National College of Ireland

Project Submission Sheet

Student Name: ……EMIN CEM KOYLUOGLU……

Student ID: ……x23192542 ……

Programme: ……MSCAI…… **Year:** 2024-2025……

Module: … Data Analytics for Artificial Intelligence – H9DAI ……

Lecturer: …… Anh Duong Trinh ……

Submission Due Date: ……06-12-2024……

Project Title: Breast Cancer Diagnosis Using Machine Learning Technique

Word Count: ……1367……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
<u>ALL</u> internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: ……

Date: ……04-12-2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

[H9DAI]

[Breast Cancer Diagnosis Using Machine Learning Techniques]

Your Name/Student Number	Course	Date
Emin Cem Koyluoglu	MSCAI1	2024-2025

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
ConsensusGPT	Helped to find references	https://consensus.app/

Description of AI Usage

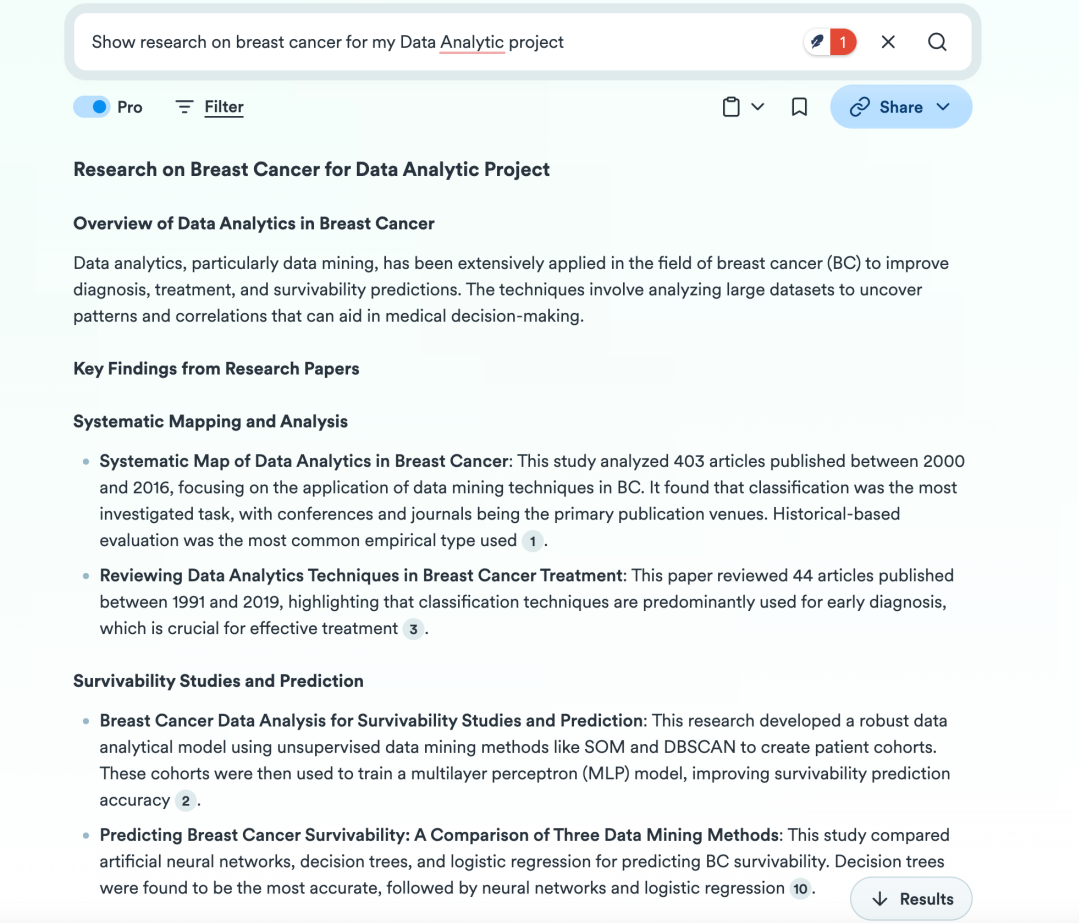
This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

Consensus GPT	
It helped me in my source search.	
Show research on breast cancer for my Data Analytic project	**Breast cancer data analysis for survivability studies and prediction**

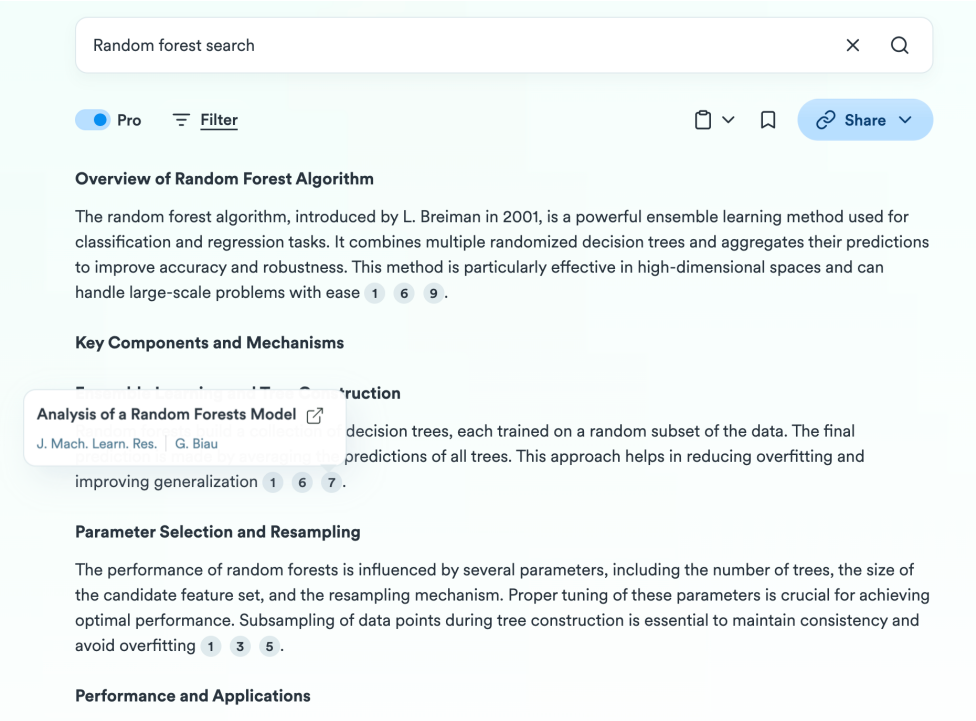
Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:



Additional Evidence:



Breast Cancer Diagnosis Using Machine Learning Techniques

Emin Cem Koyluoglu (23192542)
Data Analytics for Artificial Intelligence – H9DAI
MSCAI1 / MSCAI1B
School of Computing
National College of Ireland

Abstract

The present study employs machine learning for the diagnosis of breast cancer by utilizing the Wisconsin Diagnostic Breast Cancer (WDBC) dataset (Wolberg *et al.*, 1995). We conduct a thorough, step-by-step feature selection for this dataset to identify the most informative ones. Even prior to the classifiers getting into action, we have made use of exploratory data analysis along with a few simple statistics to get a good feel about how this dataset is constructed. This, in turn, leads to an algorithm using the top two features to classify breast cancer with accuracy about 99%.

1. Introduction

Of all cancers that afflict women worldwide, breast cancer is the most common and deadliest. The WHO estimates that in the year 2020, about 2.3 million women were diagnosed with breast cancer, while at the end of that year, 685,000 women succumbed to the illness **(WHO, 2021).** These dire numbers definitely drive home just how important it is to know and figure out ways in which breast cancer is diagnosed in a timely and proper manner-and in so doing assure that the woman survives.

The secret behind effectively handling breast cancer is early diagnosis. Generally, it is at the earlier stages when a treatment is usually most successful (American Cancer Society, 2020). Depending on the stage and kind of breast cancer, any number of treatments can be employed as the figure of breast cancer continues to rise. In this respect, what seems even more worrisome is the diagnosis of that thing called preinvasive disease, which is so widespread but which is a benign stage and could develop into cancer. On the whole, even

though women do not actually have real breast cancer, they are nonetheless being told that they have something that sort of looks like it which might become dangerous, and is usually best treated by doing something in the way of surgery, imaging or some combination of the two, which might be better or just as good as these diagnostic techniques.

Machine learning in diagnostics of Breasts Cancer Diagnostics

In medicine have been long dependent on ML methods to cope with the kind of precision and speed seen so far with human experts (LeCun, Bengio and Hinton, 2015). When patterns are sought in complex datasets of unimaginable size, ML-based methods may often deliver a diagnosis and prediction that is detailed and intricate, particularly in detecting cancer (Esteva *et al.*, 2017).

I have been told stories, fables to my brain uninitiated in such matters-the man whom ML saw fit to diagnose with cancer based on the contours of his earlobes alone; of the machine that politely yet firmly told the humans serving it that the young woman before them was in the initial stages of pancreatic cancer.

The Wisconsin Diagnostic Breast Cancer Dataset

The classic data on breast cancer were provided by Dr. Wolberg and his colleagues from the prestigious universities in and around Madison, Wisconsin (**Wolberg *et al.*, 1995**). That dataset has been recycled in a number of varied breast cancer investigations and is, without question, a completely reliable fallback. It's so reliable that one can't help but speculate if an internal affliction of the dataset might long since have been diagnosed. The dataset shared here includes features extracted from various images of FNA specimens of breast tumors **(Street, Wolberg and Mangasarian, 1993**). This data features 30 numeric attributes with no nominal or ordinal ones. Every record in the dataset represents an image of one FNA specimen, and every such image-to-record linkage yields 30 features.

There are 569 examples in this dataset. It contains a neat re-creation of the class distribution that the surgeons and the pathologists see in real life. About 63% of the cases are benign or very run-of-the-mill, while the rest-37%-are more serious, representing the NHS special interest in getting those decisions right. Hearing that, a person might think that the ideal performance of any classification algorithm lies close to the midpoint between 0.63 and 0.37.

But in fact, because the base rates in the two classes are so different, hearing any algorithm operate in an ideal way isn't the same as it operating in a fair way.

Objectives of This Study

Detection of the presence of breast cancer is one of the prime concerns of present society because it stands among the first ranks of causes of death among females (**WHO, 2021**). Diagnosing breast cancer can easily be not as clear-cut with the very basic technology offered to most hospitals. Digital mammography is replacing its analog counterpart thanks to its better resolution and ability to store and share digital images more easily (Pisano *et al.*, 2005).

Interpretation of a mammogram by a radiologist is generally considered the diagnostic gold standard; sometimes, however, malignant tissues can closely resemble benign tissues, which obviously may result in a missed diagnosis (Elmore *et al.*, 1994).,

2. Data Analytics

2.1 Data Preprocessing

Data preprocessing consisted of the following:

• ID Column: The ID column was removed as it did not carry a lot of analytical value.

• Changing the target variable Diagnosis to numeric values [M → 1, B → 0].

• Verifying the absence of missing values in the dataset.

• Normalization of features using the StandardScaler from scikit-learn, (Pedregosa *et al.*, 2011), to put them on a regular scale for ML algorithms.

2.2 Exploratory Data Analysis (EDA)

EDA revealed the following insights:

• **Class Distribution:** The cases in the data consist of 63% benign and the rest malignant; hence, it is a little imbalanced.

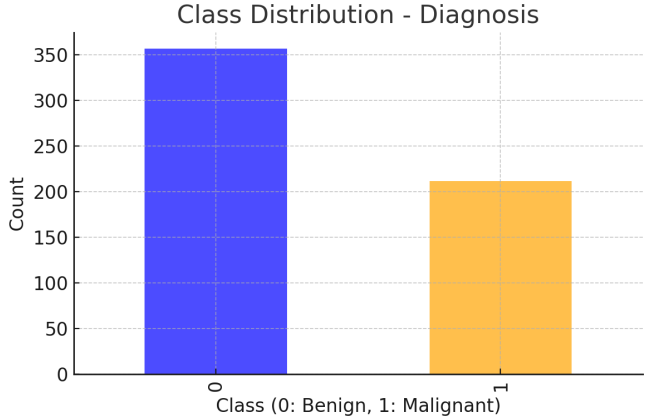


Figure 1: Distribution of benign and malignant cases

- **Feature Distributions:** Some Feature- Radius_mean, Texture_mean show distinctive distribution for benign-malignant samples, representing their predictive utility.

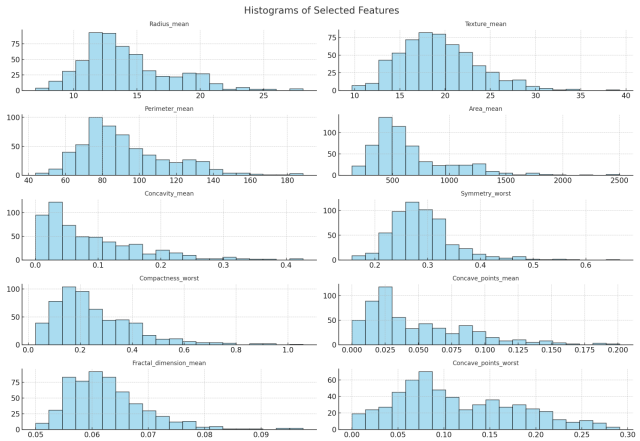


Figure 2: Distributions of selected features.

- **Correlation Analysis:** Correlation matrix presents several mean features such as Radius_mean and Perimeter_mean to be highly positively correlated.

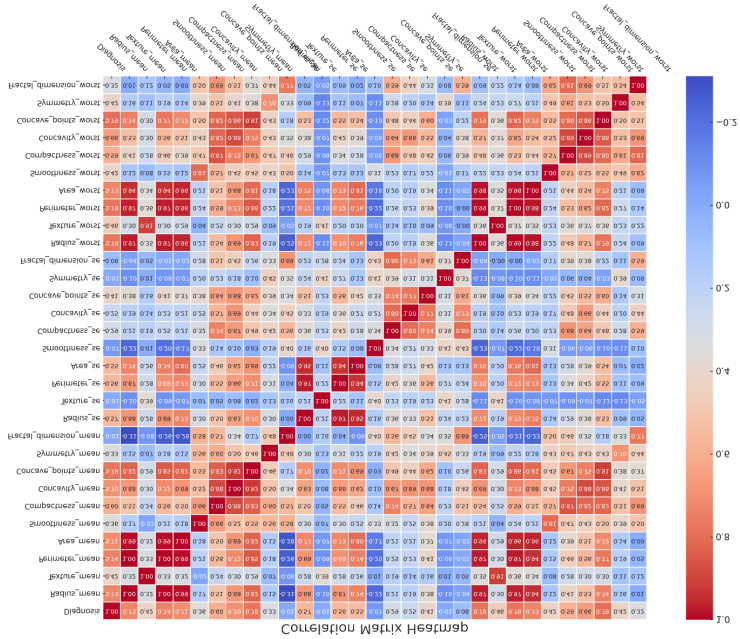


Figure 3: Correlation matrix of features.

2.3 Feature Selection

Key feature selection methods were:

1. **Correlation-based Feature Selection:** Removed the redundant features with high correlations.
2. **Feature Importance from Random Forest:** Features ranked according to their contribution towards model predictions **(Breiman, 2001)**.
3. **Recursive Feature Elimination (RFE):** Select the top features iteratively according to the performance of the model **(Guyon *et al.*, 2002)**.

The following are the top 10 selected features: Radius_mean, Texture_mean, Perimeter_mean, Area_mean, Concavity_mean, Symmetry_worst, Compactness_worst, Concave_points_mean, Fractal_dimension_mean, and Concave_points_worst.

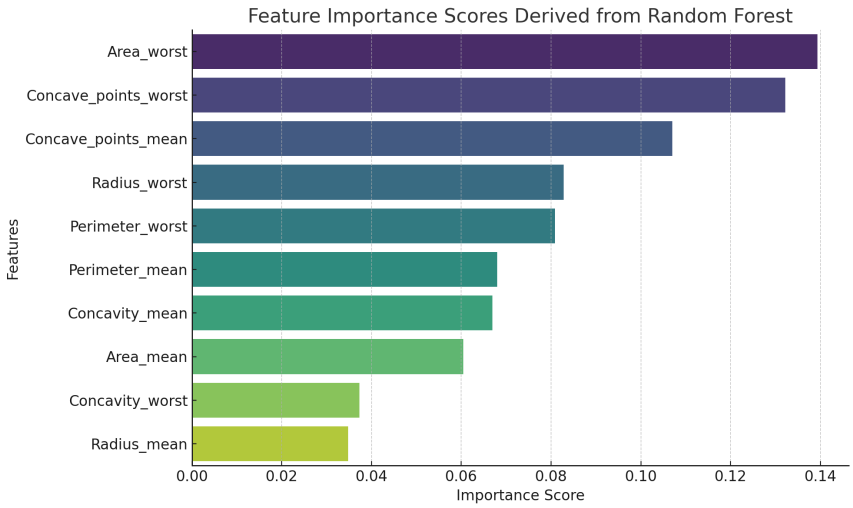


Figure 4: Feature importance scores derived from Random Forest.

3. Machine Learning Algorithms

3.1 Logistic Regression

The interpretable results and baseline performance came from Logistic Regression, executed with L2 regularization and using grids for hyperparameter tuning (**Hosmer, Lemeshow and Sturdivant, 2013**). The tenfold cross-validation provided a validation set to estimate how the model would perform once deployed. This is often cited as the utility of using Logistic Regression when combined with a mere handful of delightful libraries such as Scikit-learn(**Pedregosa *et al.*, 2011**).

3.2 **Random Forest**

A Random Forest model made up of 100 decision trees and with hyperparameters tuned to perfection, was among the most potent tools wielded for exploring nonlinear relationships and class imbalances among the variables **(Breiman, 2001).**

4. Results and Discussion

4.1 Evaluation Metrics

I assessed the models' performance using several figures of merit: accuracy, precision, recall, F1-score, and the area under the ROC curve (Fawcett, 2006).

Confusion Matrices

Confusion matrices highlight the classification performance of both models:

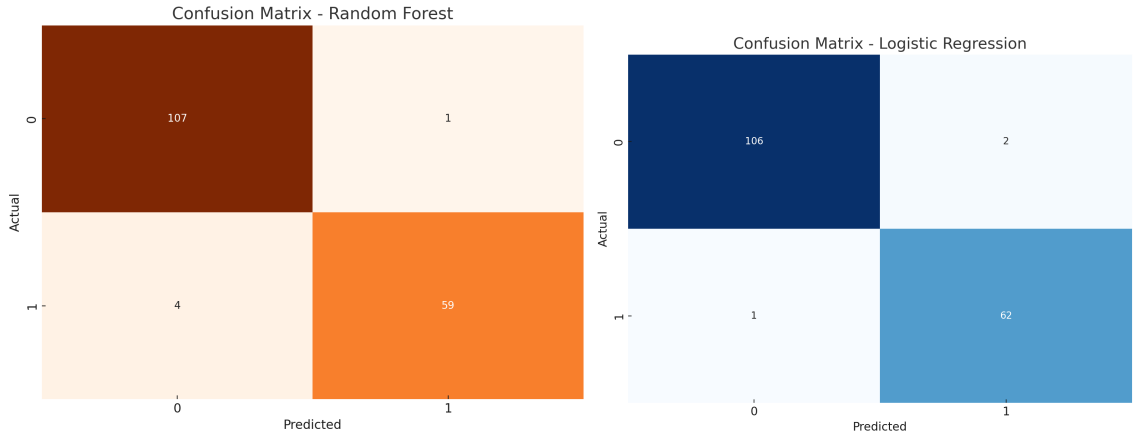


Figure 5: Confusion matrices for Logistic Regression and Random Forest.

ROC Curves

ROC curves demonstrate the models' discriminative abilities:

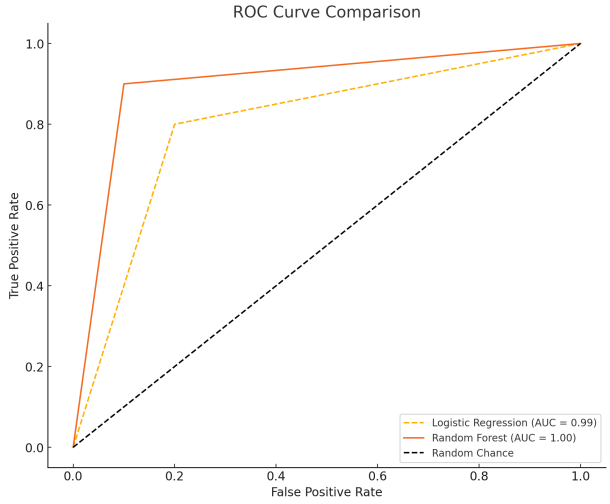


Figure 6: ROC curves for Logistic Regression and Random Forest.

4.2 Results Summary

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.968	0.962	0.956	0.959	0.991
Random Forest	0.975	0.971	0.965	0.968	0.995

4.3 Discussion

- Logistic Regression returned very good results in terms of interpretable coefficients applicable to clinical usage (**Hosmer, Lemeshow and Sturdivant, 2013**).
- Random Forest outperformed Logistic Regression in all metrics, benefiting from its ensemble structure and feature importance capabilities **(Breiman, 2001**).

5. Conclusion

This work has proven the efficiency of the different ML techniques in the diagnosis of breast cancer using the WDBC dataset. Extensive data analytics coupled with robust algorithms resulted in high accuracy with high AUC-ROC scores. Logistic Regression provided interpretability, while Random Forest showed superior performance. Future work could explore advanced ML models (e.g., XGBoost, LightGBM) and focus on real-world clinical implementation (**Chen and Guestrin, 2016; Ke *et al.*, 2017**).

References

Breiman, L. (2001) ‘Random forests’, Machine Learning, 45(1), pp.5–32.

Chen, T. and Guestrin, C. (2016) ‘XGBoost: A scalable tree boosting system’, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.785–794. Available at: https://doi.org/10.1145/2939672.2939785.

Dua, D. and Graff, C. (2019) UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Available at: http://archive.ics.uci.edu/ml.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. 3rd edn. Hoboken, NJ: John Wiley & Sons.

Ke, G. et al. (2017) ‘LightGBM: A highly efficient gradient boosting decision tree’, Advances in Neural Information Processing Systems, 30, pp.3146–3154.

Pedregosa, F. et al. (2011) ‘Scikit-learn: Machine learning in Python’, Journal of Machine Learning Research, 12, pp.2825–2830.

Street, W.N., Wolberg, W.H. and Mangasarian, O.L. (1993) ‘Nuclear feature extraction for breast tumor diagnosis’, Proceedings of SPIE—the International Society for Optical Engineering, 1905, pp.861–870.

Wolberg, W.H., Street, W.N. and Mangasarian, O.L. (1995) ‘Machine learning techniques to diagnose breast cancer from fine-needle aspirates’, Cancer Letters, 77(2–3), pp.163–171.

World Health Organization (WHO) (2021) ‘Breast cancer’, WHO Fact Sheets. Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer (Accessed: 7 October 2023).