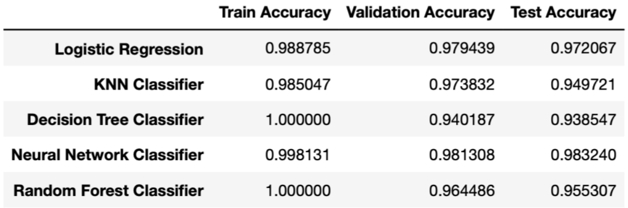


	Train Accuracy	Validation Accuracy	Test Accuracy
Logistic Regression	0.988201	0.973442	0.982301
KNN Classifier	0.979351	0.967559	0.973451
Decision Tree Classifier	1.000000	0.967603	0.893805
Neural Network Classifier	0.994100	0.976383	0.973451
Random Forest Classifier	1.000000	0.967515	0.973451

Figure 11. *Performance on data with row dropping.*

4.2.3 Class Balancing

Class balancing could introduce bias in the analysis. Therefore, with the aim of deciding whether applying class balancing or not, performance was measured for both situations. For class balancing, we used a method called *SMOTE*. This is an oversampling method that synthetically adds new samples for the minority class. Results showed that class balancing improved performance on unseen data overall. Therefore, we maintained the balanced dataset for further steps.



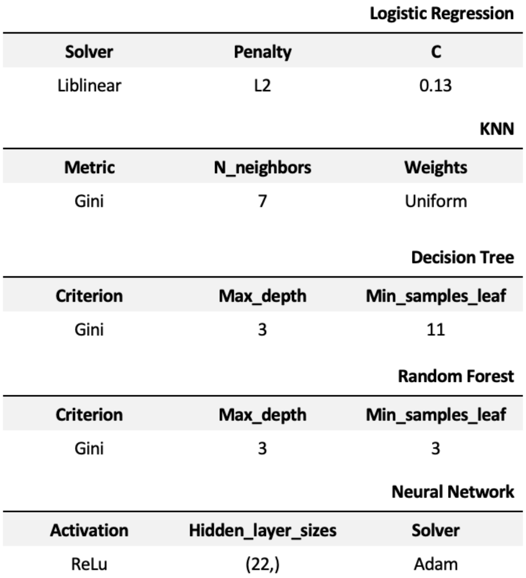
	Train Accuracy	Validation Accuracy	Test Accuracy
Logistic Regression	0.988785	0.979439	0.972067
KNN Classifier	0.985047	0.973832	0.949721
Decision Tree Classifier	1.000000	0.940187	0.938547
Neural Network Classifier	0.998131	0.981308	0.983240
Random Forest Classifier	1.000000	0.964486	0.955307

Figure 12. *Performance on SMOTE-balanced data.*

4.3 Hyperparameter Optimisation

Supervised learning was the approach chosen for classification. This is, a model was trained with the labelled data so that the underlying pattern of this data could serve the model to predict labels correctly by minimizing a cost function that takes into account both predicted and current labels.

Performance of five different models was compared in this work, namely Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Neural Network (NN), and Random Forest (RF) classifiers. The first step was to optimize hyperparameters for all models on accuracy score. Before that, data was divided into train and test sets and later standardised. By using training set, validation curves were plotted to determine the hyperparameters domains for which overfitting wouldn't occur. After that, hyperparameter optimization was performed by means of the method *GridSearchCV*, provided by *scikit-learn* library. For each feature, three hyperparameters were tuned. The remaining ones were set as default.



		Logistic Regression
Solver	Penalty	C
Liblinear	L2	0.13

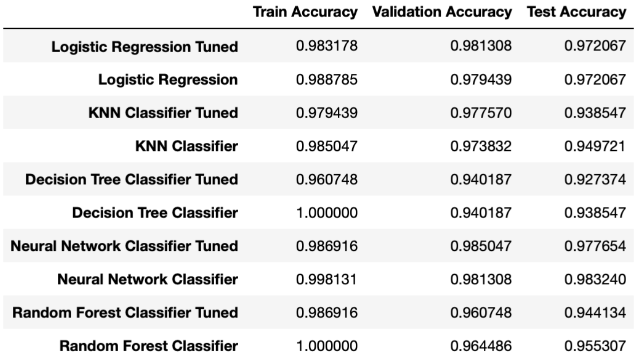
		KNN
Metric	N_neighbors	Weights
Gini	7	Uniform

		Decision Tree
Criterion	Max_depth	Min_samples_leaf
Gini	3	11

		Random Forest
Criterion	Max_depth	Min_samples_leaf
Gini	3	3

		Neural Network
Activation	Hidden_layer_sizes	Solver
ReLu	(22,)	Adam

Figure 13. *Hyperparameters election after GridSearchCV.*

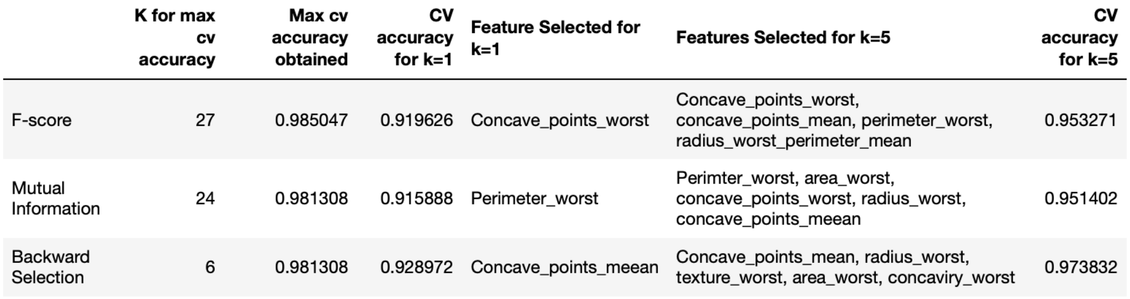


	Train Accuracy	Validation Accuracy	Test Accuracy
Logistic Regression Tuned	0.983178	0.981308	0.972067
Logistic Regression	0.988785	0.979439	0.972067
KNN Classifier Tuned	0.979439	0.977570	0.938547
KNN Classifier	0.985047	0.973832	0.949721
Decision Tree Classifier Tuned	0.960748	0.940187	0.927374
Decision Tree Classifier	1.000000	0.940187	0.938547
Neural Network Classifier Tuned	0.986916	0.985047	0.977654
Neural Network Classifier	0.998131	0.981308	0.983240
Random Forest Classifier Tuned	0.986916	0.960748	0.944134
Random Forest Classifier	1.000000	0.964486	0.955307

Figure 14. *Accuracy scores for models with non-tuned and tuned hyperparameters.*

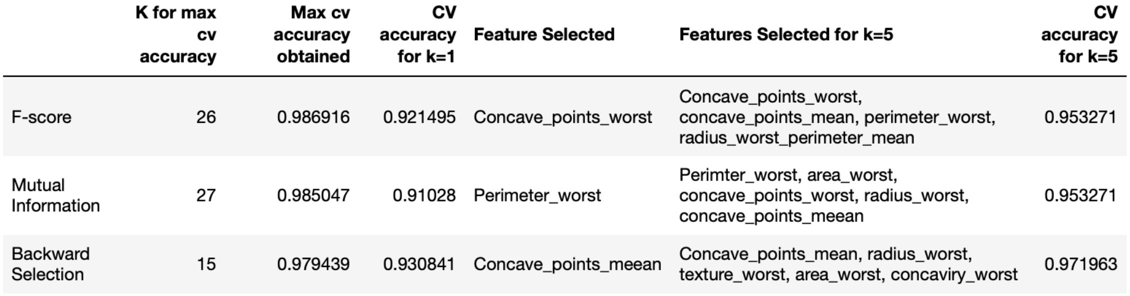
Results from models with hyperparameter tuning were very similar to those with no hyperparameter tuning. In general, we obtained really good performance metrics for every model (>0.9 accuracy, precision, recall, f1 on train, validation, and test), which led us to think that these variables were indeed useful for malignancy and non-malignancy prediction. In addition, hyperparameter tuning seemed to reduce overfitting for DT and RF classifiers.

The models that best performed on unseen data were LR and NN classifiers.



	K for max cv accuracy	Max cv accuracy obtained	CV accuracy for k=1	Feature Selected for k=1	Features Selected for k=5	CV accuracy for k=5
F-score	27	0.985047	0.919626	Concave_points_worst	Concave_points_worst, concave_points_mean, perimeter_worst, radius_worst_perimeter_mean	0.953271
Mutual Information	24	0.981308	0.915888	Perimeter_worst	Perimter_worst, area_worst, concave_points_worst, radius_worst, concave_points_meean	0.951402
Backward Selection	6	0.981308	0.928972	Concave_points_meean	Concave_points_mean, radius_worst, texture_worst, area_worst, concaviry_worst	0.973832

Figure 15. *Feature selection effect on LR classifier metrics.*



	K for max cv accuracy	Max cv accuracy obtained	CV accuracy for k=1	Feature Selected	Features Selected for k=5	CV accuracy for k=5
F-score	26	0.986916	0.921495	Concave_points_worst	Concave_points_worst, concave_points_mean, perimeter_worst, radius_worst_perimeter_mean	0.953271
Mutual Information	27	0.985047	0.91028	Perimeter_worst	Perimter_worst, area_worst, concave_points_worst, radius_worst, concave_points_meean	0.953271
Backward Selection	15	0.979439	0.930841	Concave_points_meean	Concave_points_mean, radius_worst, texture_worst, area_worst, concaviry_worst	0.971963

Figure 16. *Feature selection effect on NN classifier metrics.*

4.4 Interpretability

Explainable AI can be summed up as the process to understand the predictions of an machine learning model. The idea is to make the model as interpretable as possible, which will essentially help in testing its reliability and causality of features.

There are specific libraries dedicated to this topic such as SHAP or LIME. SHAP is more reliable and accurate but LIME is faster, so LIME is sometimes preferrable. In fact, due to the size of our data, using SHAP would have beeb prohibitibe. Hence, we analysed interpretability locally by choosing a small set of samples to see how model predicted on these records.

By inspection, the features that repeated the most among the most important features for each model are worst and mean values for texture, radius, perimeter, concave points and area**, given worst values more importance in general. Fractal dimension and smoothness mean and worst values are also frequent as the main predictors.

These results led us to think that worst values are very useful when determining if nuclei characteristics correspond to malignant cells in tissue samples. Mean values as well, whereas standard error values did not seem to contribute so much to the predictions except for those of area, radius and perimeter.

4.5 Feature Selection

For every dataset, it is possible that some features are irrelevant for the analysis, or these might even introduce some bias. By using so many features for modelling, the ability of the model to generalize could be reduced. Moreover, a high dimensional dataset is commonly computationally expensive. In order to avoid these issues, we performed feature selection.

Three feature selection methods were compared:

1. Filter Methods F-statistic.
2. Filter Methods: Mutual Information.
3. Wrapper Methods: Backward Selection.

Filter methods evaluate the relevance of each variable by individually examining the intrinsic properties of the data. These are the type of methods where individual features are ranked according to predefined relevance score which increases as the relationship between the feature and the target variable increases. The top k features are then selected. Filter methods are easy and fast to implement, but the downside of these methods is that interactions among features features are not considered.

Wrapper methods use combinations of variables to determine predictive power. This is, these methods use a machine learning algorithm of interest as a black box to score subsets of variables according to their predictive power. Unlike filter methods, wrapper methods look for interactions and dependencies between features. However, it is usually computationally expensive since, for each subset evaluation, a new model must be created and the algorithm must be trained and tested to obtain its performance metrics. Moreover, it's usually difficult to interpret and prone to overfit.

Since LR and NN algorithms seemed to perform the best on the data, we analysed how these three different feature selection methods affected their accuracy score when compared to the original non-tuned model. Based on results, several conclusions were drawn:

1. Validation accuracy for just 1 feature selected was, approximately 0.9 for all models.
2. Models with reduced number of features can be able

to perform better than that with all the original features.

3. The highest accuracy score with the highest reduction in number of features (k = #features selected) was achieved by backward feature selection. With only five features, cross-validation accuracy score was 0.97 for both Neural Network and Logistic Regression Classifiers, and 0.93/0.92 for k=1, respectively.
4. Backward feature selection seems to work better on our data.
5. Mean and worst values for features concave points and worst values for area, radius and perimeter seemed to be relevant for our analysis, since these were among the features selected for each method for k < 5.

4.6 Feature Extraction

Feature extraction aims to reduce the number of features in the dataset by creating new features from the existing ones (and then discarding the original features). This new reduced set of features should be able to summarize most of the information contained in the original set of features.

The difference between feature selection and feature extraction is that feature selection aims to rank the importance of the existing features in the dataset and discard less important ones. No features are created and it takes into account labels. On the other hand, feature extraction methods do not use labels but only independent features and aims to create new features by combining existing ones.

Dimensionality reduction techniques allow to transform data from a high-dimensional space into a low-dimensional space so the low-dimensional space retains some meaningful properties of the original data. It allows to eliminate redundant information and therefore avoid some of the issues caused by the curse of dimensionality.

PCA (Principal Component Analysis) is an unsupervised orthogonal linear transformation technique that aims to find the directions of maximum variance or minimum projection error in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one. The objective of PCA is not only to improve the model performance but to obtain similar results with a lower dimension, which will help the model not to overfit and also to reduce computation times. Hence, we looked at PCA from the perspective of improving or providing similar results to those of the non-tuned original model. As in feature selection, since models that performed the best on data were LR and NN, these models were used in this section to check performance before and after PCA.

Based on results, various conclusions were drawn:

1. PC1 accounted for 45.08% explained variance and, approximately, 0.92 accuracy was obtained by selecting only this component.
2. By selecting only PC1 and PC2 together, 0.95-0.96 accuracy was obtained for both models and these components accounted together for 64.1% of the explained variance.
3. Selecting a number of features lower than the original number of features could increase validation accuracy for both models (#PC=9 for Logistic Regression and #PC=14 for Neural Network).
4. For all models, reducing #PC from 30 to 8 lead to virtually the same results than the original model (Fig 19).
5. PCA had a great ability to reduce dimensionality of the dataset while maintaining performance of the model.

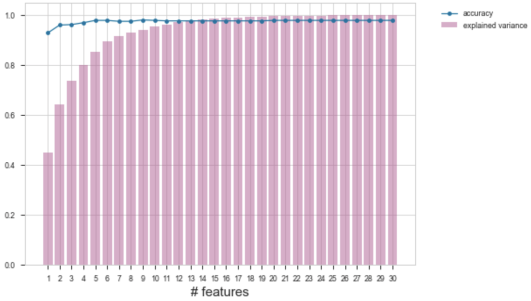


Figure 17. *Principal components, cumulative explained variance and accuracy score obtained for Logistic Regression classifier.*

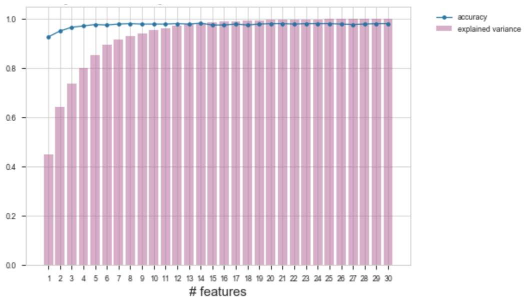
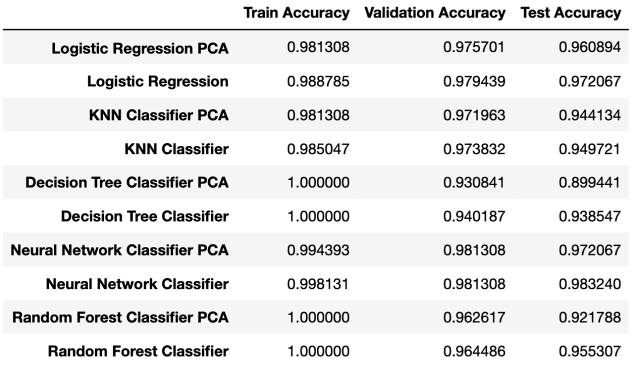


Figure 18. *Principal components, cumulative explained variance and accuracy score obtained for Neural Network classifier.*



	Train Accuracy	Validation Accuracy	Test Accuracy
Logistic Regression PCA	0.981308	0.975701	0.960894
Logistic Regression	0.988785	0.979439	0.972067
KNN Classifier PCA	0.981308	0.971963	0.944134
KNN Classifier	0.985047	0.973832	0.949721
Decision Tree Classifier PCA	1.000000	0.930841	0.899441
Decision Tree Classifier	1.000000	0.940187	0.938547
Neural Network Classifier PCA	0.994393	0.981308	0.972067
Neural Network Classifier	0.998131	0.981308	0.983240
Random Forest Classifier PCA	1.000000	0.962617	0.921788
Random Forest Classifier	1.000000	0.964486	0.955307

Figure 19. *Accuracy scores after selecting first 8 principal components for modelling.*

PCA is also useful for feature selection since eigenvectors tell us about the direction of variance in a way that first principal component points in the direction of maximum variance. As we can see in Fig 20, and consistently with previous conclusions from feature importance, mean and worst values for concavity and concave points contributed the most to this component, so we could say that these features contributed the most to the maximum variance of the data. This result was consistent to that of feature selection methods, and with previous deductions made during the analysis.

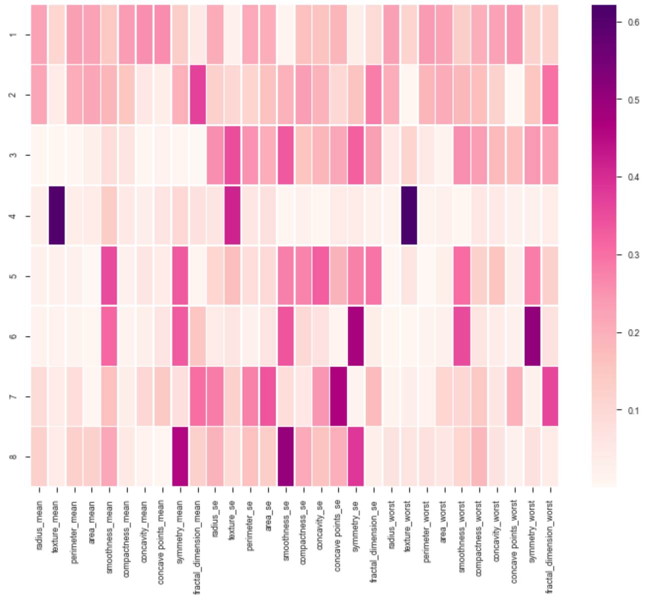


Figure 20. *Eigenvector's coefficients for each of the first eight principal components.*

5 Conclusion

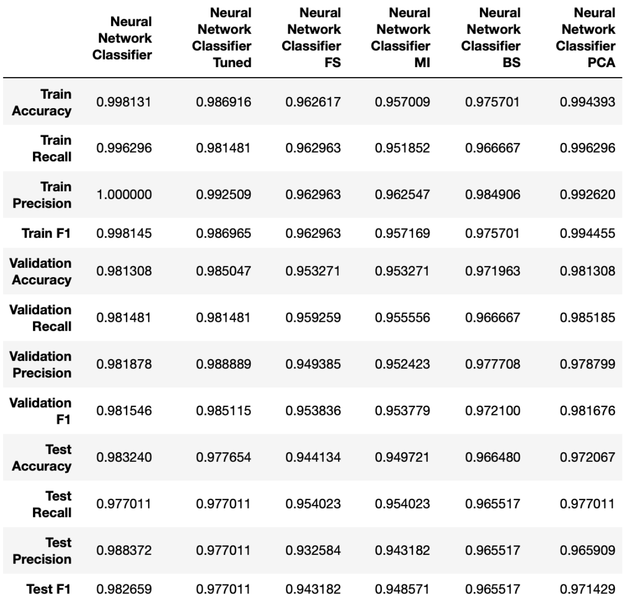
In the lights of the results obtained throughout the analysis of the data, it can be asserted that the *Breast Cancer Wisconsin (Diagnostic) Dataset* has an exceptional quality for its purpose:

1. Data has no missing values: neither explicit nor implicit missing values were present in the dataset.
2. Initially, no significant outliers were found: the elimination of those values firstly considered as outliers for being far from the population didn't change neither the results nor the distribution of the data. Moreover, by consulting available literature, we didn't find almost any paper where outliers were handled.
3. With only one feature, a validation accuracy and f1-score of approximately 0.9 was obtained for LR and NN models.
4. From analysis and visualization, we could crearly see that values for malignant samples are higher than those from benignant ones. This provides very useful information with respect to the behavior of data between these two types of tissue, and for further analysis and interpretation of the results. By plotting histograms for both diagnostics, we could see how different their distributions are.
5. Even with high correlation among features, good results for linear models such as LR.

Among models, Logistic Regression, KNN Classifier, Decision Tree Classifier, Neural Networks Classifier and Random Forest Classifier, all the models obtained, at least, validation accuracy score 0.9, being LR and NN classifiers the models that showed a slightly better cross-validation scores over the rest of the models. This could be considered a great accuracy score by taking into account that basic pre-processing of the data was needed, and that this value is also for non-tuned models.

It seems that worst and mean values predominated over standard error values, so these features might be of special importance when determining a diagnosis. Mean and worst values for concavity and concave points were frequent among most relevant features obtained through the different steps followed, therefore these features could provide useful information when determining the malignancy of the tissue.

By significantly reducing the dimensionality of the dataset, high performance scores could be obtained, even better than those of the original dataset. This way, it can be induced that the dataset contains some irrelevant/redundant features, and that PCA and Feature Selection Methods such as those based on F-score, Mutual Information and Backward Selection work proficiently on our data.



	Neural Network Classifier	Neural Network Classifier Tuned	Neural Network Classifier FS	Neural Network Classifier MI	Neural Network Classifier BS	Neural Network Classifier PCA
Train Accuracy	0.998131	0.986916	0.962617	0.957009	0.975701	0.994393
Train Recall	0.996296	0.981481	0.962963	0.951852	0.966667	0.996296
Train Precision	1.000000	0.992509	0.962963	0.962547	0.984906	0.992620
Train F1	0.998145	0.986965	0.962963	0.957169	0.975701	0.994455
Validation Accuracy	0.981308	0.985047	0.953271	0.953271	0.971963	0.981308
Validation Recall	0.981481	0.981481	0.959259	0.955556	0.966667	0.985185
Validation Precision	0.981878	0.988889	0.949385	0.952423	0.977708	0.978799
Validation F1	0.981546	0.985115	0.953836	0.953779	0.972100	0.981676
Test Accuracy	0.983240	0.977654	0.944134	0.949721	0.966480	0.972067
Test Recall	0.977011	0.977011	0.954023	0.954023	0.965517	0.977011
Test Precision	0.988372	0.977011	0.932584	0.943182	0.965517	0.965909
Test F1	0.982659	0.977011	0.943182	0.948571	0.965517	0.971429

Figure 21. *Neural Network classifier metrics comparison between methods followed throughout the analysis. Feature selection with #features = 5 and PCA with #PC = 8.*

High accuracy, recall, precision and f1-score values obtained for NN classifier demonstrate that artificial intelligence could precisely help classifying samples according to the malignant or non-malignant characteristics of their nuclei in a cost- and time-efficient way. Manual classification of samples requires on-site expertise and results of the analysis may differ with level of expertise of examiner. The pathologist examine the tissue structure, distribution of cells in tissue, regularities of cell shapes and determine benign and malignancy in image. This is very time consuming and more prone to intra- and inter-observer variability [9]. Consistency, scalability, interpretability, and accuracy of machine learning models demonstrate promising results for classification of tissues and tailored therapies.

Acknowledgments

Special thanks to the teachers of the Biomedical Engineering Degree at Rey Juan Carlos University, who provided us with the knowledge and tools to develop this project.

References

[1] WHO, "Breast cancer." https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] "Breast cancer," *Encyclopædia Britannica*, Mar 2021.

[3] C. Ming, V. Viassolo, N. Probst-Hensch, P. Chappuis, I. Dinov, and M. Katapodi, "Machine learning techniques for personalized breast cancer risk prediction: comparison with the bcrat and boadicea models," *Breast Cancer Research*, vol. 21, 06 2019.

[4] K. Gupta and N. Chawla, "Analysis of histopathological images for prediction of breast cancer using traditional classifiers with pre-trained cnn," *Procedia Computer Science*, vol. 167, pp. 878–889, 2020. International Conference on Computational Intelligence and Data Science.

[5] M. Z. Alom, C. Yakopcic, M. Nasrin, T. Taha, and V. Asari, "Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network," *Journal of Digital Imaging*, vol. 32, 02 2019.

[6] UCI, "Breast cancer (diagnostic) data set," 1992.

[7] W. Wolberg, N. Street, D. Heisey, and O. Mangasarian, "Computerized breast cancer diagnosis and prognosis from fine-needle aspirates," *Archives of surgery (Chicago, Ill. : 1960)*, vol. 130, pp. 511–6, 06 1995.

[8] M. Kass, A. P. Witkin, and D. Terzopoulos, "Snakes: Active contour models.," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.

[9] A. Belsare and M. Mushrif, "Histopathological image analysis using image processing techniques: An overview," *Signal Image Process Int J*, vol. 3, 11 2011.