Universidad Autónoma
de Madrid

Escuela Politécnica Superior

23
24

Bachelor thesis

Comparative Analysis of Classifiers for Breast Cancer Detection with Visualizations

Iván Sotillo del Horno

www.uam.es

Universidad Autónoma de Madrid

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Bachelor as Ingeniería Informática (modalidad bilingüe)

BACHELOR THESIS

Comparative Analysis of Classifiers for Breast Cancer Detection with Visualizations

Author: Iván Sotillo del Horno
Advisor: Alejandro Bellogín Kouki

marzo 2024

Francisco Tomás y Valiente, n^{o} 1
Madrid, 28049
Spain

Iván Sotillo del Horno
Comparative Analysis of Classifiers for Breast Cancer Detection with Visualizations

Iván Sotillo del Horno

PRINTED IN SPAIN

A mi madre y a mi abuela, cuya lucha contra el cáncer de mama me ha inspirado a realizar este trabajo.

Resumen

Esta tesis presenta un análisis comparativo de clasificadores para la detección del cáncer de mama y el uso de Inteligencia Artificial Explicable (XAI) para interpretar los resultados. En la fase inicial se realizará la construcción y optimización de los modelos de clasificación, estos clasificadores analizarán los resultados de las biopsias de aguja fina y clasificarán las muestras como benignas o malignas.

Posteriormente, se realiza una comparación de rendimiento comparando métricas como la puntuación F1 o la *recall*. El objetivo es identificar el mejor clasificador de acuerdo a nuestras métricas. Una vez encontrado el mejor modelo, nos adentramos más en él para entender cómo funciona. Para esto, utilizaremos SHAP (SHapley Additive exPlanations), un método de XAI que nos permite ver la importancia de cada característica y cómo contribuyen a la decisión final del modelo. Esto nos permitirá no solo clasificar las muestras, sino también entender por qué el modelo ha tomado esa decisión, lo que puede ser un avance en la comprensión de los modelos de IA para fines médicos.

Palabras clave

Detección de Cáncer de Mama, Clasificadores, Análisis Comparativo, Interpretabilidad, SHAP, IA Explicable, Visualización

Abstract

This thesis presents a comparative analysis of base and ensemble classifiers for breast cancer detection and the use of eXplainable AI (XAI) to interpret the results. The initial phase involves constructing and optimizing the classifier models, these classifiers will analyze the results from fine needle biopsy aspirations and classify the samples as benign or malignant.

Following this, a performance comparison is conducted comparing metrics such as the F1 score or the recall. The aim is to identify the best classifier regarding our metrics. Once the best classifier model is found, we dive deeper into it to understand how it works. For this, we will use SHAP (SHapley Additive exPlanations), a method of XAI (eXplainable AI) that allows us to see the importance of each feature, and how they contribute to the final decision of the model. This will allow us to not only classify the samples but also to understand why the model has made that decision which can be a step forward in understanding AI models for medical purposes.

Keywords

Breast Cancer Detection, Classifiers, Comparative Analysis, Interpretability, SHAP, eXplainable AI, Visualization

Table of Contents

Lists

List of algorithms

List of codes

List of equations

List of figures

List of tables

1

Introduction

1.1. Motivation

Breast cancer is the most common cancer type among women [1]; in 2020, there were more than 2.26 million women diagnosed with breast cancer [1], being the second leading cause of death among women in the United States [2]. Early detection is a crucial step for improving survival rates. With the current analysis techniques of FNAB (Fine Needle Aspiration Biopsy), we have a sensitivity (ability of a test to identify positive cases correctly) of 0.927 [3]. Therefore, there is a need for a more accurate interpretation of those tests.

Machine learning is a branch of artificial intelligence that focuses on developing algorithms that can learn from data and extract patterns from it to be then able to generalize it to unseen data. In this case, we care about classifiers, whose potential is in the ability to learn from a dataset and then on unseen data being able to classify it as one class or another; in this case, we will be able to classify as benign or malign the results of a fine needle aspiration.

The potential of classifiers in breast cancer detection is immense. However, the effectiveness of the different classifiers can vary; this is why it is crucial to understand how each classifier works, how to tweak it, and how to make them as precise and effective as possible, which is the goal of this thesis.

Finding the best possible classifier for this problem would impact cancer detection tasks, facilitating healthcare professionals in their diagnostic responsibilities and, ultimately, improving patient outcomes.

1.2. Objectives

1.3. Structure of the document

2

State of the Art

2.1. Base and Ensemble Classifiers

2.1.1. Base Classifiers

2.1.2. Ensemble Classifiers

2.2. Classifier Optimization

2.3. Evaluation of Classifiers

2.4. Explainable AI

2.4.1. Explainable AI

2.4.2. SHAP

2.5. Web Application Development

3

Design and Implementation

In this section the design of the project will be shown, showing the structure of the project, its requirements, and the implementation of the different parts of the project. Additionally, the implementation of each part will be explained.

3.1. Project Structure

The structure of the project is divided into four main modules corresponding to the main stages of the project: dataset, classifiers, Explainable Artificial Intelligence (XAI), and web application. The dataset module is responsible for loading the dataset and splitting it into a 70 % training set and a 30 % test set. Then in the classifiers module, we do the Exploratory Data Analysis (EDA), data preprocessing, and building and optimizing the classifiers. The XAI module is responsible for the implementation of the SHAP algorithm and the analysis of the results. Finally, the web application module is responsible for the development of the web application. The structure of the project is shown in Figure 3.1. Now we will describe each module in more detail.

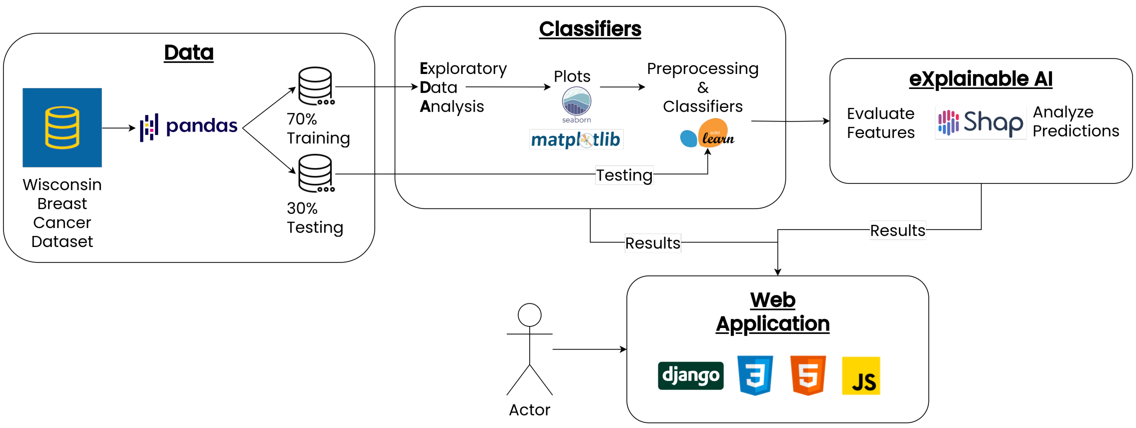

Figure 3.1: Structure of the project

1. **Dataset module**: This module is responsible for loading the dataset and splitting it into a 70 % training set and a 30 % test set. The dataset used in this project is the Breast Cancer Wisconsin dataset

[4], which contains the parameters of the cell nuclei of the sample obtained by a Fine Needle Aspiration Biopsy (FNAB) of breast masses. The dataset is loaded using the *pandas* library. The dataset is then split into a 70 % training set and a 30 % test set using *scikit-learn*. From now on, only the training will be used for the Exploratory Data Analysis, data preprocessing, and building and optimizing the classifiers. The test set will be used to evaluate the classifiers. This ensures that we avoid data leakage. An example of a FNAB is shown in Figure 3.2.

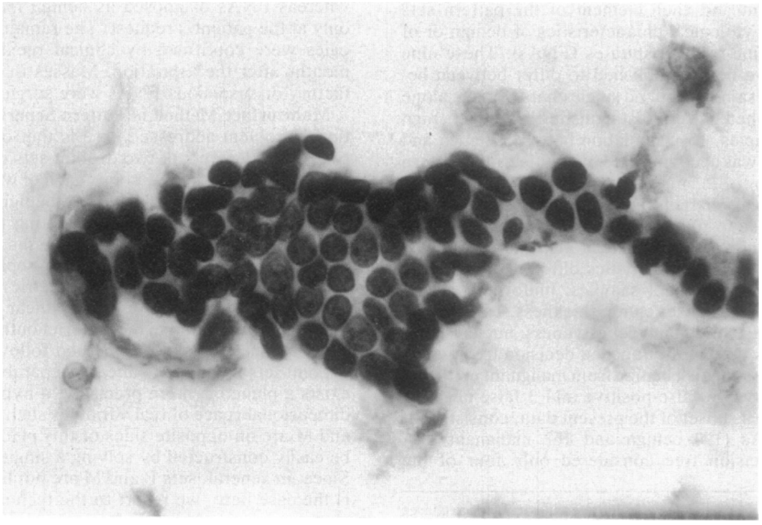

Figure 3.2: Fine Needle Aspiration Biopsy [5]

2. **Classifiers module**: This module is responsible for the EDA, data preprocessing, and building and optimizing the classifiers. The EDA is carried out using *matplotlib* and *seaborn* to obtain descriptive and visual statistics of the dataset. The data preprocessing is carried out using *scikit-learn* to scale and normalize the data and to apply Principal Component Analysis (PCA), this is made using pipelines that facilitate the process of building and optimizing the classifiers and their reproducibility. The building and optimization of the classifiers are carried out using *scikit-learn* using the training set to build and optimize the classifiers. Once the classifiers are built and optimized, they are evaluated using the test set.

3. **XAI module**: This module is responsible for the implementation of the SHapley Additive exPlanations (SHAP) algorithm and the analysis of the results. This is implemented using the *shap* library, which is a Python library that allows us to calculate the SHAP values of the classifiers, this values will be used to understand the importance of the features for each classifier and to understand how a prediction was made. This is what makes this project more interpretable than just using classifiers since this will allow a doctor to understand why a prediction was made.

4. **Web application module**: This module is responsible for the development of the web application. This website was made using *Django* as the backend, *HTML*, *CSS*, and *JavaScript* as the frontend. The website is mainly a way of visualizing the results of the project, it allows the user to see a list of all classifiers and their metrics, which features are the most important for that classifier, and how some

predictions were made. It also allows the user to compare two classifiers and see the differences in their metrics and the importance of the features.

3.2. Requirements

In this section, we will define the requirements of the project. We will divide them into functional (what the system should do) and non-functional (how the system should do it) requirements.

3.2.1. Functional Requirements

Dataset

- **FR-1.-** The application should allow the use of the Wisconsin breast cancer dataset or one with the same features.
- **FR-2.-** The dataset should be split into a training and testing dataset using a 70-30 ratio.

Classifiers

- **FR-3.-** The application should employ the features provided in the dataset for classifier comparison and XAI.
- **FR-4.-** The application should work with multiple classifiers for breast cancer detection.
- **FR-5.-** The application should preprocess the dataset before feeding it to the classifiers. This may include steps such as normalization, and feature selection, focusing on reproducibility and repeatability.
- **FR-6.-** The application should use a test dataset for classifier comparison.
- **FR-7.-** The application should use evaluation metrics to determine the effectiveness of the classifiers taking into account the nature of the problem.

Explainable Artificial Intelligence

- **FR-8.-** The application should show which features are the most important for a classifier.
- **FR-9.-** The application should show how sure the classifier was of its prediction.
- **FR-10.-** The application should show which features determined if the sample was benign or malignant.

Web Application

- **FR-11.-** The web application should have a home page with a slideshow with information.
- **FR-12.-** The web application should display a list with all the classifiers.
- **FR-13.-** The web application should show a details page about each classifier, showing its metrics, most important features, and examples of decisions made by the classifier.
- **FR-14.-** The web application should compare any two classifiers showing their metrics and most important features.

3.2.2. Non-Functional Requirements

Classifiers

- **NFR-1.-** The programming language should be Python.
- **NFR-2.-** The classifiers should be created and trained in a Jupiter Notebook.
- **NFR-3.-** The Pandas library should be used for data manipulation.
- **NFR-4.-** The Scikit-learn library should be used for the classifiers.
- **NFR-5.-** The Matplotlib and Seaborn libraries should be used for the plots.
- **NFR-6.-** The Shap library should be used to understand the feature importance of the models and how a sample decision was made.

Web Application

- **NFR-7.-** The web application should be designed for desktop usage.
- **NFR-8.-** The web application should have a pink color palette and a modern design.
- **NFR-9.-** The web application should be implemented in Django.
- **NFR-10.-** The web application should have an intuitive interface.

3.3. Exploratory Data Analysis

3.3.1. Descriptive Statistics

3.3.2. Data Visualization

3.4. Data Preprocessing

3.4.1. Scaling and Normalization

3.4.2. Principal Component Analysis

3.5. Building and Optimizing Classifiers

3.6. SHAP Implementation

3.7. Web Application Development

4

Experiments and Results

4.1. Exploratory Data Analysis

4.2. Classifier Comparison

4.2.1. Base Classifiers Comparison

4.2.2. Ensemble Classifiers Comparison

4.2.3. Choosing the Best Classifier

4.3. SHAP Analysis

4.3.1. Global Interpretability

4.3.2. Local Interpretability

5 Conclusions and Future Work

5.1. Conclusions

5.2. Future Work

Bibliography

[1] WCRF International, “Breast cancer statistics | World Cancer Research Fund International.”

[2] American Cancer Society, “Breast Cancer Statistics | How Common Is Breast Cancer?.”

[3] Y.-H. Yu, W. Wei, and J.-L. Liu, “Diagnostic value of fine-needle aspiration biopsy for breast mass: a systematic review and meta-analysis,” *BMC Cancer*, vol. 12, p. 41, Jan. 2012.

[4] O. M. William Wolberg, “Breast Cancer Wisconsin (Diagnostic),” 1993.

[5] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction,” *BMC Medical Research Methodology*, vol. 19, pp. 1–18, Dec. 2019. Number: 1 Publisher: BioMed Central.

Acronyms

EDA Exploratory Data Analysis.

FNAB Fine Needle Aspiration Biopsy.

PCA Principal Component Analysis.

SHAP SHapley Additive exPlanations.

XAI Explainable Artificial Intelligence.

Appendices

Universidad Autónoma
de Madrid