

Universidade do Minho
Escola de Engenharia

Rute Salomé Pires de Castro

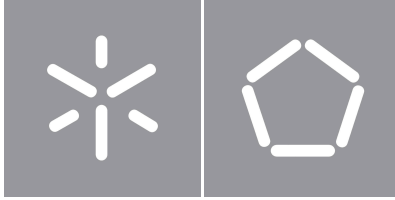
Development of an Artificial intelligence
Breast Cancer diagnostic tool

Development of an Artificial intelligence
Breast Cancer diagnostic tool

Rute Castro

UMinho | 2023

outubro de 2023



University of Minho
School of Engineering

Rute Salomé Pires de Castro

Development of an Artificial intelligence Breast Cancer diagnostic tool

Masters Dissertation
Master's in Bioinformatics

Dissertation supervised by
Óscar Manuel Lima Dias
Débora Carina Gonçalves de Abreu Ferreira

Copyright and Terms of Use for Third Party Work

This dissertation reports on academic work that can be used by third parties as long as the internationally accepted standards and good practices are respected concerning copyright and related rights.

This work can thereafter be used under the terms established in the license below.

Readers needing authorization conditions not provided for in the indicated licensing should contact the author through the RepositoriUM of the University of Minho.



CC BY-NC-ND

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Acknowledgements

Along the sometimes winding path of this investigation, two names stand out as beacons that guided and illuminated my path: Oscar Dias and Débora Ferreira. To both of them, my deepest gratitude.

To the supervisor Débora Ferreira, for her wisdom and tireless dedication, for believing in me when I doubted myself, and for being more than an advisor, but a mentor in the truest sense of the word. To the supervisor Oscar Dias, whose insight and enthusiasm for the subject were contagious and inspiring. His ability to simultaneously challenge and encourage was fundamental to my academic and personal growth.

Although I express my specific gratitude to these two pillars of my academic journey, I would also like to extend a sincere thank you to everyone who, directly or indirectly, contributed to this journey. Every conversation, every constructive criticism, every word of encouragement, played a role in shaping this work. I would like to thank my family for their unconditional love and for always believing in me, even in the most challenging moments.

And to everyone who crossed my path during this period, who in some way touched my life and left their mark, my most sincere thanks.

On each page of this thesis, in each line of research, there is a little bit of each of you. And for that, I am eternally grateful.

Statement of Integrity

I hereby declare having conducted this academic work with integrity.

I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, Braga, october 2023

Rute Salomé Pires de Castro

Abstract

Breast cancer poses a global healthcare challenge due to its prevalence and multifactorial etiology. Despite advances in early detection and treatment, therapeutic approaches like standardized chemotherapy often fall short, primarily due to tumor heterogeneity and genetic variations. In light of this, personalized treatments are imperative.

To address diagnostic and subtype classification challenges, this study utilized Metabric datasets and applied machine learning models after pre-processing, feature selection and hiperparameters otimization. The K-Nearest Neighbors (KNN) model achieved a 94% accuracy rate in cancer detection with a high recall rate, minimizing false negatives. For subtype classification, the Deep Neural Network (DNN) model excelled with a 95% accuracy and an average macro recall value of 0.96.

The study in the realm of cancer prediction, only two genes were commonly observed. In the part of predict of the subtype revealed a significant overlap of 32 genes with the PAM50 list, affirming genetic consistency in breast cancer.

In conclusion, the integration of KNN and DNN models significantly enhances both diagnostic and therapeutic strategies in breast cancer management. These models not only demonstrated high accuracy but also provided insights into gene-specific roles, potentially revolutionizing personalized, data-driven treatment pathways.

Keywords breast cancer, machine learning models, metabric, prediction, cancer subtype, gene expression

Resumo

O cancro da mama representa um desafio global para a saúde devido à sua prevalência e etiologia multifatorial. Apesar dos avanços na detecção e tratamento precoces, abordagens terapêuticas como a quimioterapia padronizada muitas vezes ficam aquém, principalmente devido à heterogeneidade do tumor e às variações genéticas. Diante disso, tratamentos personalizados são imprescindíveis.

Para enfrentar os desafios de diagnóstico e classificação de subtipos, este estudo utilizou conjuntos de dados Metabric e aplicou modelos de *machine learning* após pré-processamento, seleção de recursos e otimização de hiperparâmetros. O modelo *K-Nearest Neighbor* alcançou uma taxa de precisão de 94% na detecção de cancro com uma alta taxa de recall, minimizando falsos negativos. Para classificação de subtipos, o modelo *Deep Neural Network* se destacou com uma precisão de 95% e um valor médio de recuperação macro de 0,96.

No estudo no domínio da previsão do cancro, apenas dois genes foram comumente observados. Na parte de previsão do subtipo revelou uma sobreposição significativa de 32 genes com a lista PAM50, afirmando consistência genética no cancro de mama.

Em conclusão, a integração dos modelos KNN e DNN melhora significativamente as estratégias diagnósticas e terapêuticas no tratamento do cancro da mama. Esses modelos não apenas demonstraram alta precisão, mas também forneceram insights sobre funções específicas de genes, revolucionando potencialmente os caminhos de tratamento personalizados e baseados em dados.

Palavras-chave cancro da mama, modelos de aprendizagem de máquina, metabric, previsão, subtipo de cancro, expressão genética

Contents

I	Introductory material	1
1	Introduction	2
1.1	Context and Motivation	2
1.2	Objectives	3
1.3	Thesis overview	4
2	State of the Art	5
2.1	Breast Cancer	5
2.2	Omics data	6
2.2.1	Omics databases	7
2.2.2	METABRIC	9
2.3	Artificial Intelligence	13
2.3.1	Supervised machine learning models	15
2.3.2	Unsupervised Machine Learning models	23
2.3.3	Deep Learning	26
2.3.4	Machine learning applied to studying Breast Cancer	28
3	Materials and Methods	30
3.1	Methodology for Cancer Diagnosis	31
3.1.1	Preprocessing	31
3.1.2	Feature selection	31
3.1.3	Models optimization	33
3.1.4	Models evaluation	37
3.2	Methodology for Cancer Subtype	37
3.2.1	Preprocessing	37

3.2.2	Feature selection	38
3.2.3	Models optimization	39
3.2.4	Models evaluation	39
4	Results and Discussion	41
4.1	Determining the best feature selection method	41
4.2	Cancer Diagnosis	43
4.3	Breast Cancer (BC) subtype prediction	56
5	Conclusions and future work	75

List of Figures

1	T-distributed Stochastic Neighbor Embedding of cancer diagnosis part	33
2	T-distributed stochastic neighbour embedding of Cancer Subtype part	39
3	Comparison between different Machine Learning (ML) models, using three different feature selection methods: F-value, Principal Component Analysis (PCA) and Total(all features) and considering four different models: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Linear Regression (LR) and Decision Tree (DT).	42
4	(continued) Comparison between different Machine Learning (ML) models, using three different feature selection methods: F-value, Principal Component Analysis (PCA) and Total (all features) and considering four different models: Random Forests (RF), adaboost, Artificial Neural Networks (ANN) and Deep Neural Network (DNN).	43
5	Confusion Matrix of K-Nearest Neighbors (KNN) model	45
6	Confusion Matrix of adaboost model	48
7	Confusion Matrix of Artificial Neural Networks (ANN) model	50
8	Confusion Matrix of Deep Neural Network (DNN) model	52
9	Confusion Matrix of Support Vector Machines (SVM) model	58
10	Confusion Matrix of Linear Regression (LR) model	61
11	Confusion Matrix of Random Forests (RF) model	64
12	Confusion Matrix of adaboost model	67
13	Confusion Matrix of Artificial Neural Networks (ANN) model	69
14	Confusion Matrix of Deep Neural Network (DNN) model	71

List of Tables

1	Omics Databases	8
2	Characteristics of METABRIC datasets	9
3	Clinical features	11
4	Hyperparameters used in CustomRandomizedSearchCV to search for the optimal parameters	35
5	Optimal parameters for Machine Learning (ML) models obtained through the CustomRandomizedSearchCV class for prediction if a individual has cancer	44
6	Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of K-Nearest Neighbors (KNN) model	46
7	Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of adaboost model	49
8	Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of Artificial Neural Networks (ANN) model	51
9	Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of Deep Neural Network (DNN) model	53
10	Optimal parameters for Machine Learning (ML) models obtained through the CustomRandomizedSearchCV class for prediction of cancer subtype "three-gene classifier" . . .	57
11	Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of Support Vector Machines (SVM) model	59
12	Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of Linear Regression (LR) model	62
13	Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of Random Forests (RF) model	65

14	Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of adaboost model	68
15	Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of Artificial Neural Networks (ANN) model	70
16	Score metrics of Deep Neural Network (DNN) model	72

Acronyms

TCGA	The Cancer Genome Atlas
GEO	Gene Expression Omnibus
HER2	Human Epidermal Growth Factor
PR	Progesterone Receptor
ER	Estrogen Receptor
NST	No Special Type
BC	Breast Cancer
NGS	Next Generation Sequencing
WGS	Whole-Genome Sequencing
WES	Whole-Exome Sequencing
NB	Naive Bayes
DT	Decision Tree
RT	Regression Tree
DNN	Deep Neural Network
CNN	Convolutional Neural Network
ML	Machine Learning
DL	Deep Learning
AI	Artificial Intelligence
NN	Neural Network
ANN	Artificial Neural Networks
KNN	K-Nearest Neighbors
SVM	Support Vector Machines
RF	Random Forests
PCA	Principal Component Analysis
LR	Linear Regression

LOR Logistic Regression

RNN Recurrent Neural Networks

LSTM Long Short-Term Memory

GRU Gated Recurrent Unit

t-SNE T-distributed stochastic neighbour embedding

CNA Copy Number Alteration

Part I

Introductory material

Chapter 1

Introduction

1.1 Context and Motivation

The total collection of DNA instructions present in a cell makes up the genome. The human genome is made up of 23 pairs of chromosomes that are found in the cell's nucleus and one tiny chromosome that is found in the mitochondria. Everything a person needs to grow and operate is encoded in their DNA. Whether DNA contains millions or billions of letters, its transmission down the generations is the primary method for passing down organismal properties.

One of the greatest scientific achievements in history is the Human Genome Project. A multinational team of scientists undertook the project as a biological journey to thoroughly examine the whole DNA (genome) of a particular specie group. The Human Genome Project, which began in October 1990 and ended in April 2003, is best known for its achievement of creating the very first genome sequence for humans. This achievement provided fundamental knowledge about the human blueprint and has since sped up research into human biology and improved medical care [1, 2]. Microarrays and Next Generation Sequencing (NGS) methods, including Whole-Genome Sequencing (WGS), Whole-Exome Sequencing (WES) and Targeted Sequencing are used to generate this sort of data [3].

Although the models cannot be considered completely accurate, they may help doctors and other health science professionals to make the best judgments when treating cancer patients [4]. Our bodies have millions of cells. Cancer starts when cellular changes cause these cells to grow excessively, creating a lump called a primary tumour [5]. Globally, 23% of cancer cases are Breast Cancer (BC), which also kills more women than any other disease [6]. According to the most current data, BC became the most commonly diagnosed cancer in the world in 2020, with 2.3 million new cases and 684,996 fatalities. In South America, Africa, and Asia, BC is more common. Additionally, patients with BC have poorer prognoses and higher rates of recurrence, compared to other cancers, and the current rate of patients living longer than five years is still low [7, 8].

Given these findings, it is critical to learn more about how the subtypes emerge and disappear as well as how genes that govern unchecked cell proliferation are affected by mutations and other forms of functional loss. For instance, deletions and insertions to discover new targeted treatments.

As a result, ML provides an approach to fill these gaps and identify differences in gene expression datasets. It is possible to use ML algorithms to understand genomics activity and find relevant genes as indicators for cancer classification and early diagnosis [9].

K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Tree (DT) and Naive Bayes (NB) are some of the examples of Machine Learning (ML) algorithms that have already been tested for this type of cancer to create predictive breast cancer models with datasets containing gene expression data (as well as other types of data) [4]. This work is particularly important in the context of genomics, which is the study of the entire genome and allows for the identification of genetic variations that have a significant role in illness, therapy response, or patient prognosis.[10].

The main mission is not only to offer something intuitive, but above all to help in the projection of an accurate diagnosis for an individual. The motivation behind this approach is based on the perspective of providing health professionals, in particular physician and pathologists, with a tool (a model, the optimal one) that can act as a second opinion, strengthening confidence in the established diagnosis. This additional confirmation can be crucial, accelerating clinical decisions and, consequently, increasing the effectiveness and success of therapeutic interventions.

1.2 Objectives

The main objective of this project is to develop a BC diagnostic framework. The framework will be based on training and validating ML models based on gene expression data to predict distinct biological activities relevant to our diagnosis. The work will address the following scientific/technical objectives:

- Review the relevant literature on ML models and their applications in cancer classification tasks;
- Search for relevant datasets;
- Create different ML models for breast cancer detection and classification, evaluating their performance in different case studies;
- Develop ML pipelines that encompasses data collection, data preprocessing, feature selection, model training, and validation the model.

1.3 Thesis overview

Chapter 2 addresses the disease, BC with the aim of providing a biological perspective. It also explores omic data and the databases that store this type of data. The second part of this chapter aims to provide the basics of Artificial Intelligence (AI) and its applications. Also, the basics of ML and its applications and the most used supervised and non-supervised models.

Chapter 3 serves to unveil the architecture of the pipelines developed for the project, elucidating each step and the intrinsic details and explain the dataset used.

In chapter 4, the results are presented and discussed.

Chapter 5 presents the main conclusions and future work.

Chapter 2

State of the Art

2.1 Breast Cancer

About 10% of BC are hereditary and linked to a family history due to genetic susceptibility. The primary genetic modifications that result in BC include mutations such as minor insertions or deletions, amplifications, and rearrangements [11].

According to its histological and molecular properties, BC is now categorized into five subtypes [12]:

- Triple negative: absence of expression of the Human Epidermal Growth Factor (HER2), Progesterone Receptor (PR), Estrogen Receptor (ER), high ki67 index (a measure of tumor cell proliferation and growth), and No Special Type (NST);
- Enriched with HER2 (non-luminal): Targeted therapy such as trastuzumab, pertuzumab, lapatinib, neratinib, and trastuzumab emtansine (T-DM1) are effective against HER2 expression, high ki67 index, and NST [13];
- HER2+ luminal type B: expresses ER, PR, and HER2, has a high ki67 index, NST, but pleomorphic cells; responds to targeted therapy;
- HER2- luminal type B: expresses ER and PR but not HER2. High ki67 index, high-risk gene expression signature, NST, and micropapillary (carcinoma made up of tiny tumor cell clusters inside of gaps resembling dilated vascular canals);
- Luminal type A: Comparatively high ER, PR, and HER2 expression to type B, also has a low ki67 index, low risk gene expression signature, NST, tubular cribriform, and classical lobular. It is the most prevalent subtype of BC.

Regarding risk factors, it is known that alcohol use, diet, obesity, and physical inactivity are linked to a

higher possibility of developing this condition, as well as early menstrual beginning, a delayed menopause, and no breastfeed [14].

Mammography is the first particular diagnostic test that is being offered for BC screening [7]. After a nodule has been confirmed, an ultrasound and a needle biopsy are often conducted [15]. There are a number of treatment options available if a BC malignant tumor is confirmed, including surgery (mastectomy), which is done when the tumor is already big and is followed by adjuvant radiation to prevent recurrences. When there is a localized tumor, breast conserving surgery is done, then adjuvant radiation. Adjuvant chemotherapy and/or neoadjuvant chemotherapy are often used in conjunction with surgery to reduce tumor size. When the BC-hormone is positive, hormone treatment is used. This includes endocrine therapy, targeted therapy, and bone-modifying medications [12].

2.2 Omics data

There are various types of omics data, such as epigenomics (which studies epigenetic alterations), transcriptomics (studying the expression of genes through mRNA), proteomics (studying peptides and proteins), metabolomics (studying the substances and metabolites created by metabolism), microbiomics and genomics (studying the genome through DNA analysis) [3].

These omics enable the generation and analysis of incredibly large volumes of data. Omics data become important as patients may react differently to the same therapy. Precision medicine aims at selecting the patient's best suitable treatment based on genotype and gene expression data [16].

The best effective treatment with the fewest side effects is what is often meant by the term "precision medicine". All fields of pathology can profit from the use of precision medicine, but oncology is the most advantageous. As there are several BC subtypes, numerous genes may be mutated in various subgroups [17].

Knowing the specific genes that have been altered in a tumour might be the difference between administering an ineffective treatment, with substantial side effects, or an effective therapy that has minimal side effects. Precision medicine's ability to succeed relies on genomic data. This omics provides the information required to build databases that allow for the personalized treatment of each patient. The two primary omics studied in relation to cancer are genomics that is the study of the entire genome in this case with the purpose of identifying genetic variations that have a significant role in illness, therapy response or patient prognosis and transcriptomics [10]. The best course of treatment can be given by knowing a patient's genotype and gene expression [18].

Omics data supports the development of new drugs, and assist in selecting the appropriate medication, and open the door to the work of diagnosing and categorizing the stage and subtype of cancer.

It is feasible to pinpoint the genes and proteins that cause diseases, by examining the genome and how genes are expressed. When opposed to evaluating random chemicals, the process of creating or identifying a medicine is much simpler by knowing the target protein.

Understanding the causes of illnesses and how to cure them requires the study of mutations. Presently, a variety of methods are used to identify mutations. Some of the current approaches for finding mutations include DNA sequencing techniques WES and WGS, together with bioinformatics tools [19].

NGS offers a superior molecular and genomic examination of tumors compared to Sanger sequencing. DNA/RNA extraction, DNA/RNA quantification, library preparation and enrichment, NGS sequencing, base calling, alignment, quality control, variant calling, and annotation are some of the phases in the NGS workflow [20]. Its use is mostly applied on cancer research leading to a change in cancer diagnosis. Consequently, this technique helps in the decision of which genes that are relevant for a certain condition [21].

2.2.1 Omics databases

NGS offered less expensive methods to study the human genome. The quantity of data pertaining to the fields of genetics rose significantly with the decrease in sequencing costs. The quantity and quality of genomic data began to get more efficient funding from research institutions and businesses. For the analysis of individual variation and a wide range of disorders, especially cancer, the study of many genomes became crucial.

Large volumes of data were and are being created as the cost of getting data decreased thanks to NGS. Because databases act as data repositories and enable data exchange with society, databases are crucial. It makes sense to study the data that already exists rather than invest time and money in gathering new data.

Although new data is always beneficial, analyzing current information might lead to new findings. The community has access to a variety of omics datasets like Gene Expression Omnibus (GEO) (more generic) to cancer-specific databases like The Cancer Genome Atlas (TCGA) [22].

Table 1: Available omics databases and the types of data they store

Type of Database	Name	Characteristics
Genomics	Genomics of Drug Sensitivity in Cancer (GDSC) [23]	Identifies biomarkers that have the potential to identify groups that react better to certain anticancer therapy
Genomics	GenBank [24]	Data from DNA of genetic sequences
Transcriptomics	ArrayExpress [25]	Data repository for functional genomics using microarray and sequencing databases
Transcriptomics	Gene Expression Omnibus (GEO) [26]	Global repository for functional genomics data storage and distribution
Proteomics	Human Protein Atlas (HPA) [27]	Mapping of all human proteins in cells, tissues and organs through various omics technologies
Proteomics	Uniprot [28]	Contains protein sequence and annotation data from several European research institutes
Compounds	PubChem [29]	For chemical molecules including information such as IDs and characteristics
Compounds	ChEMBL [30]	Bioactive chemical compounds having drug-like characteristics are hand-curated in this database
Multiple Omics	Cancer Dependency Map (DepMap) [31]	Detection of genetic and pharmacological correlations in cancer, with the objective of finding genetic targets and predicting therapeutic susceptibility
Multiple Omics	The International Cancer Genome Consortium (ICGC) [32]	Has data from the 3 main omics: genomics, epigenomics and transcriptomics related to cancer genomes
Multiple Omics	Catalogue Of Somatic Mutation In Cancer (COSMIC) [33]	Manually curated data and GWS data on cancer somatic mutations
Multiple Omics	GDC Data Portal [34]	Contains diverse omics cancer data from various projects

2.2.2 METABRIC

METABRIC, or the Molecular Taxonomy of Breast Cancer International Consortium, is a groundbreaking project in breast cancer research. Born from a collaboration between UK and Canadian researchers, this initiative aimed to delve deeper into the molecular complexities of breast cancer than any study before it. The goal was to categorize breast cancer on a molecular level to develop more effective, targeted treatments. This dataset is a goldmine of information, containing both clinical and genomic data from nearly 2000 patients. But it's not just about basic clinical facts like age at diagnosis or hormone status. METABRIC also includes in-depth molecular analysis, essentially serving as a crossroads for four key types of data: genetic expression in cancer patients, their clinical data, copy number alterations, and gene expression data from cancer-free individuals [35].

When it comes to the technical side, high-end genotyping platforms like Illumina microarrays were used to scrutinize the samples. The data gathering process was rigorous, involving complex sequencing, bioinformatics analysis, and validation steps. If you're interested in using this dataset, be prepared for a stringent access process due to data sensitivity, including paperwork and justification for data use. This is the dataset that was used and access can be requested here: <https://ega-archive.org/>. Due to the contract signed to have access to it, it is not possible for me to give access to it [36, 37]. The resume of the characteristics of METABRIC datasets are in table 4.

Though it might not have an ideal number of samples, METABRIC is unmatched in the depth and richness of its data. Despite its limitations, researchers value it for the unparalleled insights it offers.

Table 2: Characteristics of METABRIC datasets

Dataset	Description
Copy Number Alteration (CNA) dataset	22544 features
	2174 patients
Clinical dataset	24 clinical features
	2509 patients
Breast cancer genetic expression dataset	20603 features
	1981 patients
Non-cancerous Genetic Expression dataset	48803 features
	144 individuals

2.2.2.1. Breast cancer Copy Number Alteration (CNA) dataset

Copy Number Alterations Copy Number Alteration (CNA), containing precise information on 22,544 genes for 2,174 patients. While the dataset has some missing values, its significance is paramount. CNAs result in either an increase or a decrease in the number of copies of DNA segments and have a significant role in the development and progression of cancers, including breast cancer [35].

To generate this comprehensive CNA data, METABRIC employed high-resolution genotyping microarrays, specifically the Affymetrix SNP 6.0. This platform offers detailed coverage of the genome. The raw data gathered underwent meticulous bioinformatics processing, including algorithms like Circular Binary Segmentation, to identify areas of the genome where copy numbers were altered. This raw information then passed through additional preprocessing steps, including z-score normalization and quality probe filtering, to prepare it for statistical evaluation. The inclusion of the CNA dataset in cancer predictive models serves multiple purposes. First, it enhances the depth and complexity of the models by capturing molecular details that simpler datasets may overlook. Second, the CNA information allows the identification of genomic patterns related to specific tumor subtypes, thereby improving the accuracy and personalization of predictive models. Finally, the CNA dataset's role in influencing therapeutic response gives the predictive models a heightened level of clinical relevance [35, 37].

2.2.2.2. Breast cancer clinical dataset

This dataset covers several variables related to the clinical status of patients identified in table 3. It contains 24 features about clinical data (including the one with the patient identifier) and the respective information for all 2509 patients, however, part of the information is missing [35].

This data set is crucial for advancing breast cancer treatment. It helps personalize treatments, make more accurate prognoses, and test new therapies in scenarios that mimic real life. Furthermore, by bringing together clinical data with genomic analyses, we gain a more complete understanding of the disease. This, in turn, allows doctors to adjust treatment plans based on actual results, making care more effective. It is an essential tool for both research and clinical care [38, 37].

Table 3: Clinical features of the patients' clinical data dataset

Clinical data	Description
Patient Identifier	Identifier to uniquely specify a patient
Lymph nodes examined positive	N° of positive lymphnodes
Nottingham prognostic index	Formula that combines tumor size, lymph node stage and histological grade to predict the prognosis of BC patients
Cellularity	Tumor Content
Chemotherapy	yes, no
Cohort	Group of individuals who share a common characteristic
ER status measured by IHC	ER status measured by Immunohistochemistry
HER2 status measured by SNP6	HER2 status measured by SNP6
Hormone Therapy	yes, no
Inferred Menopausal State	post, pre
Sex	Female, masculine
Integrative Cluster	Refers to a classification method that combines multiple types of data to identify distinct subgroups within a disease or condition
Age at Diagnosis	Age at Diagnosis
Overall Survival (Months)	Overall survival in months since initial diagnosis
Overall Survival Status	1:DECEASED, 0:LIVING
Pam50 + Claudin-low subtype	Pam50 + Claudin-low subtype
3-Gene classifier subtype	Categorizes cancer subtypes based on the expression of three specific genes
"Patient's Vital Status"	Living, Died of Disease, Died of Other Causes
Primary Tumor Laterality	For tumors in paired organs, designates the side on which the cancer originates
Radio Therapy	Radio Therapy
Tumor Other Histologic Subtype	Text to describe a tumor's histologic subtype or mixed diagnosis that is different from previously specified options
Type of Breast Surgery	mastectomy, breast conserving
Relapse Free Status	Relapse Free Status (Recurred = cases where there was loco-regional relapse, distant relapse or death-specific death)
Relapse Free Status (Months)	Time until last follow-up or relapse (in months)

2.2.2.3. Breast cancer genetic expression dataset

This dataset contains 20603 genes and the respective information of 1981 patients, however some information is missing. These data was Obtained through advanced RNA sequencing techniques and using high-density microarray platforms. Each tumor sample was carefully subjected to a rigorous process involving RNA extraction, amplification and hybridization to ensure maximum quality and accuracy of the data obtained. The expression of each gene is quantified numerically. Additionally, the dataset includes unique identifiers for each patient, making it easy to integrate with other types of data, such as clinical information and CNA [37].

The dataset in question is crucial for several aspects of breast cancer research and treatment. First off, it helps identify specific subtypes of breast cancer, allowing for more personalized treatments. It also serves as a resource for finding new therapeutic targets by highlighting genes that differ between tumor and normal tissues. Understanding how the disease progresses is another perk, as the data can shed light on the aggressiveness of the tumor and its potential to spread or resist treatments. When incorporated into machine learning models, this dataset elevates the game. It fine-tunes the accuracy of classification and prognosis, thanks to its detailed gene expression information. By blending this data with other types like clinical information, the models become multidimensional, revealing intricate interactions that could be crucial for treatment plans. The enriched predictive capabilities offer a fuller understanding of the disease, making the dataset indispensable for both researchers and clinicians [38, 37, 35].

2.2.2.4. Non-cancerous Genetic Expression dataset

This dataset contains 48803 genes and the respective information of 144 individuals. It was obtained in the same way as the breast cancer gene expression dataset. Once again, some information was missing [36].

While having a limited number of samples might seem like a drawback, it's actually essential because it acts as a control group. This helps in several ways. First, it allows for direct comparison with cancer patient data, helping to pinpoint genes that act differently when the disease is present. It also helps to standardize the data, so you can be more confident in interpreting gene expression levels in cancer cases. In terms of machine learning, including this dataset is a game changer. It significantly boosts the accuracy of models designed to distinguish between healthy and diseased states. The control group also provides a way to validate the model's results, making them more robust and reliable. Additionally, the dataset allows for more nuanced analyses that consider variations in gene expression between healthy and sick individuals. This extra layer of context is invaluable for interpreting results [36].

So, all in all, without this dataset, any predictive model for breast cancer would be less accurate and less clinically useful. Its inclusion not only makes the analysis stronger but also increases the real-world applicability of the models developed.

2.3 Artificial Intelligence

The field of study of AI emerged in the 1950s with the aim of developing machines capable of performing tasks that require human intelligence. Scientists and engineers at the time began to investigate the possibility of creating systems that could recognize speech and images, learn, reason and perform other cognitive functions.[39].

A team of computer scientists headed by John McCarthy, who hosted the inaugural AI workshop at Dartmouth College, originally used the term "artificial intelligence" in 1956. The phrase "research and design of electronic systems that may be characterized as intelligent" was used to describe AI. The rise of digital computers and the development of information technology are key factors in the history of AI since they made it possible to handle massive quantities of data and carry out intricate computations [39, 40].

The earliest AI algorithms were built using rule programming and symbolic logic. AI systems, such as machines able to play chess games or assist with medical diagnostic systems, were developed using these methods. New algorithms, such as Neural Network (NN) and other ML algorithms were created throughout time [41].

AI examples that are often used include [40, 42]:

- Virtual assistants:like Google Assistant, Siri, and Alexa to comprehend inquiries, provide answers, and carry out tasks;
- Voice and Speech Recognition: examples include virtual assistants, smart speakers, and mobile;
- Computer vision:includes tools for object identification, face recognition, and picture analysis in medicine;
- Games: include simulation, strategy, and smart board games;
- Autonomous driving:drones, autonomous vehicles, and commercial robots are examples;
- Recommendation Systems:like suggestions for music, movies, and books based on your preferences and use patterns;
- Robotics: include appliances for the kitchen, bathrooms, and operating rooms;

- ML: includes techniques for classification, clustering, prediction and optimization.

Similar to how people tackle problems, AI systems are employed to carry out complicated tasks. A branch of AI known as ML is the capacity of a computer to mimic intelligent human behavior. It is a branch that focuses on examining and deciphering data patterns and structures to facilitate learning, reasoning, and decision-making without the involvement of humans. This is accomplished without explicitly programming all the rules and instructions to execute certain jobs. Alternatively, ML algorithms "study" from data by detecting patterns and trends in the data using statistical methods and algorithms. It contains both basic and complex algorithms like NN, LR, and DT [43, 44].

While ML is a more general discipline that includes many kinds of ML algorithms and approaches, Deep Learning (DL) is a subset of ML that focuses on deep NN [45].

ML comes in a variety of forms, such as reinforcement learning, transfer learning, unsupervised learning, and supervised learning [44, 46, 47]:

- Supervised learning: the system is taught using labeled data in which each sample has a label or classification. The aim is predicting the label for future cases;
- Unsupervised learning: when an algorithm is taught with unlabeled data with the intention of finding patterns or structures in the data;
- Reinforcement learning: is the process through which an algorithm picks up knowledge via positive or negative feedback, such as rewards or penalties. Examples include Deep Deterministic Policy Gradient (DDPG), Sarsa, and Q-Learning;
- Transfer learning: refers to the application of knowledge developed while completing one task to assist in the solution of another, related issue.

ML model's parameters are the variables that are changed throughout the model training phase to enable the model to generate precise predictions. An ML model has two primary categories of parameters [48]:

1. The values of a template's parameters that determine the template's behaviour and structure. During the training phase, they may either be automatically taught or manually tweaked. For instance, the number of trees in a RF or the number of layers in a NN [49];
2. The training parameters affect how the model is trained. They contain parameters such as the rate of learning, regularization and dropout rate [50].

ML has a wide array of applications. It is used for the purpose of predicting future patterns, categorizing information, and identifying irregularities. Cluster analysis is a method used to identify and analyze patterns within datasets, while recommendation systems are designed to provide personalized suggestions for users based on their individual preferences and characteristics. Optimization methods are used to enhance the performance of intricate systems, while data analytics provide significant insights of value. The field of robotics use machine learning techniques to enhance adaptability, whereas sentiment analysis is utilized to assess emotional states conveyed in textual data. ML plays diverse and vital functions in several domains [51, 52].

2.3.1 Supervised machine learning models

Supervised learning is when an algorithm is trained with labelled data, where each sample includes a label or classification. Predicting the label for new cases is the objective. Applications for supervised learning include voice recognition, stock price forecasting, and picture categorization. The training phase and the prediction phase make up the two fundamental components of a supervised learning model [53].

The algorithm is taught during the training phase using a collection of labelled data. The model searches for a connection between inputs (data attributes) and labels during training (outputs). To achieve this, model parameters may be modified using optimization methods like gradient descent.

The prediction phase: after training the model, we can use the model to make predictions on new data. This is done by providing the model input data and using the fitted parameters to produce an output prediction.

There are several supervised learning algorithm types, and each has a unique set of training and prediction methodologies. Data for supervised learning are often split into two partitions, training data and testing data. The model is trained using the training set, and its effectiveness is assessed using the test set. If the model can properly predict the labels for the test data, it is deemed successful. In general, it operates as follows [54, 55]:

1. Obtaining data is the initial stage. The model will be trained using a collection of labelled data;
2. Preprocessing: before the data can be utilized to train the model, it must be cleaned, normalized, and prepared and perform feature selection to carry out the search for the best hyperparameters;
3. Select the model: decide on the best model for the job, then customize it to fit the data set;
4. Model training: using the training dataset, the model is trained with its parameters adjusted to reduce prediction error;

5. Model validation: to make sure the model generalizes effectively to new data, it is required to assess the model's performance using a validation dataset;
6. Using the model: the model may be used to anticipate the labels for new instances once it has been trained and verified.

2.3.1.1. Feature selection

Feature selection is a data preprocessing approach that aims to choose the most useful characteristics for a ML model. The objective is to determine the features that offer the most information on the output while eliminating duplicate or unnecessary features. This may assist enhance model performance by accelerating training and reducing the likelihood of overfitting. PCA and genetic algorithms are two prevalent feature selection strategies [56, 57, 58].

However, there is another important feature selection method in gene expression data, the f-value (ANOVA F-test). While PCA is a dimensionality reduction technique that looks for new "features" that are linear combinations of the original ones to capture the maximum variance in the data, the f-value is used in a different method known as the ANOVA (Analysis of Variance) test.

The f-value is a metric that arises when performing hypothesis tests in analysis of variance (ANOVA). It is used to compare changes between the means of two or more groups. In the context of feature selection, each feature is evaluated individually to verify its relationship with the response variable [59]. Features that have a significantly large f-value indicate that there is a significant difference in the means of the response variable for the groups defined by that feature i.e. if the groups have very different means, the numerator will be large, resulting in a high F-value. If the means are similar, the F value will be close to 1. The higher the F value, the more likely it is that there is a significant difference between the group means[60, 61]. The basic formula for the F-value in ANOVA is the between-group variance divided by the within-group variance [62, 63].

The selection of a feature selection approach is dependent on the nature of the data and the underlying ML issue. In BC investigations, these strategies may be used in combination with ML algorithms to select the most relevant cancer detection features [64, 65, 66].

There are several techniques for selecting features. Built-in feature selection techniques are those that are immediately included in an ML model's training procedure. Typically, they are determined by a measure, such as the variable significance score, the regression coefficient, or the discriminatory power of a decision tree. The purpose of these approaches is to select the optimal variables for a given model such that only relevant features are retained and all others are eliminated for example: Lasso (Least Absolute

Shrinkage and Selection Operator) in LR. Filter methods are a group of feature selection techniques that compare the response of each feature to the response of the target variable. They rank the characteristics according to statistical criteria, such as the correlation between the features and the response or the significance of the features in model construction. These classifications are used to select a subset of relevant attributes for the model. The absence of attention to relationships between characteristics and the potential of picking unnecessary or duplicate features are disadvantages of these techniques. Wrapper approaches are an approach to feature selection that use ML algorithms to assess the efficacy of feature selection. These algorithms evaluate many feature combinations and select the optimal ones based on the specified ML model. They are computationally more costly than filter approaches, but in general, they are more accurate as they account for the interaction between features and how it impacts model performance [67, 68].

As previously mentioned, the last step results in the selection of the optimal model, based on the metrics tables produced and the confusion matrices.

Each metrics table contains:

- Precision: This metric indicates the proportion of positive hits (class predictions) that were actually correct. It is calculated as:

$$\frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \quad (2.1)$$

A high precision indicates that the class was predicted well relative to incorrect predictions of the same class.

- Recall (or Sensitivity): Shows the proportion of actual positives that were correctly identified. It is calculated as:

$$\frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad (2.2)$$

A high recall indicates that most of the actual class observations were correctly identified by the model.

- F1-score: It is the harmonic average of Precision and Recall and provides a balance between these two metrics. It is particularly useful if classes are unbalanced. It is calculated as:

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.3)$$

A high F1-score indicates a good balance between accuracy and recall.

- Support: Refers to the actual number of occurrences of the class in the specified dataset. It indicates how many observations belong to each category in the test set.
- Accuracy: It is the total proportion of correct predictions in relation to all predictions made. Gives an overview of the effectiveness of the model.

$$\frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{True Positives (TP)} + \text{False Positives (FP)} + \text{True Negatives (TN)} + \text{False Negatives (FN)}} \quad (2.4)$$

- Macro avg: Calculates the arithmetic mean for each metric (precision, recall, and f1-score) without considering class imbalance.
- Weighted avg: Calculates the average of each metric taking into account the class imbalance. Uses the "support" metric as a weight for each average.

2.3.1.2. Support vector machines

An example of a supervised learning method used for classification and regression is the SVM. It operates by identifying a line or hyperplane that best delineates the various dataset classifications. The separating hyperplane that optimizes the margin between the various classes is what the SVM looks for. The margin is the separation between each class's nearest points, or support vectors, and the hyperplane. The only points that have an impact on where the separating hyperplane is located are these support vectors, which are also the points that are most nearby. Binary classification, which divides data into two classes, and multi-class classification may both be accomplished using SVM (splitting data into more than two classes). Additionally, it may be used for the regression problem. Through the inclusion of a kernel function, which puts the data into a feature space of greater dimension where a linear separation may be found, SVM are also able to handle datasets that are not linearly separable. SVM are extensively utilized in a variety of industries, including bioinformatics, image recognition, and natural language processing. SVM are a sort of supervised learning algorithm that may be used to categorize problems, including the detection of BC [69].

Data from imaging tests, such as mammograms and ultrasounds, as well as information regarding tumour features, such as size, shape, and texture, are often utilized to train an SVM model for the diagnosis of BC. The algorithm is trained using this data to spot patterns that point to BC [70].

The SVM model may be used to categorize new instances of BC as malignant or benign after being trained. As a result, it may be used as a support tool to assist physicians in making treatment choices or in identifying individuals who are more likely to acquire BC. SVM has even been employed by researchers

to categorize BC based on gene expression data. To train the algorithm, they chose a number of genes that have previously been linked to BC [71, 72].

2.3.1.3. k-Nearest Neighbors

A supervised learning classification technique called KNN is based on the notion that a data point is categorised based on most of its nearest neighbours.

The two primary steps for KNN to function are [73]:

- Training: In this phase, a training dataset with each data point labelled with a distinct class is given to the algorithm to load into memory. No further training is done;
- Classification: The method locates the k nearest data points (the neighbours) in the training set when a new data point is supplied.

A smoother model that is less likely to be overfitted is produced by larger values of k , while a more complicated model that is more likely to be overfitted is produced by lower values of k .

The majority of the classes of the k closest neighbours are used to give the class to the new data point. The new point will be categorized as class A if $k=3$ and two of its neighbours are class A and one neighbour is class B. Then, a test dataset is used to evaluate the KNN model, and accuracy is calculated. The algorithm may be modified to use alternative distance metrics to determine how close data points are to one another or to use weights to give the closest neighbours more weight. KNN is a straightforward algorithm that is simple to comprehend and use, but it has some drawbacks, such as the fact that it needs the storage of the full training dataset and the search for neighbours may be laborious for big data sets [74, 73].

Several studies have employed the KNN to categorize BC, and in some instances, the findings have been encouraging. For instance, in 2017, researchers utilized KNN to categorize BC based on genetic traits. The research revealed that the KNN algorithm has a 97.5% classification accuracy rate. The amount of k and the characteristics used, however, might have an impact on the algorithm's performance, and it may not always perform as well as more sophisticated algorithms. It is also critical to take into account that the algorithm might be affected by outliers and data noise, and huge datasets can increase its computing cost [75, 76, 77].

2.3.1.4. Linear and logistic regression

For the LR a continuous target variable may be predicted using the supervised ML approach of LR using one or more input variables. Finding the best straight line between the data points is the core tenet of LR. The data and an optimization method are used to estimate the coefficients. Finding the coefficient values that reduce the discrepancy between the target variable's anticipated values and actual values is the goal [78].

The model may be used to generate predictions on fresh data when the coefficients are computed. It is a common practice in several industries, including engineering, finance, and economics, to employ this simple yet effective strategy. Although it is an excellent place to start for regression tasks, LR makes the assumption that independent and dependent variables are always linear, which may not be true for all data sets [79].

Based on certain input characteristics, ML may utilize LR to forecast the probability of developing BC. Risk factors including age, family history, and genetics might be included in these input variables, sometimes referred to as predictors. Predicting BC recurrence is one use of LR in BC research. The association between certain genetic markers and the chance of BC recurrence was modelled by the scientists using LR. To model the link between the predictors and the goal, LR is utilized (recurrence or BC risk). Based on your input variables, the model may be used to forecast the risk of BC in new patients after it has been trained [80, 81].

For the Logistic Regression (LOR) the likelihood of a particular class may be predicted using the supervised ML technique of LOR using a collection of features. It is based on the logistic equation model, which represents the association between the dependent variable (Y) and one or more independent variables (X). Finding the optimal values for the coefficients ($b_0, b_1, b_2, \dots$) that maximize the likelihood of correctly predicting the class is the goal. The logistic function also referred to as the sigmoid function, is used by the LOR model to simulate the likelihood of a binary outcome [82, 83]. Numerous applications of LOR include predicting patient recovery in medicine, predicting consumer purchases in marketing, predicting loan default in finance, predicting recidivism in criminology, and many more.

It is extremely helpful for issues requiring binary classification, such as the diagnosis of BC, where the class may be either "cancer" or "non-cancer". It is feasible to forecast a patient's risk of developing cancer using a number of factors and to apply a predetermined cutoff rate to decide whether the patient is diagnosed with cancer or not. Additionally, the possibility that a tumour is benign or malignant may be predicted using algorithms, based on the tumour's size, shape, density, and levels of gene expression [84, 85, 86].

The fact that LOR is a linear model, which only accounts for linear correlations between input and output variables, should also be noted. Other models, such as DT or NN, may be better suitable for interactions that are more complicated.

2.3.1.5. Decision and Regression Trees

Using regression and decision trees for both classification and regression issues, supervised ML algorithms in the form of trees are used. They operate by building a tree-like representation of choices and their outcomes [87]. In a DT, the branches stand in for the choices made in response to the input characteristics, while the leaves represent the anticipated class labels. The procedure begins at the tree's base and recursively divides the data into classes based on a characteristic that optimizes class separation. The final predictions are performed by moving up the tree from the root to a leaf node [88].

The branches of a regression tree indicate the judgments based on the input characteristics, while the leaves represent the projected continuous value. Similar to a DT, the technique operates by attempting to reduce the variance of the target variable rather than maximising class separation [89, 90]. Both DT and Regression Tree (RT) can handle categorical and numerical data, and they are both simple to comprehend and analyze. But since DT are prone to overfitting, procedures like pruning or ensemble methods are often used to solve this issue. These algorithms are effective for processing huge and complicated datasets since they can handle both numerical and categorical data.

In relation to BC, these algorithms have already been used to diagnose BC using mammographic images and also using a patient database [91, 92, 93].

2.3.1.6. Artificial Neural Networks

The structure of the human brain inspires ANN. They consist of layers of linked neurons, each of which performs basic computations and transmits the results to other neurons [94].

Adjusting the weights of connections between neurons based on training examples is the learning process of a neural network. Using optimization methods such as gradient descent, weights are modified to reduce the discrepancy between the expected output and actual data. ANN learns to recognize patterns in the training data and changes the weights to provide correct outputs during training. After training, the ANN may make predictions on fresh data based on the patterns it has learnt. The network receives input data, which the neurons process to generate an output. This procedure is performed several times to fine-tune the weights of the network [95, 96].

ANNs are used in a variety of applications, such as classification, pattern identification, natural lan-

guage processing, and prediction. ANNs may be used to categorize tumours as benign or malignant based on criteria like size, shape, and image intensity in the setting of BC. In addition, they may be used to forecast disease progression and recurrence risk. In the context of BC, ANNs are commonly used for the categorization and detection of lesions, as well as the identification of genetic markers that are crucial for cancer prediction and therapy. ANNs are composed of layers of linked neurons that are trained using vast amounts of data to fulfil certain tasks. In BC gene expression datasets, ANN may be trained to predict the presence or absence of cancer and to identify the disease's most significant genes. In addition, ANN may be utilized to combine clinical and genetic data for more precise prognosis and therapy of BC [97, 98].

However, it is essential to note that ANN are tools whose efficacy relies on the quality and amount of data available for training. In addition, since ANN are very complicated models, it is essential to review your findings thoroughly to verify that they are accurately understood and not misapplied.

2.3.1.7. Ensembles

A ML technique called ensemble learning mixes a number of distinct models to increase the reliability and accuracy of the output. To produce a forecast, many models are trained and combined rather than simply one. There are three primary categories: Random Forests (RF), Boosting and Bagging [99].

By building several independent training models from random samples and replacing the training data, the Bagging (Bootstrapped Aggregation) ensemble approach seeks to increase the stability and accuracy of ML models. By combining the predictions from the several models, it is expected that the final model would be less prone to overfitting and more resistant to data noise. Since each model is trained using a distinct sample of data, it may identify various patterns in the data. The final model may be more accurate than a model trained on all the data by incorporating these patterns. Typically, the various model forecasts are polled or averaged to get the final projection. Bagging is often employed with DT and other overfitting-prone ML models [100].

Boosting is a sort of ensemble ML technique that makes use of weighted combinations of base models to enhance the performance of a poor base model. The approach first fits a base model to the training data, after which it fits additional models to the residual errors that the base model had previously detected. A weighted combination of all base and extra models are created by weighing the additional models according to how well they are able to forecast base model mistakes. Boosting is used to fix base model flaws and increase its capability to generalize to new inputs [100].

RF is an ensemble learning strategy that mixes numerous decision trees to build a final model. It works by separately building multiple DT, and then integrating the findings to get a final forecast. Multiple

voting is used in this situation when the predicted class is shared by the majority of trees. The benefit of employing RF is that it is less prone to overfitting and can cope with irrelevant or correlated variables without hurting model accuracy [101].

Ensemble methods, which include classification, regression, and anomaly detection, are ML approaches that aggregate many distinct models to increase the accuracy and stability of findings. Ensemble techniques may be used to incorporate several elements, such as mammographic pictures, clinical data, and gene expression profiles, to increase diagnostic accuracy in the detection of BC [102, 103, 104].

2.3.2 Unsupervised Machine Learning models

Unsupervised learning is a type of ML in which the algorithm is trained on a dataset that has no labels or classifications. Without previous instruction, the data does not have a preset output, hence the goal is to detect patterns or structures in the data. Data compression, anomaly detection, and customer group identification are a few examples of uses [105].

There are several kinds of unsupervised learning algorithms, and each has a unique set of strategies and procedures for identifying patterns in data. Algorithms for unsupervised learning include, among others, dimensionality reduction, and clustering [106].

Since there are no labels to predict, an unsupervised learning model just has one essential component, the training phase. During the training phase, an unlabeled dataset is used to train the algorithm. The model searches for structures or patterns in the data during training. Techniques like DL, dimensionality reduction, and clustering may be used to accomplish this. Whether labels or training responses are present or absent is the primary distinction between supervised and unsupervised learning. When the model has been trained, it may be utilized to explain the input data. This may be achieved by categorizing the incoming data into several groups or by finding important data properties [107, 106].

In brief, when we have labelled data and know the anticipated outcome, we use supervised learning. However, when we do not have labelled data and do not know the expected outcome, we use unsupervised learning.

Typically, the stages are as follows [108]:

1. Collecting data: Collecting an unlabeled data set is the initial phase in the process. This data set will be utilized to train the model;
2. Preprocessing: Prior to using the data to train the model, it is important to clean, normalize, and prepare it;

3. Model selection: Select the best model and set it up for the dataset and particular job at hand;
4. Model training: Using the dataset as a training set, the model is trained to seek structures or patterns in the data;
5. Description of Data: After training, the model may be used to characterize the input data, clustering it into subsets or highlighting distinctive features;
6. How to utilize the model: The model may be used for tasks like grouping data or dimensionality reduction, as well as identifying patterns or structures in the data.

K-means

K-means is an algorithm for grouping comparable data into "k" distinct groups. It works by creating "k" random starting centroids and then allocating each data point to the centroid that is closest to it. Then, the centroids are re-evaluated based on the average of the given data points, and this procedure is continued until the centroids no longer change or until a stopping condition is satisfied. The outcome is "k" separate clusters of comparable data [109].

K-means is extensively used in data analysis, particularly in data mining and ML applications. It is a quick and effective method that may be used to identify patterns and trends in massive data sets. In addition, it is often used to execute data collations in several areas, including healthcare, marketing, and finance. K-means is one of ML's most well-known and widely-used clustering techniques. This technique seeks to identify clusters or groupings of related data within a dataset. K-means can be used to group samples of tumours with comparable genetic traits in the setting of BC. This may be helpful for identifying subgroups of cancers with unique features, which may be significant for the creation of individualized treatment regimens. It is essential to note, however, that K-means is only a data exploration tool, and that the generated findings must be confirmed by additional methods in order to get a more trustworthy conclusion [110, 111].

T-distributed stochastic neighbour embedding

T-distributed stochastic neighbour embedding (t-SNE) is a non-linear two-dimensional visualization technique used in ML. It is often used to study and analyze complicated data in a manner that is readily comprehended by people. It is particularly effective for displaying high-dimensional data, such as genetic data, in order to find patterns and correlations among samples. t-SNE relies on retaining closeness

between strongly correlated points in high-dimensional space while projecting these points onto a two-dimensional representation. This enables t-SNE to show the underlying data structure in a form that is obvious and simple to comprehend [112].

t-SNE is commonly used in ML applications, including BC research. By using t-SNE to the genetic data of BC patients, researchers may find patterns and links between various kinds of tumour cells and their genetic traits, therefore facilitating the identification of novel treatment options and the customization of therapy. However, it is essential to emphasize that t-SNE visualization is a supplementary tool for data analysis and should not be utilized as the only diagnostic method [113, 114].

Principal Component Analysis

PCA is a method for reducing the dimensionality of a given dataset. This implies that instead of dealing with each data feature or column, PCA finds and extracts the principal components that represent the majority of the data's variability. These primary components are derived from the data's correlation or covariance matrix. PCA has the benefit of reducing the dimensionality of data without sacrificing too much information. This is beneficial for a variety of applications, including data visualization, feature selection, enhancing the performance of ML models, and spotting outliers [115, 116].

PCA is often used with other ML methods to enhance performance. PCA may be used, for instance, to decrease the dimensionality of the data prior to performing a classification or clustering technique. This may be used to examine vast volumes of gene expression data pertaining to BC. These fundamental components may be used for data exploration and visualization, as well as feature selection and categorization. PCA is a crucial tool for identifying complicated patterns in gene expression data and identifying possible BC biomarkers [117, 118].

Hierarchical clustering

Hierarchical Clustering is an ML data clustering approach that aims to identify hierarchical patterns within data. The concept is that, as the process progresses, the data is sorted into more dense clusters until all of the points are clustered into a single cluster or a limited number of different clusters. Agglomerative and divisive are the two primary techniques of Hierarchical Clustering. The Agglomerative method begins with each point being a distinct cluster and, over time, the closest clusters are combined into one. The Divisive method, on the other hand, begins with all points in a single cluster, which is then split into smaller clusters over time [119, 120].

Hierarchical Clustering has been used to identify subgroups of BC patients with comparable features

based on genomic data, including gene expression. This may be useful for gaining a better understanding of the distinctions between cancers and developing individualized treatment plans. To assure its validity, Hierarchical Clustering, like other clustering algorithms, needs careful parameter modifications and interpretation of findings [121, 122].

2.3.3 Deep Learning

DL is a subfield of ML that deals with models like DNN. DL employs a sophisticated system of algorithms that makes it possible to handle unstructured data, including documents, photos, and text.

DNN consist of numerous hidden layers of nodes, allowing models to learn complex, nonlinear input-output interactions. The objective is to discover hierarchical representations of the data, which will enable the models to execute sophisticated ML tasks such as classification, object identification, and natural language processing. These models are trained on vast volumes of data and can effectively identify and forecast from that data without explicit instructions. DL has been utilized extensively in areas such as computer vision, natural language processing, voice recognition, and others, and is increasingly being employed to handle complicated and challenging issues in domains such as medicine. As with ML, research into BC using DL is continually changing and new applications are being created. These procedures must be evaluated and integrated with other methods to guarantee accuracy and safety.

Deep Neural Networks

DNN are a family of deep ML models based on artificial NN with a large number of hidden layers. They are intended to learn hierarchical representations of complicated data and for the model to automatically learn from them [123].

DNNs are constructed of many layers of neurons, with each layer including multiple interconnected neurons. The input data goes through each layer's weights and biases, and the layer's output is transferred to the subsequent layer. During training, weights and biases are modified to reduce model error, enabling the model to acquire knowledge from the data [124, 125].

There are several applications for DNNs, including classification, image recognition, natural language processing, and sentiment analysis. They are particularly beneficial for issues involving complex data, such as pictures and signals, since they enable the model to catch crucial patterns and characteristics that are difficult for shallower models to grasp [126].

DNNs have been used in the context of BC for the categorization of tumor types, identification of tumors in mammography pictures, and survival prediction, among others. They are capable of learning intricate

gene expression and imaging patterns, making them a great tool for interpreting BC data [127, 128, 129].

Convolutional Neural Networks

Convolutional Neural Network (CNN)s are an evolution from ANN class used to solve image processing issues such as computer vision and pattern recognition. The processing layers of a CNN are comprised of convolution, pooling, and fully linked layers. Convolution layers use filters to identify visual patterns and extract pertinent features. Pooling layers minimize the dimensionality of data while maintaining vital information. Finally, layers with complete connectivity are employed for classification and prediction. CNNs are frequently utilized in computer vision applications such as face identification, license plate recognition, and medical picture categorization because of their excellent accuracy in object recognition tasks [130]. CNNs are a particular design of ANN that are often used for computer vision problems. Utilizing seizure operations (filters) to collect spatial patterns in picture data is the core concept behind CNN. This enables the model to learn important characteristics automatically from the data [131]. Filters are applied to small picture subregions known as "receptors". The outcome of these seizure processes is applied to an activation function that determines the significance of a particular characteristic. The outputs of these processes are then aggregated to produce a new set of features, which are then processed with additional filters to provide more intricate outputs. This procedure is repeated several times, resulting in convolution and maximum pooling layers (a technique for dimensionality reduction). At the conclusion of this procedure, the learnt characteristics are fed into a fully connected layer that is responsible for classification or prediction. The ultimate result of image classification jobs is a probability distribution across the available classes [132]. CNNs are specially used in image-related tasks like image classification, object identification, and image segmentation and have been used effectively for the identification of mammographic lesions and the categorization of forms of BC in relation to BC. They have also been used to evaluate pictures from tissue biopsies to aid in cancer detection and disease severity evaluation [133, 134].

Recurrent Neural Networks

Recurrent Neural Networks (RNN) are a type of DL models intended to process time series data. RNNs feature an internal memory structure that enables the analysis of previous data, which is important for forecasting future occurrences and categorizing time series [135].

An RNN comprises a layer of recurrent neurons that keep an internal memory and an output layer that generates a prediction. Each temporal step in the input sequence is individually processed by the recurrent neurons, and the output information from the preceding step is utilized as input for the subsequent step.

Several RNN variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), are intended to handle disappearing and bursting gradients during training [136, 137]. In healthcare applications, RNNs have been used effectively to evaluate time series of medical data, such as electrocardiogram signals or patient monitoring data, in order to anticipate illness or assist with diagnosis. In addition, RNNs have been used with other models to increase the precision of BC diagnosis using mammography images [138, 139].

2.3.4 Machine learning applied to studying Breast Cancer

These are a few applications of ML in the diagnosis of BC. New applications are being created as a result of ongoing research in this field. Therefore, using ML, gene expression data may be used to predict a patient's likelihood of having BC by examining the activity levels of various genes, in blood or tissue samples. The degree of gene activity referred to as gene expression, may be used to define a patient as having BC or not and can provide details about the biology of the tumour's underlying structure [140].

By examining the activity levels of various genes in tumour samples, gene expression data may be utilized to forecast the stage of BC. The degree of a gene's activity referred to as its expression level may provide details about the biology of the tumour and be used to categorize it into distinct phases. The complexity of the illness, the amount and quality of the dataset, and the particular algorithm utilized may all have an impact on the model's performance. It is crucial to remember that identifying BC and forecasting the stage remain challenging tasks. Further study is required to properly understand the underlying causes of BC and how this might be predicted using gene expression data since this is a complicated endeavour [141].

In several areas of BC detection and therapy, ML is applied. A few scenarios include [142, 143, 144, 145, 146, 147]:

- Diagnosis: using ML algorithms, radiologists can diagnose worrisome lesions by classifying breast pictures as benign or cancerous;
- Prognosis prediction: Using both genetic and clinical data, one can forecast the likelihood of a BC recurrence and assist choose the best course of action;
- Identification of biomarkers: (proteins or genes) linked to illness using ML algorithms may aid in the creation of novel therapeutics;
- Drug identification: is the process of identifying the best medications for each patient based on their genetic inheritance and medical history using ML algorithms;

- Identification of high-risk patients: using ML algorithms, it is possible to identify patients who are at a high risk of getting BC based on risk variables including age, family history, and lifestyle choices.

Chapter 3

Materials and Methods

For implementation, visual studio code, the Python language in version 3.8 and the scikit-learn library was used. It is a ML library for the Python programming language. All code developed is properly hosted and available for consultation in the GitHub repository through the link:

<https://github.com/ruteeecaastrooo/DevelopmentofanAIBCdiagnostictool>.

Due to the demand for high computational capacity inherent in training the models and the subsequent research and optimization of hyperparameters, a server provided by the IT department of the University of Minho was used. This server, endowed with advanced resources, for example in terms of RAM and CPU, allowed a significant reduction in the time required to train the models and conduct research and parameter improvement.

The project's architecture was conceived in a modular way. Each pipeline is designed to operate sequentially or, alternatively, be used autonomously. This modularity gives the freedom to load specific datasets, whether of genetic expression or clinical data, the latter being able to incorporate information about Copy Number Alteration (CNA). Through these pipelines, it is possible to preprocess the data, apply feature selection methods, and segment the dataset into training and test subsets. Once processed, the dataset is ready to feed various ML algorithms, spanning both traditional models and deep learning frameworks. Additionally, a specific module is dedicated to the optimization of hyperparameters, aiming to maximize the performance of the models. Finally, the last pipeline culminates in the application of the most promising models, producing metrics tables, confusion matrices and, consequently, allowing the selection of the optimal model.

3.1 Methodology for Cancer Diagnosis

3.1.1 Preprocessing

The first step for the creation of the pipeline for data processing began with the conversion of gene identifiers, specifically transitioning from the Illumina format to the Entrez format. The function employed for this task was designated as "illuminaToEntrez", utilizing an online API to perform the conversion. Following this, the gene expression data for healthy subjects was imported from a file named "normals_ExpressionMatrix.txt". The dataset needed a reorientation to align samples in rows and genes in columns. Subsequent to this transposition, the gene names were adapted to the Entrez format, employing the aforementioned function. Some gene identifiers were not converted successfully so the columns for these genes were eliminated.

Afterwards, gene expression data for individuals diagnosed with cancer was imported. A similar transposition of this dataset was undertaken to ensure uniformity. The columns, which represent genes, were synchronized between the datasets originating from healthy individuals and those from cancer patients.

Before joining these datasets, the gene expression values specific to healthy individuals were subjected to normalization, specifically through the z-normalization method. This particular approach was selected owing to the fact that the dataset of cancer patients had already undergone z-score normalization.

To conclude, both datasets were merged into a single entity. An auxiliary column, termed "hasCancer", was appended to facilitate future analyses. This column was assigned a binary value: 0 for healthy individuals and 1 for those diagnosed with cancer. Upon this, any columns laden with missing values were eliminated, thereby completing the process. The dataset, in its final form, was stored in a file named "final.csv".

3.1.2 Feature selection

Feature selection was carried out using F-value, focusing on the 1500 most relevant features. This choice was supported by the "feature_importance" pipeline analysis, which demonstrated that these 1500 features were responsible for explaining approximately 99.99% of the variation in the data. The 1500 features were saved in the file "sorted_f_value_hascancer.txt". The choice was made to use only the F-value method for feature selection, specifically for the task of predicting the presence or absence of cancer. This decision was motivated by a combination of factors explained below. When other methods were explored to predict cancer subtype, it was found that the F-value method produced the best results and the nature of the

data for both tasks of predicting the presence of cancer and its subtypes was the same, justifying using the same feature selection method. Furthermore, although PCA is a robust method for dimensionality reduction, it merges original features into principal components. This makes it impossible to extract the 1500 most important features, so only what was extracted from the f-value was what curiously gave the best results. This information was very important to then cross-reference with the PAM50 gene list (it is a clinically validated panel of 50 genes commonly used to classify breast cancer into molecular subtypes). The F-value method, on the other hand, allows for the preservation of original features, facilitating a more interpretable model and providing insights into specific features.

Initially, the "final.csv" dataset was loaded into this pipeline using the pandas library. Subsequently, the dataset was balanced. This was done by filtering out instances where "hasCancer" was equal to 1 and randomly selecting 144 of those instances. This was the balancing approach decided on (Random Undersampling). The result was then combined with instances where "hasCancer" was different from 1. This step was critical to ensure that both classes were more equally represented, thus mitigating the risk of class bias.

After preparing the dataset, the features and labels were separated into different variables, x and y, respectively. The dataset was then split into training and testing subsets using the "train_test_split" function from the scikit-learn library. An 80-20 ratio was maintained between training and testing data. After that, the t-distributed Stochastic Neighbor Embedding (TSNE) visualization technique was applied to visualize the distribution of data in a two-dimensional space (Figure 1). This technique is useful for understanding the structure and relationships between data.

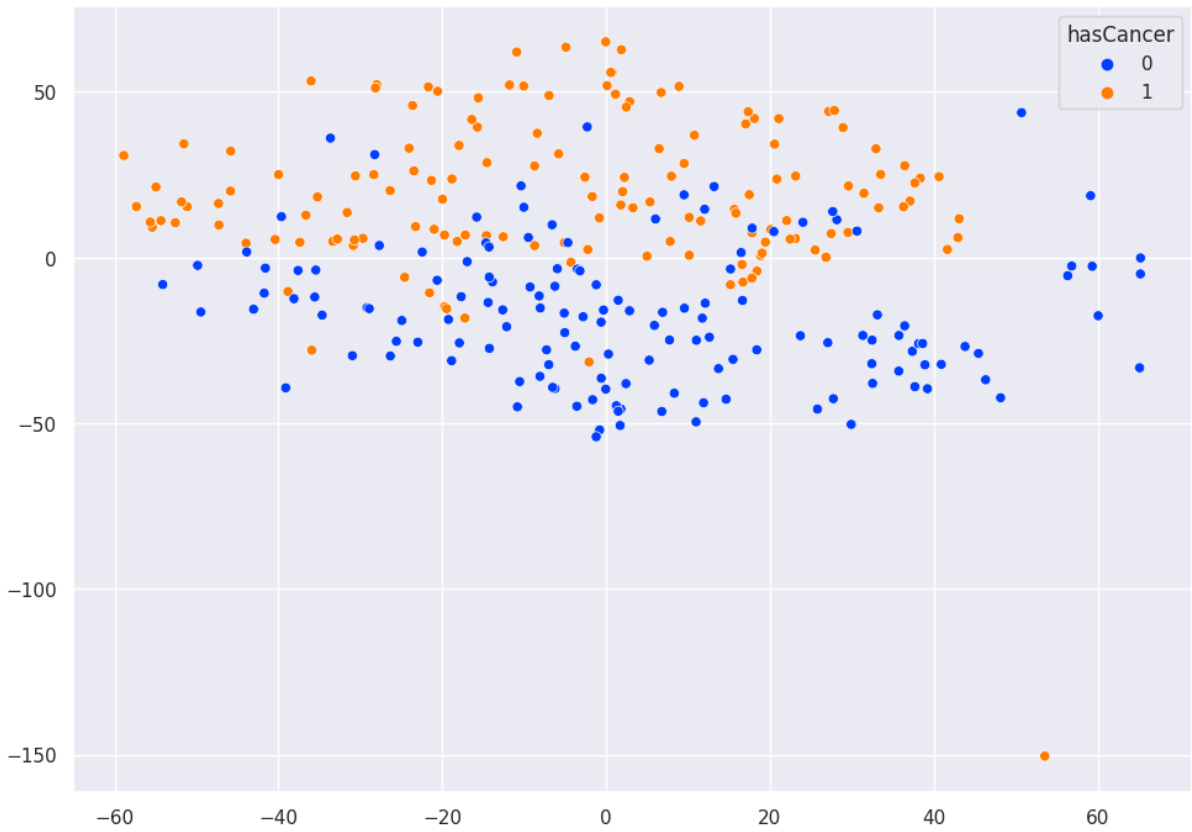


Figure 1: T-distributed Stochastic Neighbor Embedding to visualize the distribution of data in a two-dimensional space of class 0 (healthy individuals) and class 1 (individuals with cancer)

Then, the feature selection method was applied to the training data. The method chosen was the ANOVA F test, implemented by the "SelectKBest" class with "f_classif". This method evaluated the importance of each characteristic in relation to the objective variable and selected the best 1500 to keep.

The training and testing datasets were then transformed to include only these selected features. These transformed versions were saved to files using the pickle format for future use. Finally, the F values and corresponding feature names were retrieved and sorted in descending order of importance. This information was saved in a text file for later analysis.

3.1.3 Models optimization

The objective of this step was to identify the best hyperparameters for each model, thus optimizing its performance. This approach offers a systematic and efficient way to perform hyperparameter search, saving time and computational resources.

In the presented implementation, a class called "CustomRandomizedSearchCV" was developed, which

acts as an extension of the RandomizedSearchCV functionality provided by the Scikit-learn library. The purpose of this class is to optimize the hyperparameter search process for various ML algorithms. The class was designed to allow simultaneous and comparative evaluation of different models and their hyperparameter combinations.

The "CustomRandomizedSearchCV" class was initialized with several parameters, including the training and validation data, the ML algorithm to be optimized, the distribution of hyperparameters, the number of iterations allowed, and others. An important point is that this class also allows the use of multiple CPU cores to speed up the training process, defined by the "n_jobs" parameter.

Additional methods were implemented to facilitate the process. The "calc_total_parameters" method calculates the total number of possible hyperparameter combinations based on the given ranges. The "random_search_cv" method performs the random search, fitting the model to the data and finding the best hyperparameters. Other methods provide metrics, such as the best set of hyperparameters ("get_best_parameters"), the best score achieved ("get_best_score"), and the average training and testing times ("get_average_train_time" and "get_average_test_time"). However, choosing the various parameter_distribution for each model and then searching for the best hyperparameters for each model with the "CustomRandomizedSearchCV" class was done manually, by reading the documentation. (Table: 4)

For each ML algorithm including SVM, KNN, LR, DT, RF, AdaBoost, ANN and DNN, an instance of the "CustomRandomizedSearchCV" class was created. These instances were then used to find the best hyperparameters for each algorithm, using the preprocessed training and validation data.

Table 4: Hyperparameters used in CustomRandomizedSearchCV to search for the optimal parameters

Model	Hyperparameters
Support Vector Machines (SVM)	Kernel: [linear, poly, rbf] C: [log10(0.1), log10(1000), 1000] Gamma: [log10(0.001), log10(10), 1000] Max_iter: [100, 200, 500, 1000] Num_iters allowed: [1000]
K-Nearest Neighbors (KNN)	N_neighbors: [1, 30, 30] Algorithm: [ball_tree, kd_tree, brute] Weights: [uniform, distance] Metric: [euclidean, manhattan, minkowski] Num_iters allowed: [1000]
Linear Regression (LR)	Penalty: [l1, l2, elasticnet, none] Solver: [newton-cg, lbfgs, liblinear, sag, saga] Max_iter: [100, 200, 500, 1000, 2000, 3500, 5000] Multi_class: [auto, ovr, multinomial] C: [0.5, 2, 30] Num_iters allowed: [1000]
Decision Tree (DT)	Criterion: [gini, entropy] Splitter: [best, random] Max_depth: [None, 2, 4, 8, 16] Min_samples_split: [2, 4, 8, 16, 32] Max_features: [auto, sqrt, log2] Num_iters allowed: [1000]
Random Forests (RF)	Criterion: [gini, entropy] Max_depth: [None, 2, 4, 8, 16] Min_samples_split: [2, 4, 8, 16, 32] Max_features: [auto, sqrt, log2] Num_iters allowed: [1000]

AdaBoost	Algorithm: [SAMME, SAMME.R] N_estimators: [10, 50, 100, 200] Learning_rate: [log10(0.01), log10(1), 100] Num_iters allowed: [1000]
Artificial Neural Networks (ANN)	Hidden_layer_sizes: [128, 256, 512, 1024, 2048] Solver: [lbfgs, sg, adam] Learning_rate: [constant, invscaling, adaptive] Learning_rate_init: [log10(0.001), log10(1), 1000] Max_iter: [200, 500, 2000] Batch_size: [auto, 64, 512] Num_iters allowed: [1000]
Deep Neural Network (DNN)	Hidden_layer_sizes: [(512, 256, 128, 128, 128, 128, 128, 128), (256, 128, 128, 128, 128, 128, 128, 128), (512, 256, 128, 128, 128, 128, 128), (512, 256, 128, 128, 128, 128), (256, 128, 128, 128, 128), (512, 128, 128, 128, 128), (512, 128, 128, 128), (256, 128, 128, 128), (256, 128, 128), (512, 128), (256, 128)] Solver: [lbfgs, sg, adam] Learning_rate: [constant, invscaling, adaptive] Learning_rate_init: [log10(0.001), log10(1), 1000] Max_iter: [200, 500, 2000] Batch_size: [auto, 64, 512] Num_iters allowed: [1000]

3.1.4 Models evaluation

In this final pipeline the objective was to train multiple ML models, make predictions and evaluate the performance of these models through metrics and confusion matrices. Models with scores below 90% in the previous pipeline were excluded from testing here. First, previously processed datasets were loaded into the previous pipeline in pickle format. This data included training and validation sets ("X_train_validation", "y_train_validation") and test sets ("X_test", "y_test").

The best hyperparameters were added to the first model to be trained and predictions were made on the test set. Classification metrics were printed using the "classification_report" function from the scikit-learn library. Additionally, a confusion matrix was generated and visualized. For each of these models, performance metrics were calculated and confusion matrices were displayed to evaluate the effectiveness of the models in correctly classifying the test data to conclude which model was best.

3.2 Methodology for Cancer Subtype

3.2.1 Preprocessing

The first step was to load the patients' clinical dataset using the Python Pandas library. The Scipy library was also imported to enable the use of statistical normalization. Three options were made for normalization: z-score, min-max and None. Next, the CNA and gene expression datasets were also loaded and their respective gene identifier columns ("Entrez_Gene_Id") were excluded to standardize the dataset.

To deal with any missing values in the datasets, a function was defined to remove them and also print the number of rows removed. This function was applied to all three datasets: clinical, CNA, and gene expression of individuals with cancer. After that, the normalization option "none" was applied to the CNA and gene expression datasets, because the gene expression one already came with z-score normalization and as for the CNA and clinical data did not make sense to normalize given the nature of the data.

Then, the gene expression and CNA expression tables were transposed. The genes, which were initially in the rows, were moved to the columns and the patients, which were in the columns, were moved to the rows. Column names were then modified to clearly distinguish between gene expression and CNA data.

Later, these transposed datasets were merged based on the indices, which were the patient IDs. The same was done to combine the clinical dataset with this merged mRNA and CNA dataset. Finally, the dimensions of all dataframes were printed for verification and the final dataset was saved to a csv file for use in subsequent pipelines. The option was also made to save the preprocessed dataset in pkl format.

3.2.2 Feature selection

The purpose of this pipeline was to identify and reduce the data to the most relevant characteristics of the dataset. First, several Python libraries, including pandas, scikit-learn, and other visualization and analysis libraries, were imported. Next, the previously processed dataset was loaded and checked for missing values. Then the clinical characteristics were removed, except the one that was of interest for the prediction, defined at the beginning as “3-Gene classifier subtype”. The “3-Gene Classifier Subtype” column represents breast cancer subtypes based on expression of: ER, PR, and HER2.

The dataset was then divided into input variables (x) and target variable (y). The dataset was also divided into training and test sets (80/20) using the `train_test_split` function from the scikit-learn library. After that, the t-distributed Stochastic Neighbor Embedding visualization technique was applied to visualize the distribution of data in a two-dimensional space (Figure 2). This technique is useful for understanding the structure and relationships between data.

Then, the feature selection method was applied to the training data. The method chosen was the ANOVA F test, implemented by the “SelectKBest” class with “f_classif”. This method evaluated the importance of each characteristic in relation to the objective variable and selected the best 1400 to keep. This k was verified through the random forest feature importance that 1400 genes or components explained 99.88% of all information, which became the k value used (see the `feature_importance.ipynb` pipeline). The training and testing datasets were then transformed to include only these selected features. These transformed versions were saved to files using the pickle format for future use. Then, the f-values and corresponding feature names were saved and sorted in descending order of importance. This information was saved in a text file for later analysis. The 1400 features was save in the file “sorted_f_value.txt”. Subsequently, another feature selection method PCA was tested using the same procedure and the transformed data was also saved in picke format. Finally, the pipeline was executed for the third time, as was the case with f-value and PCA, but here all features were used, the number of features was not reduced as in the previous two times.

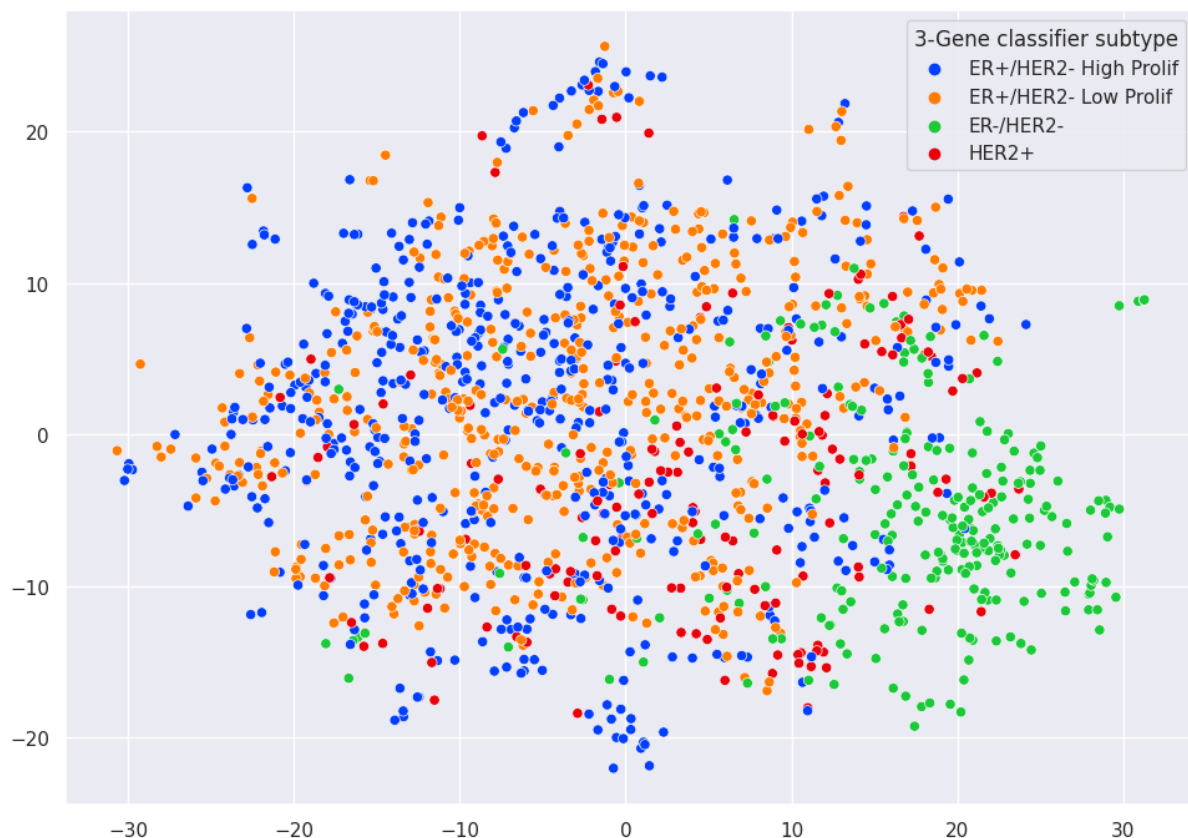


Figure 2: T-distributed stochastic neighbour embedding to visualize the distribution of data in a two-dimensional space of the four classes of “3-Gene Classifier Subtype”

3.2.3 Models optimization

The methodology adopted here was exactly the same methodology as the models optimization for Cancer Diagnosis. The only variant here was that the pipeline here was run three times as it ran with data from each feature selection method adopted. The objective was precisely to see which method in general presented the best f1-score results, so that in the future, when the nature of the data was the same and with a view to optimizing time, this method would be used exclusively.

3.2.4 Models evaluation

The methodology adopted here was exactly the same methodology as that used in the models evaluation for Cancer Diagnosis. After all the process, a small “crossgeneslistas” pipeline was created to import the txt files with the list of genes after the feature selection and their respective values to cross with the list of pam50 genes “pam50.txt”. With this, an ordered list of genes (from highest to lowest value) in common

in both lists was obtained. This was done twice, once with the feature selection list for predicting whether you have cancer and once with the feature selection list for predicting the cancer subtype.

Chapter 4

Results and Discussion

4.1 Determining the best feature selection method

The first results were obtained when testing the three feature selection methods. The scores obtained by each model using the different feature selection methods are illustrated in Figure 3 and 4. In general, the F-value method consistently proves to be the most effective for the majority of models presented, while the PCA and Total methods vary in effectiveness depending on the specific model in question. This consistency suggests that the F-value method is a robust and reliable technique for feature selection in these datasets and contexts.

For the SVM model, the best results are obtained when the F-value method is used, reaching a score of 0.934, while the PCA and Total methods reach 0.904 and 0.903, respectively. In the case of KNN, the F-value feature selection technique once again demonstrated superiority with a score of 0.870, followed by PCA with 0.639 and the Total method with 0.642. The LR model exhibits remarkable performance with the F-value method, scoring 0.923, while PCA and Total score 0.885 and 0.912, respectively. The last model in this figure, DT, achieves 0.819 with the F-value method, 0.486 PCA, followed by the , 0.744, and the Total.

Moving on to Figure 4, four other models are analyzed: RF, adaboost, ANN and DNN. In the case of the RF model, the best score is achieved with the F-value, 0.919, followed by the PCA, 0.646, and the Total method, 0.918. The adaboost model, in turn, shows solid performance with the F-value method, 0.902, while the PCA and Total methods achieve scores of 0.766 and 0.905, respectively. For the ANN model, the F-value feature selection technique stands out with a score of 0.933, followed by PCA, 0.882, and Total with 0. Lastly, the DNN model shows an impressive result with the F-value method, achieving a score of 0.936, with PCA and Total scoring 0.864 and 0, respectively. 0 means that there was no computational capacity to run the model.

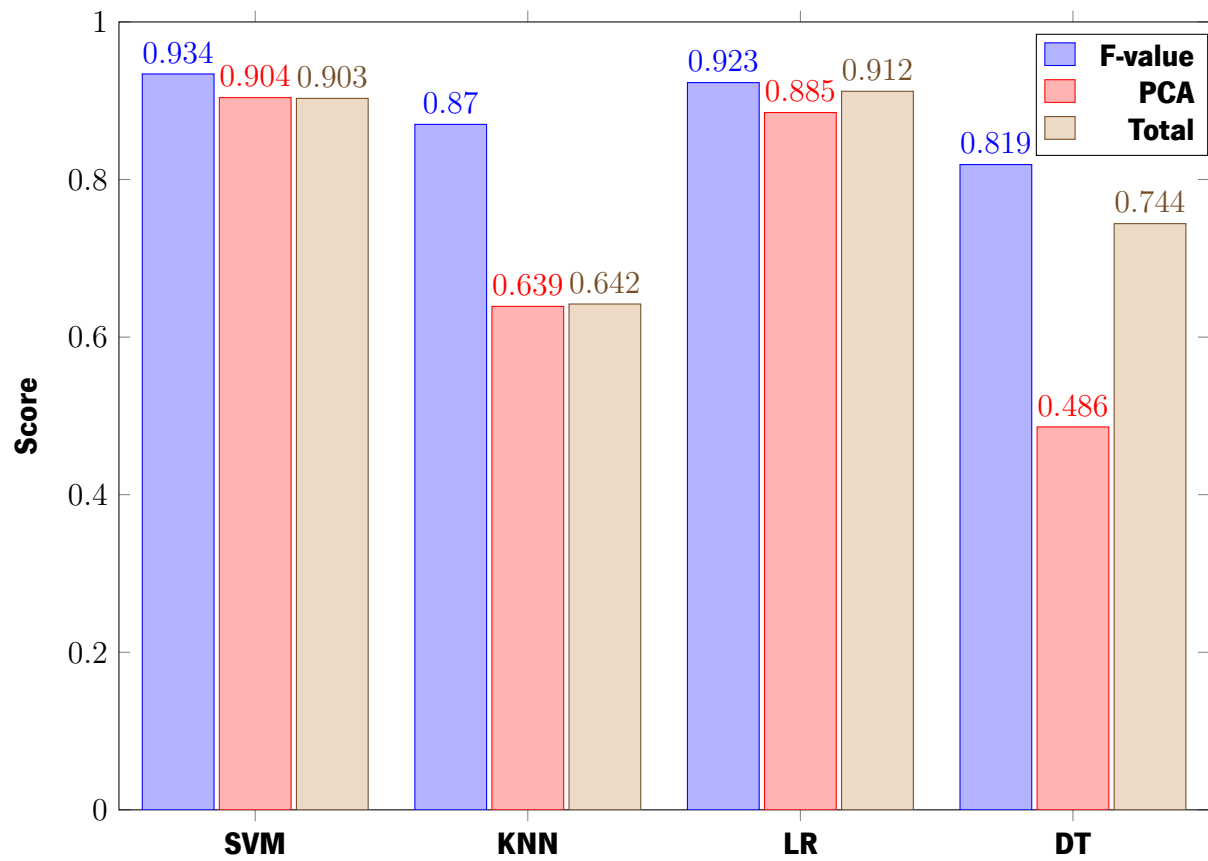


Figure 3: Comparison between different Machine Learning (ML) models, using three different feature selection methods: F-value, Principal Component Analysis (PCA) and Total(all features) and considering four different models: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Linear Regression (LR) and Decision Tree (DT).

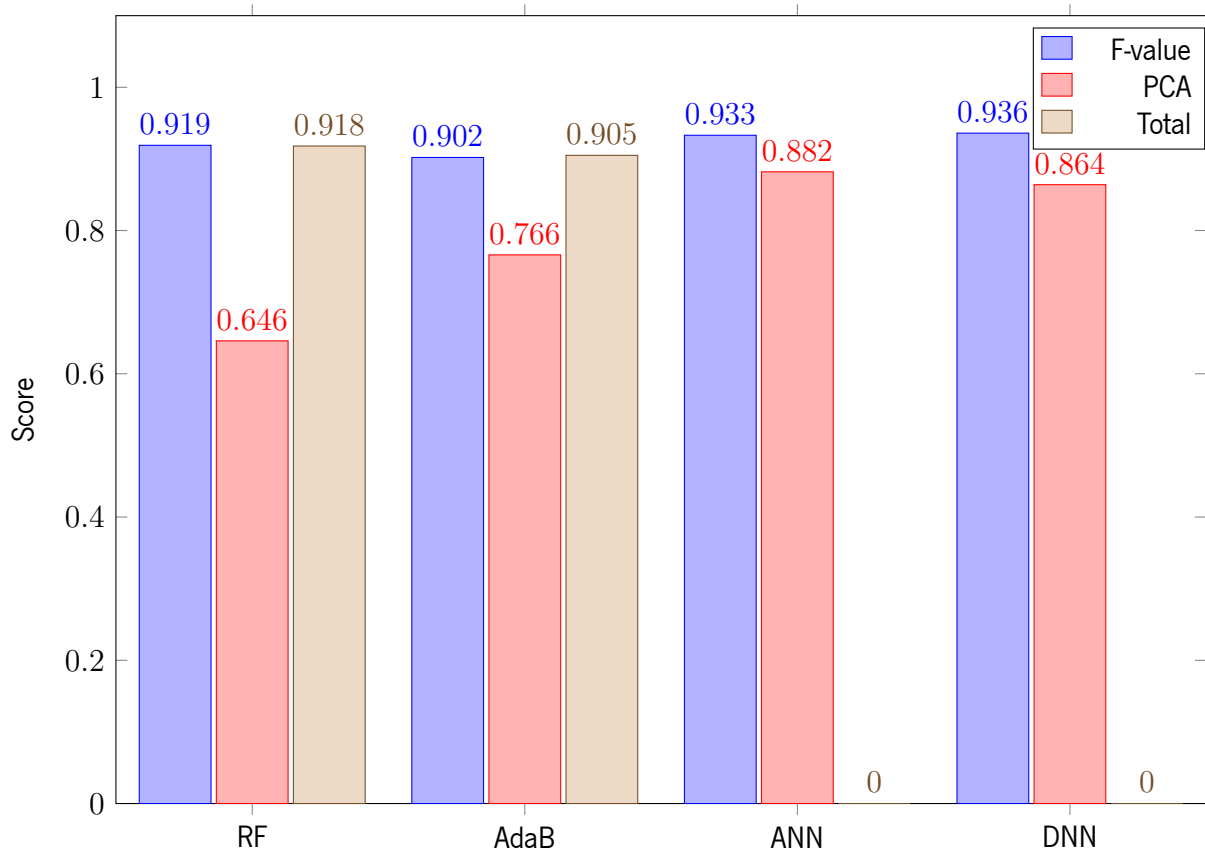


Figure 4: (continued) Comparison between different Machine Learning (ML) models, using three different feature selection methods: F-value, Principal Component Analysis (PCA) and Total (all features) and considering four different models: Random Forests (RF), adaboost, Artificial Neural Networks (ANN) and Deep Neural Network (DNN).

4.2 Cancer Diagnosis

Choosing the best feature selection method is important but choosing the right model is vital, especially in a medical context where a misclassification can have serious consequences. Below, a consolidated analysis of the performance of the various ML models is presented using their respective confusion matrix (Figures 5 to 8) and tables of results (Tables 6 to 9) after being made with the best parameters resumed in table 5 from the previous hyperparameter optimization part.

Table 5: Optimal parameters for Machine Learning (ML) models obtained through the CustomRandomizedSearchCV class for prediction if a individual has cancer

Model	Hyperparameters
KNN	Weights: [uniform] N_neighbors: [4] Metric: [manhattan] Algorithm: [ball_tree]
AdaBoost	N_estimators: [200] Learning_rate: [0.19630406500402714] Algorithm: [SAMME.R]
ANN	Solver: [lbfgs] Max_iter: [500] Learning_rate_init: [0.0019287915080207778] Learning_rate: [constant] Hidden_layer_sizes: [1024] Batch_size: [auto]
DNN	Solver: [adam] Max_iter: [200] Learning_rate_init: [0.0012650337203959038] Learning_rate: [constant] Hidden_layer_sizes: [256, 128, 128, 128] Batch_size: [512]

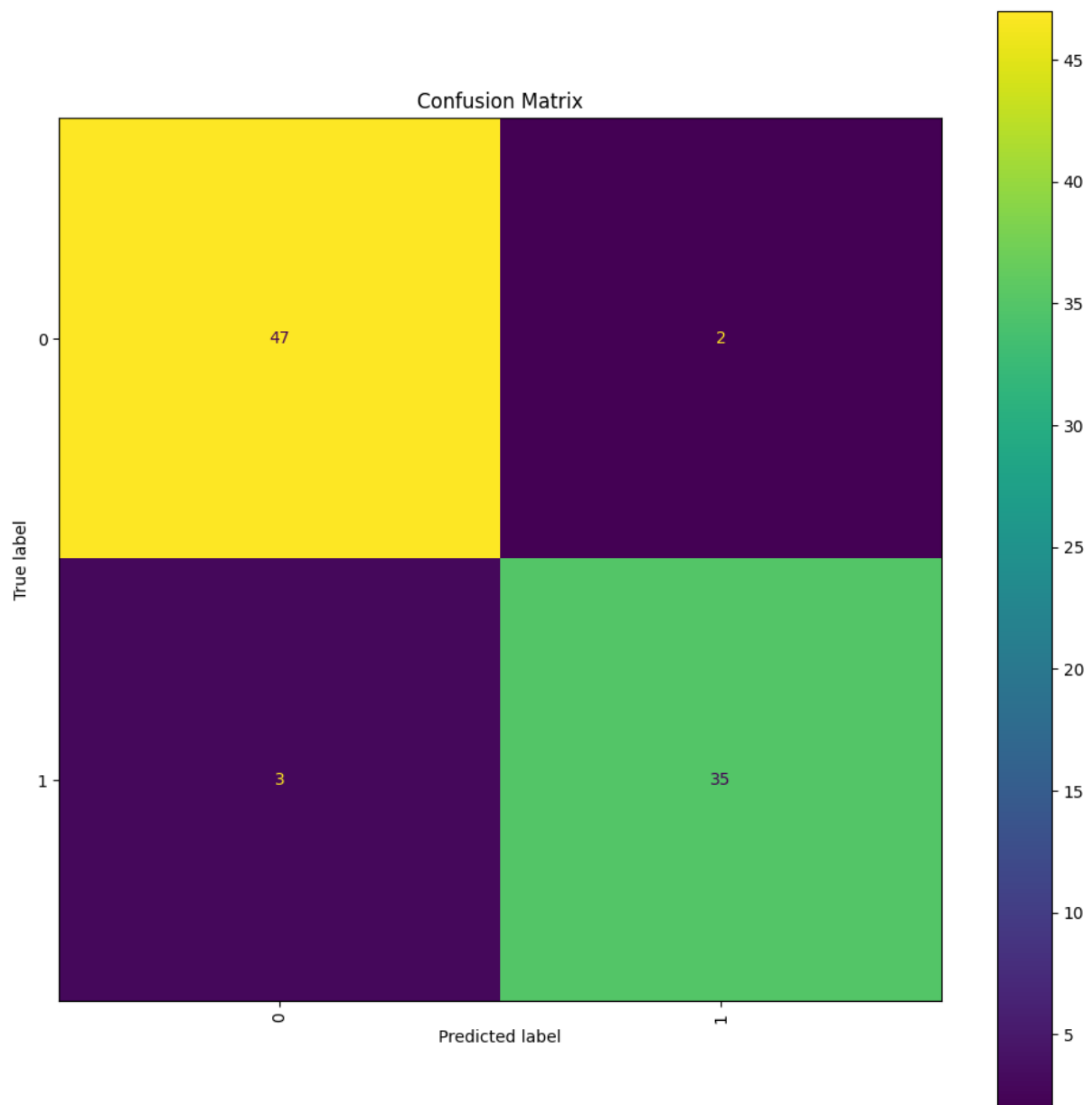


Figure 5: Confusion Matrix of K-Nearest Neighbors (KNN) model. True label represents the real state of the individual and predicted label represents the state predicted by the model for this individual. Top left corner (yellow): true negatives (TN). Top right corner (purple): false positives (FP). Bottom left corner (purple): false negatives (FN). Bottom right (green): true positives (TP).

Table 6: Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of K-Nearest Neighbors (KNN) model

	Precision	Recall	F1-score	Support
Healthy individual (0)	0.94	0.96	0.95	49
Individual with cancer (1)	0.95	0.92	0.93	38
Accuracy			0.94	87
Macro avg	0.94	0.94	0.94	87
Weighted avg	0.94	0.94	0.94	87

In the figure 5 the confusion matrix is a crucial tool in evaluating model performance and provides a detailed view of how the model behaves with respect to correct and incorrect predictions. When analyzing the confusion matrix provided, it is noted that for the Healthy individual (0) there were 47 correct predictions, that is, the cases in which the model correctly identified a healthy individual, while only 2 were incorrectly classified as Individual with cancer (1). This implies that the model has a robust ability to identify healthy individuals with a very low false positive rate. On the other hand, when looking at Individual with cancer (1), 35 predictions were correctly classified, indicating that the model effectively recognized the presence of cancer. However, 3 individuals were incorrectly classified as Healthy individual (0), suggesting that there are a number of false negatives. This information is extremely relevant as it indicates that, although the model has a high overall accuracy, there is still a small risk of not correctly identifying an individual with cancer, which, in a clinical context, can have significant implications.

In the context of the KNN model applied to detect the presence or absence of cancer, a detailed analysis of the results presented in the metrics table 6 is essential. When considering accuracy, it is observed that the model presents a notable performance for both the Healthy individual (0) and the Individual with cancer (1), with values of 0.94 and 0.95, respectively. This suggests that the model is capable of correctly identifying both cases with high accuracy, making very few type I or type II errors. Furthermore, when observing recall, the model displays a performance of 0.96 for the Healthy individual (0) and 0.92 for the Individual with cancer (1), which indicates a high capacity to identify true positives, especially for healthy individuals.

The f1-score metric, which combines precision and recall, provides a harmonized representation of the model's overall performance. For the Healthy individual (0), the f1-score is 0.95 and for the Individual with cancer (1) it is 0.93. These values demonstrate that the balance between precision and recall is well maintained in both classes.

Regarding the binary nature of the precision value, this implies that we are trying to predict one of two possible classes. This is why there are two precision values in the table, each corresponding to one of the classes (Healthy individual (0) and Individual with cancer (1)). The accuracy of each class is crucial to determining how well the model can identify each class without confusing it for another.

Last but not least, it is crucial to address the accuracy metric, which stands at 0.94. This value suggests that, of all predictions made by the model, 94% were correct. The weighted average and the macro average also have the same value as the accuracy. This is a testament to the model's excellent performance in the task, which reinforces its reliability and effectiveness in predicting the presence or absence of cancer.

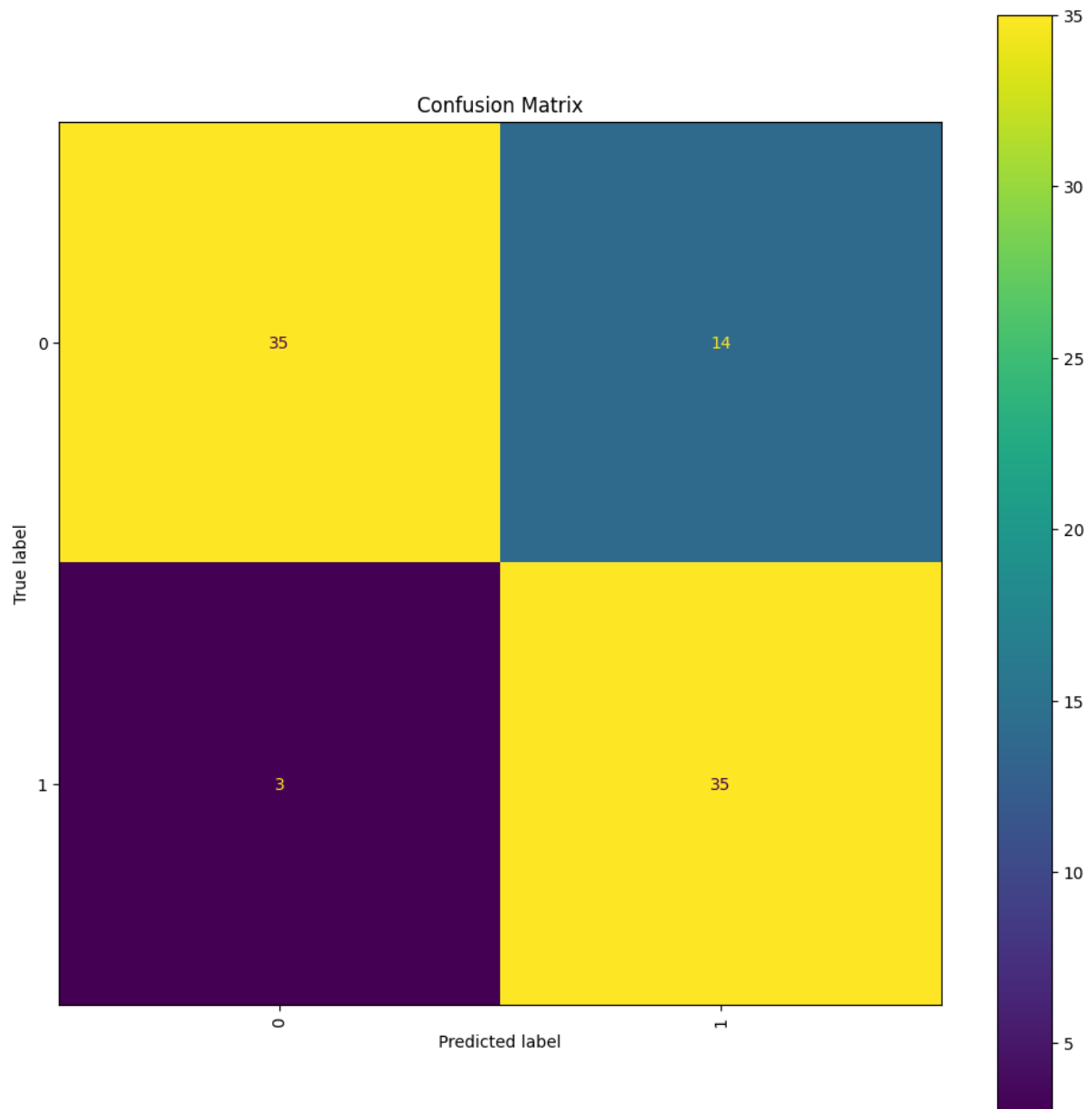


Figure 6: Confusion Matrix of adaboost model. True label represents the real state of the individual and predicted label represents the state predicted by the model for this individual. Top left corner (yellow): true negatives (TN). Top right corner (purple): false positives (FP). Bottom left corner (purple): false negatives (FN). Bottom right (yellow): true positives (TP).

Table 7: Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of adaboost model

	Precision	Recall	F1-score	Support
Healthy individual (0)	0.92	0.71	0.80	49
Individual with cancer (1)	0.71	0.92	0.80	38
Accuracy			0.80	87
Macro avg	0.82	0.82	0.80	87
Weighted avg	0.83	0.80	0.80	87

The figure 6, the confusion matrix provided for the adaboost model, it is observed that the number of true positives for the Healthy individual (0) is 35, which means that 35 healthy individuals were correctly classified by the model, while 14 Healthy individuals (0) were incorrectly classified as Individual with cancer (1). On the other hand, for the Individual with cancer (1), the model was able to correctly identify 35 individuals, while only 3 were incorrectly classified as Healthy individual (0). This result indicates that the model has considerable sensitivity in detecting cancer cases, given the high number of true positives for the Individual with cancer class (1), but shows some limitations in its ability to correctly identify healthy individuals, given the number of false positives for the individual Healthy class (0).

Analyzing the metrics table 7 of the model, it aims to predict whether an individual has cancer or not. The model displays an accuracy of 0.92 for the Healthy individual (0), suggesting that when it identifies an individual as healthy, there is a high probability of being correct. However, the recall of 0.71 for this same class indicates that the model correctly identifies only 71% of the truly healthy individuals present in the dataset. On the other hand, for the Individual with cancer (1), the precision is 0.71 and the recall is 0.92, which reveals that, despite some of the predictions being inaccurate, the model is effective in identifying the vast majority of individuals who actually have cancer. The f1-score of 0.80 for both classes suggests that the model maintains a balance between precision and recall. Regarding the averages, the macro average of 0.82 indicates that the model has an almost uniform performance for both classes, while the weighted average of 0.83 demonstrates that, when taking into account the number of samples in each class, the performance remains robust. The overall accuracy of the model is 0.80, indicating that it makes correct predictions in 80% of cases.

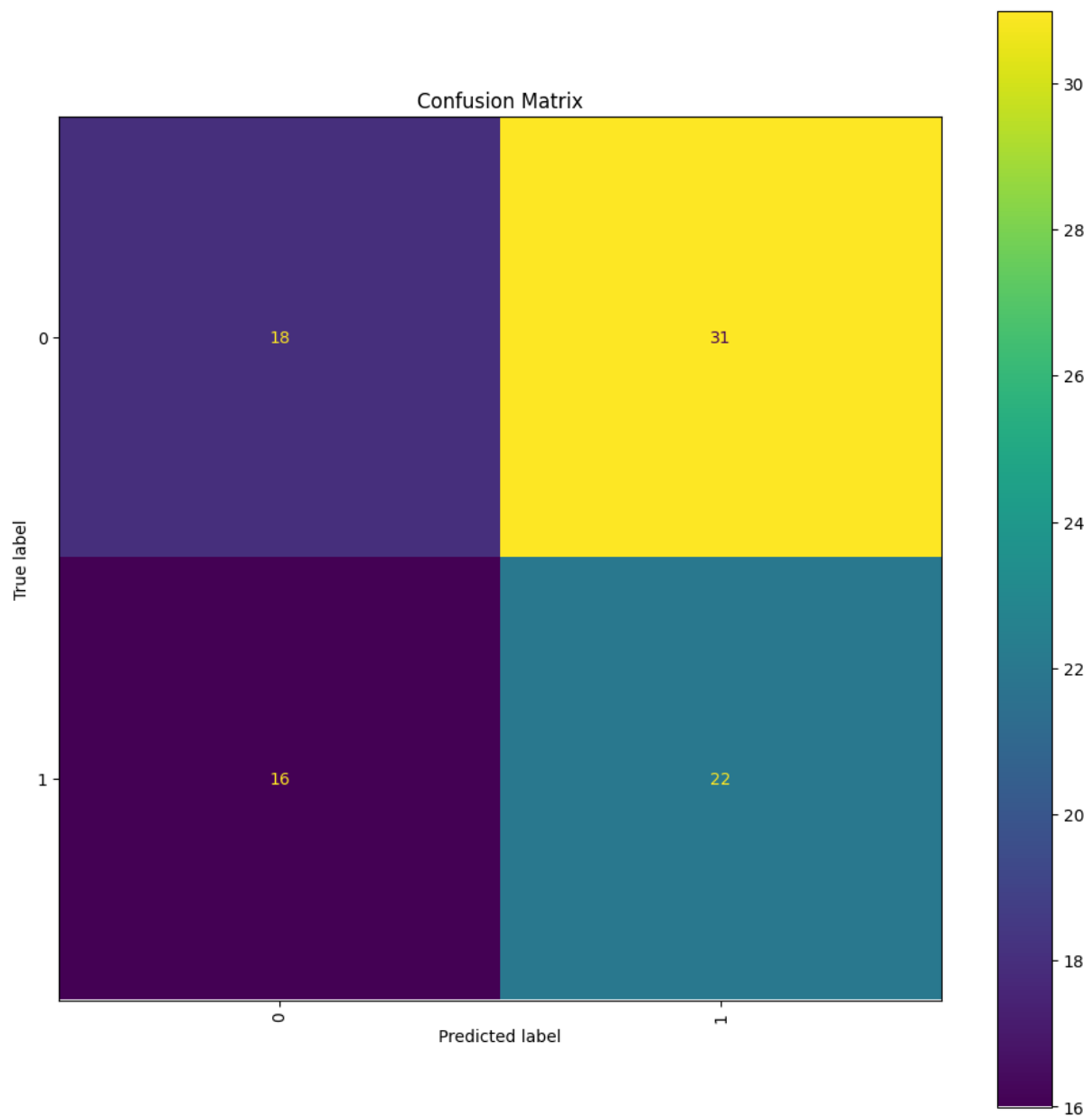


Figure 7: Confusion Matrix of Artificial Neural Networks (ANN) model. True label represents the real state of the individual and predicted label represents the state predicted by the model for this individual. Top left corner (purple): true negatives (TN). Top right corner (yellow): false positives (FP). Bottom left corner (purple): false negatives (FN). Bottom right (blue): true positives (TP).

Table 8: Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of Artificial Neural Networks (ANN) model

	Precision	Recall	F1-score	Support
Healthy individual (0)	0.53	0.37	0.43	49
Individual with cancer (1)	0.42	0.58	0.48	38
Accuracy			0.46	87
Macro avg	0.47	0.47	0.46	87
Weighted avg	0.48	0.46	0.46	87

Analyzing the confusion matrix in figure 7 provided for the model in question, it is observed that the true Healthy individuals (0), 18 were correctly classified by the model, while 31 were erroneously classified as Individuals with cancer (1). This indicates a tendency for the model to misclassify healthy individuals, potentially leading to a high number of false positives. In contrast, of the true Individuals with cancer (1), 22 were correctly identified by the model, and 16 were misclassified as Healthy individuals (0). Although the model was able to identify a significant portion of individuals with cancer, it still presents difficulties, as it makes mistakes in a considerable portion of these classifications, which can result in false negatives. False negatives, in the context of cancer diagnosis, are particularly worrying as they can lead to delays in treatment.

From the metrics table 8 provided for this model, it is clear that its predictive ability to determine whether or not an individual has cancer presents challenges. The accuracy for an individual Healthy (0) is 0.53, which indicates that when the model predicts that an individual is healthy, it is correct just over half the time. The recall of 0.37 for the same class suggests that the model correctly identified only 37% of the Healthy individuals (0) present in the test set. In relation to Individual with cancer (1), the precision is 0.42, while the recall is higher, standing at 0.58. This metric suggests that the model tends to misclassify many individuals as having cancer, but manages to capture 58% of those who actually have the disease. The f1-score, which is a harmonic mean between precision and recall, is 0.43 and 0.48 for the Healthy individual (0) and the Individual with cancer (1), respectively. These values point to a moderate performance of the model. As for the general metrics, both the macro average and the weighted average of the metrics are aligned with the global accuracy, which is 0.46, revealing consistency in the model's predictions, although it is operating below the 50% mark.

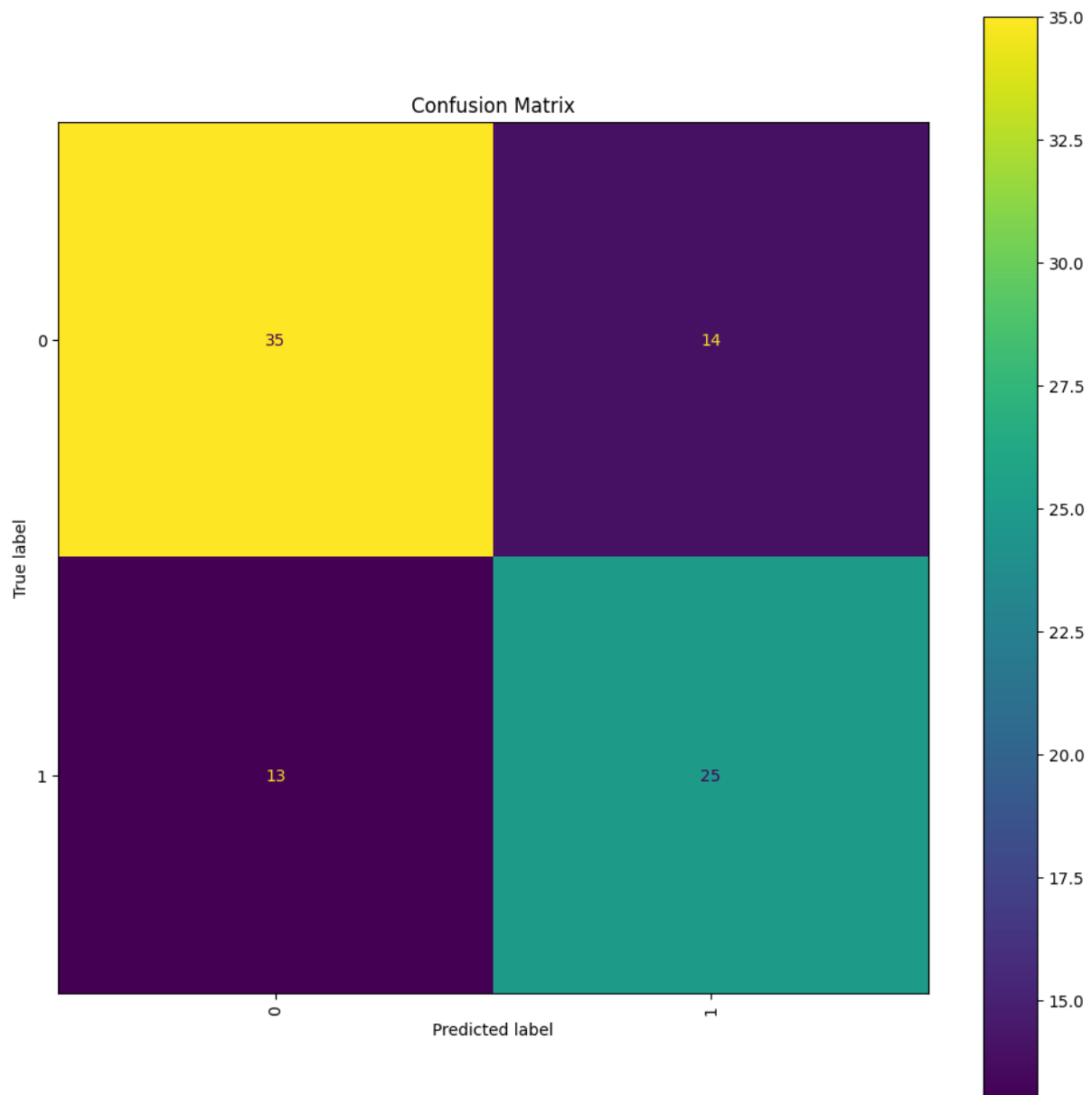


Figure 8: Confusion Matrix of Deep Neural Network (DNN) model. True label represents the real state of the individual and predicted label represents the state predicted by the model for this individual. Top left corner (yellow): true negatives (TN). Top right corner (purple): false positives (FP). Bottom left corner (purple): false negatives (FN). Bottom right (green): true positives (TP).

Table 9: Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of Deep Neural Network (DNN) model

	Precision	Recall	F1-score	Support
Healthy individual (0)	0.73	0.71	0.72	49
Individual with cancer (1)	0.64	0.66	0.65	38
Accuracy			0.69	87
Macro avg	0.69	0.69	0.69	87
Weighted avg	0.69	0.69	0.69	87

This figure 9 or matrix provides a clear visualization of where the model got it right and where it got it wrong, allowing for a deeper understanding of areas where the model may need improvements or adjustments. The top left corner, represented in yellow, shows the number of true positives for the Healthy individual (0), which is 35, that is, 35 healthy individuals were correctly classified by the model. The upper right corner, in purple, indicates false positives, that is, 14 Healthy individuals (0) were incorrectly classified as Individuals with cancer (1). Analyzing the bottom left corner, also in purple, 13 Individuals with cancer (1) were erroneously classified as Healthy individuals (0), representing false negatives. Finally, the bottom right corner, in green, shows that 25 Individuals with cancer (1) were correctly identified, corresponding to the true positives.

By looking at the metrics table 9 for the deep neural network (DNN) model, it is possible to discern several aspects about its performance. Regarding the individual Healthy (0), the model demonstrated an accuracy of 0.73, meaning that 73% of the predictions it made for that class were correct. The recall, or sensitivity, was 0.71, which reveals that the model was able to correctly identify 71% of all true Healthy individuals (0) present in the dataset. The F1-score, which is a harmonic metric between precision and recall, stood at 0.72.

As for Individual with cancer (1), the accuracy was 0.64, indicating that 64% of the predictions in this class were correct. Recall was slightly higher, reaching 0.66, which suggests that the model correctly identified 66% of true Individuals with cancer (1). The F1-score for this class was 0.65, remaining consistent with previous metrics.

Overall, the model had a weighted average precision of 0.69, and an equally weighted average recall of 0.69. The weighted F1-score followed these values. The avg macro metric also showed values of 0.69 for precision, recall and F1-score. The model's overall accuracy was 0.69, that is, it made correct predictions 69% of the time.

The results in predicting the presence or absence of cancer showed that the KNN model achieved the best results, with an accuracy of 94%. Also, the KNN model demonstrated superior performance, particularly with regard to the recall metric. With a recall of 0.96 for healthy individuals and 0.92 for cancer individuals, KNN exhibited a unique ability to minimize false negatives. Such an ability is critical in a medical environment, where failing to detect a positive case can have devastating consequences. This performance can be attributed to the simplicity and effectiveness of KNN, which makes no specific assumptions about the distribution of the underlying data, making it less prone to overfitting and more capable of generalizing well to new data, especially when the set of data is small or imbalanced [148, 149, 77, 150].

AdaBoost, although it did not perform as effectively as KNN in terms of overall precision, demonstrated a high recall of 0.92 for the "Individuals with Cancer" class, indicating a remarkable ability to identify positive cases. However, the inherent complexity and sensitivity to outliers can be disadvantageous in medical datasets, where abrupt variabilities are common [151].

The analysis also explored neural network models, where ANN proved to be inadequate for the dataset under study, with a precision of just 46% and low recall values. In contrast, DNN achieved a moderate precision of 69% and had more acceptable recall values [152, 153, 154]. However, even with these limitations, DNN exhibits a greater ability to represent complex relationships in data compared to ANN, which may explain its relatively better performance.

The lack of a balanced dataset, with only 144 samples from healthy individuals, may have impacted the robustness of the model. Because of the notable imbalance in the data set (almost 1,900 samples from cancer patients versus just 144 individuals without cancer) a subsampling approach was chosen where both classes were limited to 144, this improved the results. The results were better than those not presented here where all samples were used. Because in the first attempt used all the samples from people with cancer (1980) and all from people without cancer (144) and the results obtained in the metrics tables were very poor, obtaining accuracy equal to zero on some models. The decision not to use the "oversampling" technique was to preserve the integrity and reliability of the data in a medical context, especially since creating synthetic data in a medical context can introduce variables that deviate the model from biological reality.

This problem with the number of samples may also have affected the list of 1500 genes of the feature selection aimed at predicting the presence or absence of cancer, the number of genes in common is notably small, with only 2 genes shared with PAM50 list. The first was *ANLN*, which is involved in cytokinesis, the final process of cell division [155]. It has been associated with cell proliferation in several types of cancer,

making it an interesting target for anticancer therapies. The second was *NUF2*, which is an essential component of the centromere-associated complex, playing a fundamental role in the correct segregation of chromosomes during cell division. Its dysfunction can lead to aneuploidy, a common feature in many cancer cells [156]. The fact that all genes, including those in common, have very similar and low trait importance values may suggest that the prediction of this condition is not strongly influenced by a small set of genes, but possibly by a broader grouping of genetic and epigenetic factors [157].

The genes with the highest importance value in predicting the presence or absence of cancer are not on the pam50 list. The first was *HIGD1B*, which has been implicated in oxidative stress playing a role in cell survival [158]. Secondly, *TNNT2* was a gene associated with the heart muscle, however, its relationship with breast cancer is unknown [159]. Third, *RASL11A*, which is involved in signal transduction processes, is associated with cell proliferation [160]. Next, *OTUD6B*, which is related to the regulation of deubiquitination, has an indirect relationship with protein stability in cancer cells [161]. Lastly, *PMS2P3* is associated with DNA repair and maintenance of genomic integrity [162].

The results presented by ML models, namely in predicting the presence or absence of breast cancer, demonstrated that did not reach an optimal level of precision. This limitation can be, to a large extent, attributed to the biological complexity inherent to the disease. The absence of a specific set of genes that, in itself, is decisive in predicting breast cancer indicates that the genetics of this pathology are multifaceted and are not limited to a limited subset of genes.

Biologically speaking, the heterogeneity of breast cancer and the interaction of multiple genetic and epigenetic factors may have contributed to the difficulty of models in making accurate predictions. The PAM50 list, although recognized, may not completely cover all genes relevant to oncogenesis and disease progression. The identification of genes such as *HIGD1B*, *TNNT2*, among others, which are not present in the PAM50 list, highlights this idea.

Additionally, imbalance in the dataset, with a significantly lower representation of healthy individuals compared to cancer patients, may have hampered the models' ability to generalize appropriately. This disparity emphasizes the importance of data balancing in medical contexts, where each sample has critical relevance.

In resume, ML models were not able to fully capture the complexity and biological variability of breast cancer, leading to results that, although promising in certain aspects, still have room for optimization.

4.3 Breast Cancer (BC) subtype prediction

In the context of predicting cancer subtypes, diagnostic accuracy becomes even more critical, as each subtype may require a distinct treatment and specific management approach. Below is a consolidated analysis of the performance of different ML models in predicting cancer subtypes. This analysis is illustrated through the corresponding confusion matrices (Figures 9 to 14) and tables of results (Tables 11 to 16), after applying the best parameters identified in table 10 of the section on hyperparameter optimization.

Table 10: Optimal parameters for Machine Learning (ML) models obtained through the CustomRandomizedSearchCV class for prediction of cancer subtype "three-gene classifier"

Model	Hyperparameters
SVM	Max1_iter: [1000] Kernel: [linear] Gamma: [1.7188391428171454] C: [5.220567527846975]
LR	Solver: [saga] Penalty: [l1] Multi_class: [multinomial] Max_iter: [5000] C: [0.8620689655172413]
RF	Min_samples_split: [16] Max_features: [sqrt] Max_depth: [16] Criterion: [gini]
AdaBoost	N_estimators: [200] Learning_rate: [1.0] Algorithm: [SAMME]
ANN	Solver: [sgd] Max_iter: [500] Learning_rate_init: [0.02596655972934871] Learning_rate: [adaptive] Hidden_layer_sizes: [1024] Batch_size: [auto]
DNN	Solver: [lbfgs] Max_iter: [500] Learning_rate_init: [0.5831305113526224] Learning_rate: [invscaling] Hidden_layer_sizes: [256, 128, 128] Batch_size: [64]

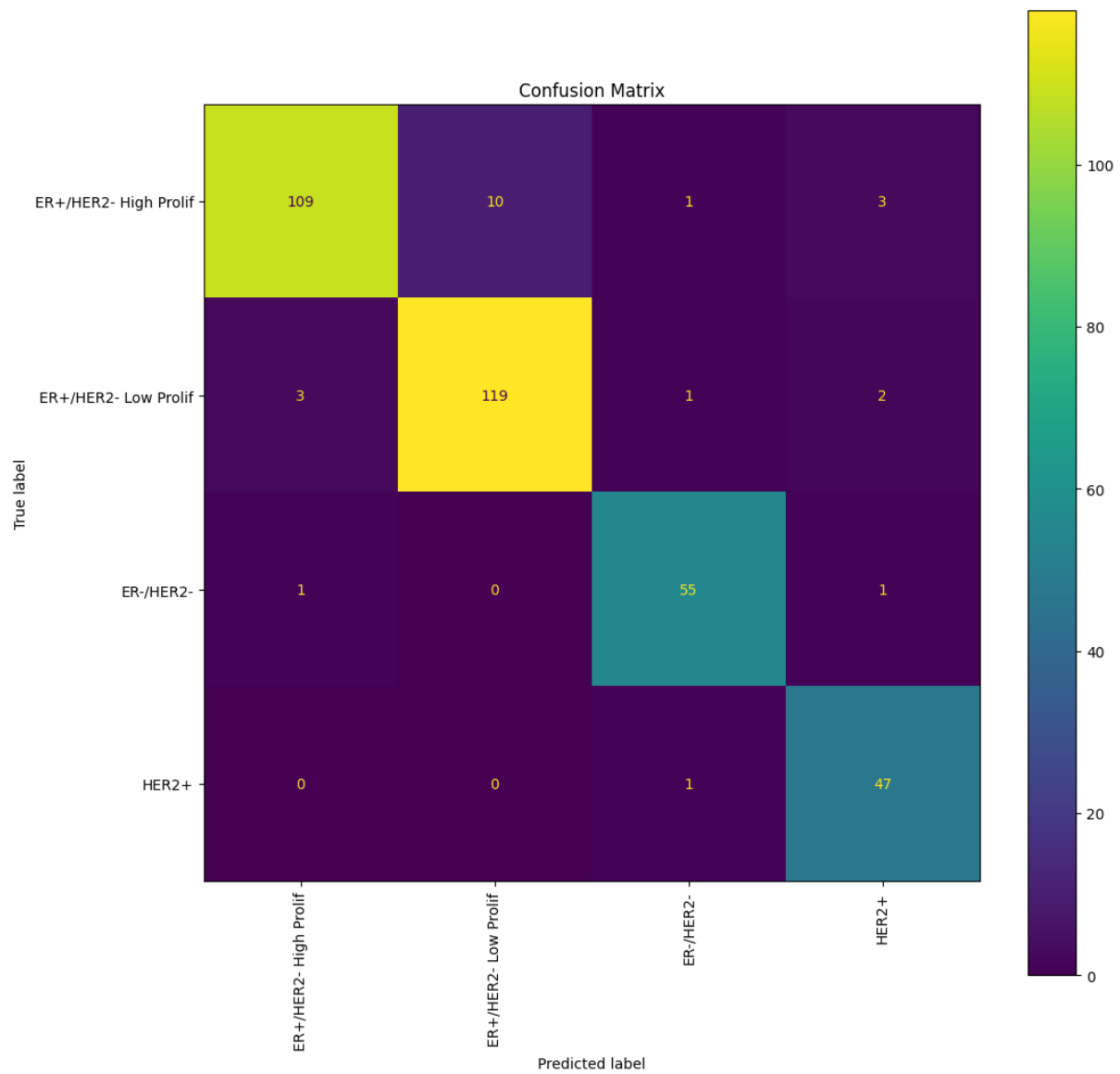


Figure 9: Confusion Matrix of Support Vector Machines (SVM) model. Vertical axis (True label): Represents the actual cancer subtypes for patients: "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif", "ER-/HER2-", and "HER2+". Horizontal axis (Predicted label): Represents the predicted cancer subtypes for patients: "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif", "ER-/HER2-", and "HER2+". Numbers in diagonal cells (yellow and blue) represent correct predictions, while numbers outside this diagonal represent incorrect predictions.

Table 11: Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of Support Vector Machines (SVM) model

	Precision	Recall	F1-score	Support
ER+/HER2- High Prolif	0.96	0.89	0.92	123
ER+/HER2- Low Prolif	0.92	0.95	0.94	125
ER-/HER2-	0.95	0.96	0.96	57
HER2+	0.89	0.98	0.93	48
Accuracy			0.93	353
Macro avg	0.93	0.95	0.94	353
Weighted avg	0.94	0.93	0.93	353

Observing the confusion matrix in figure 9 of the SVM model for BC subtype prediction, several insights can be extracted. Of the 123 true observations in the "ER+/HER2- High Prolif" category, the model correctly predicted 109 of them, but there were 10 cases that were incorrectly classified as "ER+/HER2- Low Prolif", 1 case as "ER-/HER2-" and 3 cases as "HER2+", indicating that most confusions of this subtype were with "ER+/HER2- Low Prolif". Of the 125 true observations in the "ER+/HER2- Low Prolif" category, 119 were correctly identified by the model, with errors distributed among 3 cases classified as "ER+/HER2- High Prolif", 1 case as "ER-/HER2-" and 2 cases as "HER2+". The "ER-/HER2-" category showed strong performance with 55 of the 57 true observations classified correctly and only 1 case was incorrectly identified as "ER+/HER2- High Prolif" and 1 case as "HER2+". Among the 48 true observations in the "HER2+" category, 47 were correctly predicted with only 1 case incorrectly classified as "ER-/HER2-". Overall, the model performs well in distinguishing between cancer subtypes.

When analyzing the metrics presented in table 11 for the SVM model in predicting the breast cancer subtype, several aspects deserve detailed attention. The accuracy metric, which measures the proportion of true positives to total predicted positives, reveals strong results across all categories. The "ER+/HER2- High Prolif" subtype exhibits an accuracy of 0.96, suggesting that the model is highly effective in accurately identifying this subtype. The "ER-/HER2-" subtype has a slightly lower accuracy of 0.95, still extremely robust, while "ER+/HER2- Low Prolif" and "HER2+" record accuracies of 0.92 and 0.89, respectively.

Regarding the recall metric, which measures the proportion of true positives in relation to the total number of real positive cases, the values are also impressive. The "HER2+" subtype stands out with a recall of 0.98, followed by "ER-/HER2-" with 0.96 and "ER+/HER2- Low Prolif" with 0.95. The "ER+/HER2- High Prolif" subtype has a recall of 0.89, which is still strong but slightly lower than the other categories.

The F1-score, which is the harmonic mean between precision and recall, also provides a comprehensive view of the model's effectiveness. In this case, all subtypes present values above 0.90, with emphasis on "ER-/HER2-" with 0.96 and "ER+/HER2- Low Prolif" with 0.94.

The weighted average and macro average of the metrics are also strong, both hovering around 0.93-0.94, indicating solid overall performance even taking into account the weight of each class (which is the case with the weighted average).

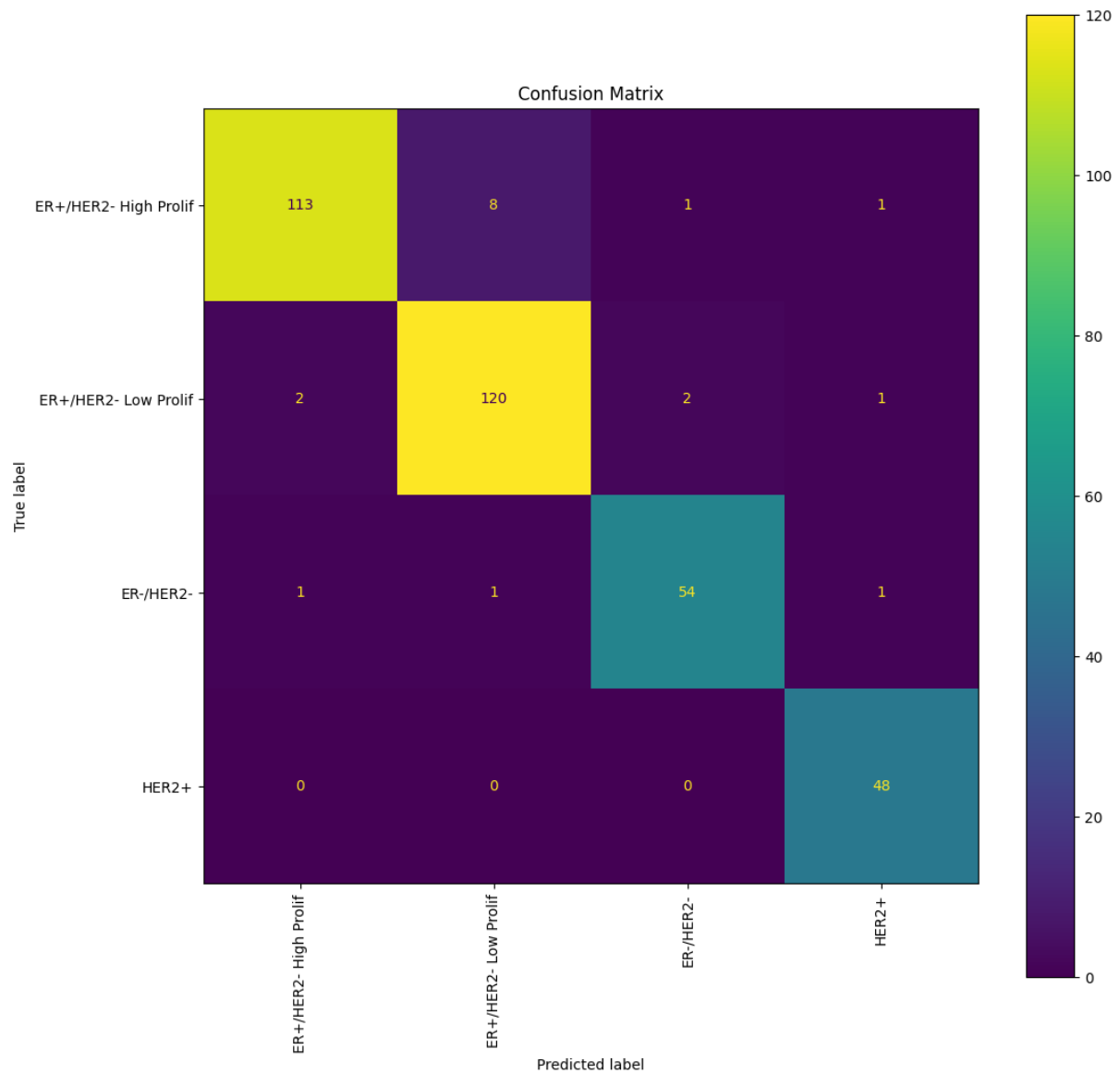


Figure 10: Confusion Matrix of Linear Regression (LR) model. Vertical axis (True label): Represents the actual cancer subtypes for patients: "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif", "ER-/HER2-", and "HER2+". Horizontal axis (Predicted label): Represents the predicted cancer subtypes for patients: "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif", "ER-/HER2-", and "HER2+". Numbers in diagonal cells (yellow and blue) represent correct predictions, while numbers outside this diagonal represent incorrect predictions.

Table 12: Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of Linear Regression (LR) model

	Precision	Recall	F1-score	Support
ER+/HER2- High Prolif	0.95	0.89	0.92	123
ER+/HER2- Low Prolif	0.91	0.94	0.93	125
ER-/HER2-	0.95	0.95	0.95	57
HER2+	0.94	1.00	0.97	48
Accuracy			0.93	353
Macro avg	0.94	0.95	0.94	353
Weighted avg	0.94	0.93	0.93	353

The matrix analysis in figure 10 allows for a deep understanding of correct and incorrect model predictions. Observing the "ER+/HER2- High Prolif" subtype, it appears that the model correctly identified 113 samples, while 8 samples were incorrectly classified as "ER+/HER2- Low Prolif", 1 sample was wrongly classified as "ER-/HER2-" and another as "HER2+". Regarding the "ER+/HER2- Low Prolif" subtype, 120 samples were correctly identified, while two were confused with "ER+/HER2- High Prolif", two with "ER-/HER2-" and one with "HER2+". For the "ER-/HER2-" subtype, 54 samples were correctly classified and two samples were incorrectly categorized, one as "ER+/HER2- High Prolif" and the other as "ER+/HER2- Low Prolif". Finally, the "HER2+" subtype presented an impeccable classification, with 48 samples correctly identified and no confusion with other subtypes.

This matrix shows that the model has an excellent ability to distinguish the "HER2+" subtype, without errors, and also a remarkable performance for the "ER-/HER2-" subtype. However, the "ER+/HER2- High Prolif" and "ER+/HER2- Low Prolif" subtypes showed a slightly higher number of incorrect classifications among themselves, suggesting that these two subtypes may present more similar characteristics, leading to greater difficulty of the model in correctly distinguishing between them.

Analyzing the score metrics presented in table 12, a picture of high accuracy in the model predictions can be discerned. The "ER+/HER2- High Prolif" subtype boasts a precision of 0.95 and a recall of 0.89, resulting in an f1-score of 0.92. This means that, of this subtype, 95% of the predictions made by the model are correct, but only 89% of the real cases of this subtype were correctly identified by the model. In turn, the "ER+/HER2- Low Prolif" subtype has a slightly lower precision, 0.91, but a higher recall, 0.94, indicating that although a smaller proportion of predictions are correct, a greater proportion of real cases were identified. The f1-score for this subtype is 0.93, which is a harmonic mean between precision and

recall and, in this context, is slightly higher than the "High Prolif" subtype. The "ER-/HER2-" subtype exhibits consistently high values across all metrics, with precision and recall both set at 0.95, leading to an f1-score also of 0.95. Finally, the "HER2+" subtype demonstrates the highest precision and recall, with values of 0.94 and 1.00, respectively, which translates into an f1-score of 0.97, the highest among all subtypes.

The global accuracy of the model is 0.93, reflecting the proportion of correct predictions among all those made. The macro and weighted averages provide a global perspective of the metrics, with both averaging around 0.94, which corroborates the high performance of the model.

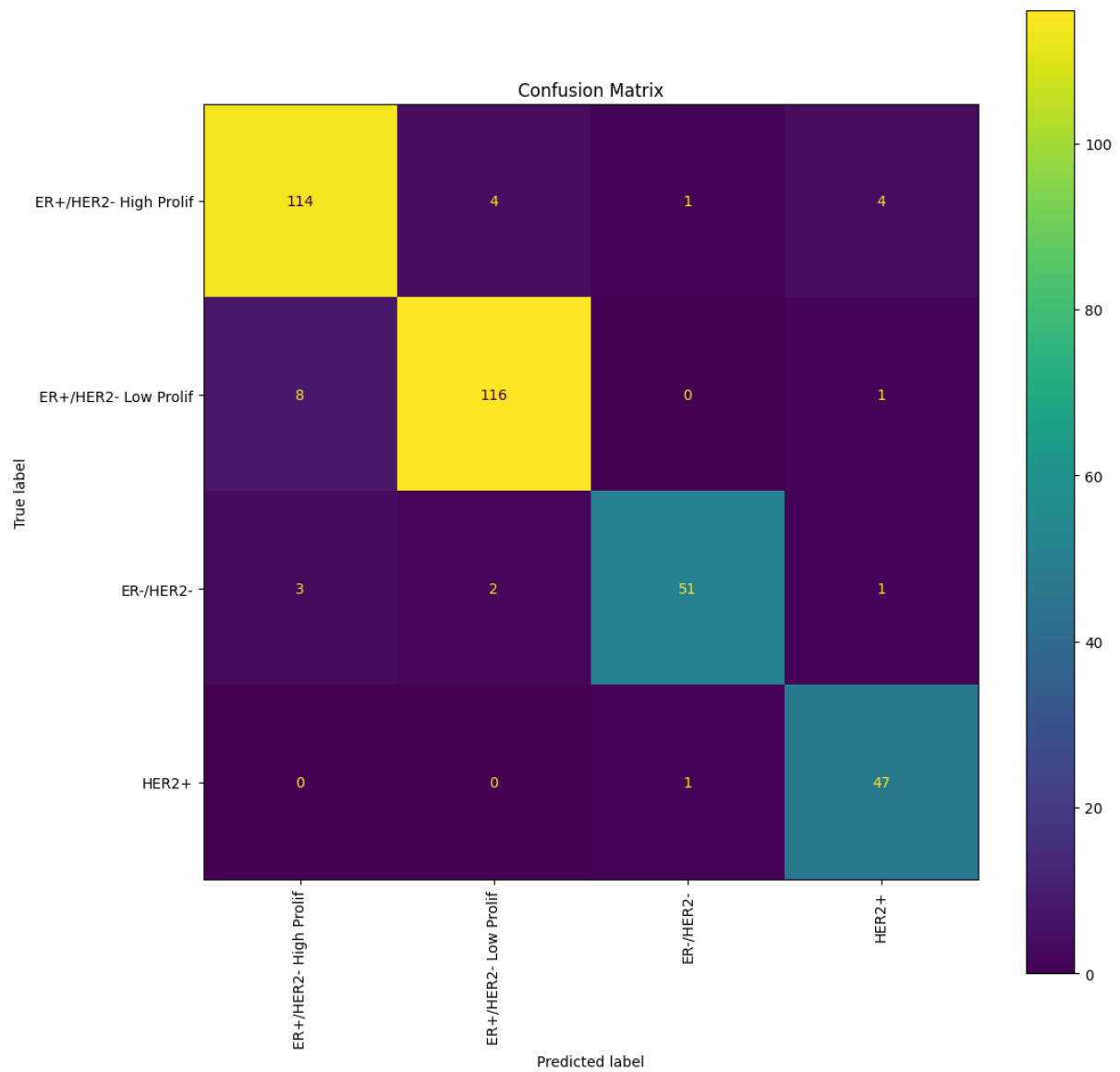


Figure 11: Confusion Matrix of Random Forests (RF) model. Vertical axis (True label): Represents the actual cancer subtypes for patients: "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif", "ER-/HER2-", and "HER2+". Horizontal axis (Predicted label): Represents the predicted cancer subtypes for patients: "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif", "ER-/HER2-", and "HER2+". Numbers in diagonal cells (yellow and blue) represent correct predictions, while numbers outside this diagonal represent incorrect predictions.

Table 13: Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of Random Forests (RF) model

	Precision	Recall	F1-score	Support
ER+/HER2- High Prolif	0.88	0.92	0.90	123
ER+/HER2- Low Prolif	0.94	0.92	0.93	125
ER-/HER2-	0.94	0.88	0.91	57
HER2+	0.88	0.92	0.90	48
Accuracy			0.91	353
Macro avg	0.91	0.91	0.91	353
Weighted avg	0.91	0.91	0.91	353

As referred, the main diagonal of the matrix in figure 11 shows the model hits, that is, the instances in which the predictions correspond to the true labels. Observing this diagonal, it appears that the model was able to correctly classify 114 samples for "ER+/HER2- High Prolif", 116 for "ER+/HER2- Low Prolif", 51 for "ER-/HER2-" and 47 for "HER2+ ". These values reinforce the high precision and recall observed in the previously discussed metrics.

However, a confusion matrix also highlights where the model made incorrect predictions. For example, 8 truly "ER+/HER2- Low Prolif" samples were classified as "ER+/HER2- High Prolif" and 4 truly "ER+/HER2- High Prolif" samples were classified as "ER+/HER2- Low Prolif". These confusions between "High Prolif" and "Low Prolif" may indicate that there are similar characteristics between these two classes that the model found difficult to distinguish. Additionally, it appears that the "ER-/HER2-" subtype had 3 samples classified as "ER+/HER2- High Prolif" and 2 as "ER+/HER2- Low Prolif", suggesting that there may be common characteristics or overlaps between these classes. Finally, the "HER2+" subtype showed a mixture of classifications, with one sample classified as "ER-/HER2-" and another as "ER+/HER2- High Prolif".

The RF model, used to predict breast cancer subtypes, shows remarkable performance, evidenced by the metrics provided in table 13. Analyzing the metrics individually by subtype, for "ER+/HER2- High Prolif", the precision is 0.88, the recall is 0.92 and the f1-score is 0.90. These values indicate that the model has an excellent ability to correctly predict this subtype, being more cautious when assigning a sample to this group, resulting in a recall greater than precision. For the "ER+/HER2- Low Prolif" subtype, the precision, recall and f1-score values are 0.94, 0.92 and 0.93, respectively, denoting an excellent balance between the metrics. The "ER-/HER2-" subtype has a precision of 0.94, recall of 0.88 and f1-

score of 0.91. The "HER2+" subtype records a precision of 0.88, a recall of 0.92 and an f1-score of 0.90.

When evaluating general metrics, such as the macro average and the weighted average, all values center around 0.91, demonstrating the robustness and balance of the model in general terms. This balance is essential to ensure that the model is not over-optimized for a specific subtype, but rather performs consistently across the various subtypes. Accuracy remains in this balance with 0.91.

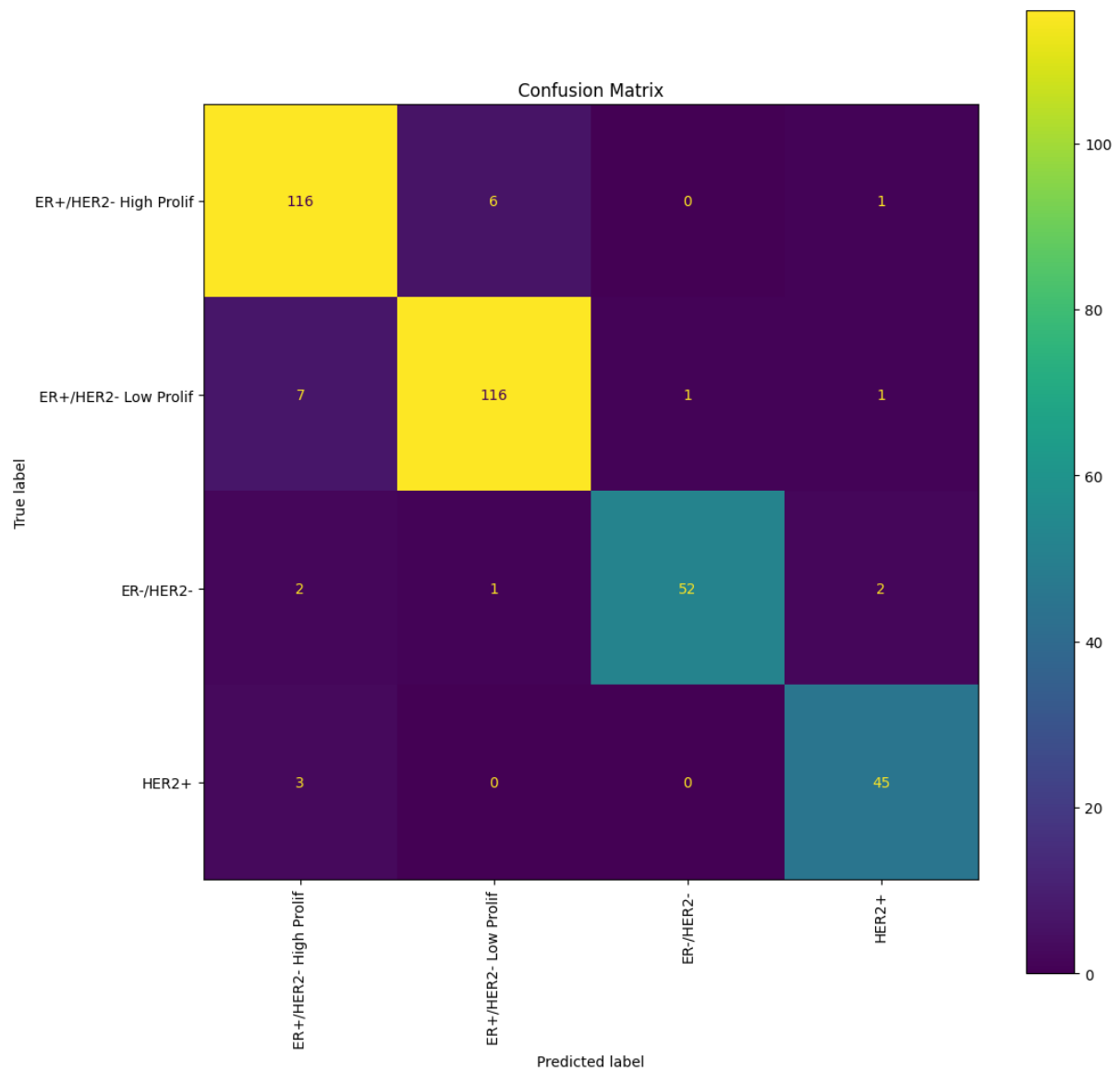


Figure 12: Confusion Matrix of adaboost model. Vertical axis (True label): Represents the actual cancer subtypes for patients: "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif", "ER-/HER2-", and "HER2+". Horizontal axis (Predicted label): Represents the predicted cancer subtypes for patients: "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif", "ER-/HER2-", and "HER2+". Numbers in diagonal cells (yellow and blue) represent correct predictions, while numbers outside this diagonal represent incorrect predictions.

Table 14: Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of adaboost model

	Precision	Recall	F1-score	Support
ER+/HER2- High Prolif	0.89	0.91	0.90	123
ER+/HER2- Low Prolif	0.90	0.90	0.90	125
ER-/HER2-	0.96	0.88	0.92	57
HER2+	0.92	0.94	0.93	48
Accuracy			0.91	353
Macro avg	0.92	0.91	0.91	353
Weighted avg	0.91	0.91	0.91	353

The confusion matrix presented in figure 12 represents the model prediction results for the different cancer subtypes. In the "ER+/HER2- High Prolif" class, 116 samples were correctly predicted, while 6 were incorrectly classified as "ER+/HER2- Low Prolif" and 1 as "HER2+". Regarding the "ER+/HER2- Low Prolif" class, 116 samples were correctly identified, 7 were incorrectly categorized as "ER+/HER2- High Prolif", 1 as "ER-/HER2-" and 1 as "HER2+". In the "ER-/HER2-" subtype, the model made the correct prediction for 52 samples, 2 were classified as "ER+/HER2- High Prolif", 1 as "ER+/HER2- Low Prolif" and 2 as "HER2+". Finally, for the "HER2+" subtype, the model correctly identified 45 samples and incorrectly classified 3 as "ER+/HER2- High Prolif". These results reflect the model's ability to correctly distinguish between different subtypes, as well as identify areas where improvements or adjustments may be needed.

The Adaboost model's metrics table 14 reveals remarkably good performance on several fronts. With a weighted average precision of 0.91 and a weighted average recall and f1-score also at 0.91, the model demonstrates substantial effectiveness in correctly classifying the different subtypes. In particular, the "ER-/HER2-" class has the highest accuracy of 0.96, indicating that the model has an exceptional ability to identify true positives within this category. However, its recall for this same class is relatively lower (0.88), suggesting a tendency for a greater number of false negatives compared to other classes.

The "ER+/HER2- High Prolif" and "ER+/HER2- Low Prolif" classes demonstrate precision, recall and f1-score values close to 0.90, which also indicates good performance of the model in these categories. On the other hand, the "HER2+" class has a precision and f1-score above 0.90, and a recall of 0.94, thus becoming the class with the highest recall, suggesting a robust ability of the model to capture most true positive cases in this category.

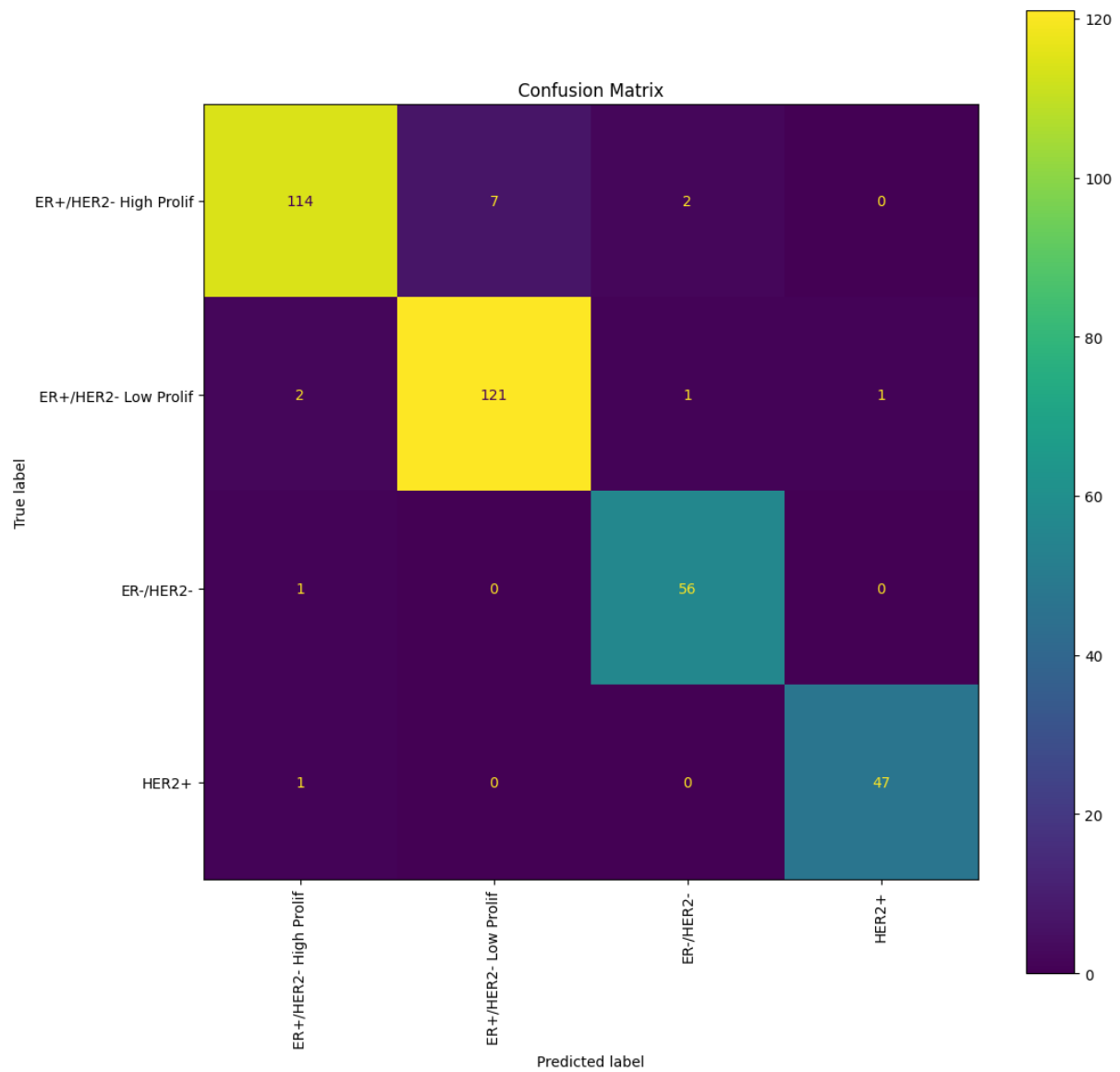


Figure 13: Confusion Matrix of Artificial Neural Networks (ANN) model. Vertical axis (True label): Represents the actual cancer subtypes for patients: "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif", "ER-/HER2-", and "HER2+". Horizontal axis (Predicted label): Represents the predicted cancer subtypes for patients: "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif", "ER-/HER2-", and "HER2+". Numbers in diagonal cells (yellow and blue) represent correct predictions, while numbers outside this diagonal represent incorrect predictions.

Table 15: Score metrics (precision, recall, f1-score, support, accuracy, macro average, weighted average) of Artificial Neural Networks (ANN) model

	Precision	Recall	F1-score	Support
ER+/HER2- High Prolif	0.94	0.92	0.93	123
ER+/HER2- Low Prolif	0.94	0.96	0.95	125
ER-/HER2-	0.95	0.96	0.96	57
HER2+	0.96	0.94	0.95	48
Accuracy			0.94	353
Macro avg	0.95	0.95	0.95	353
Weighted avg	0.94	0.94	0.94	353

The presented confusion matrix in figure 13 provides a visual representation of the single-layer neural network model predictions for classifying breast cancer subtypes. The diagonal values of the matrix indicate the correct predictions of the model, with the "ER+/HER2- High Prolif" class having 114 correct predictions, the "ER+/HER2- Low Prolif" 121, the "ER-/HER2-" 56 and to "HER2+" 47. These numbers reflect the high accuracy of the model in correctly identifying cancer subtypes.

However, there are also some glaring errors. The "ER+/HER2- High Prolif" class was confused 7 times with "ER+/HER2- Low Prolif" and 2 times with "ER-/HER2-". For "ER+/HER2- Low Prolif", there were 2 cases in which it was confused with "ER+/HER2- High Prolif" and 1 case for each of the other two classes. "ER-/HER2-" was misclassified once as "ER+/HER2- High Prolif", and "HER2+" was also confused once with "ER+/HER2- High Prolif".

Despite these confusions, it is clear that most of the model's predictions are found on the diagonal of the matrix, indicating that most of the predictions are correct.

Analysis of this model's metrics table 15 reveals high effectiveness in several classification dimensions. The average weighted precision, average weighted recall, and average weighted f1-score are all at 0.94, demonstrating excellent agreement between the model predictions and the true classes. Specifically, the "ER-/HER2-" class boasts a precision of 0.95 and a recall of 0.96, exhibiting the highest balance between identifying true positives and minimizing false negatives. The "ER+/HER2- Low Prolif" class follows closely with a precision and recall of 0.94 and 0.96, respectively, also representing a very good performance.

In contrast, the "HER2+" class presents a precision of 0.96 but with a slight decrease in recall to 0.94. Although this is a minimal difference, it suggests a slight susceptibility of the model to false negatives in this class. However, it is worth noting that all these metrics remain high, which implies a very robust

overall performance.

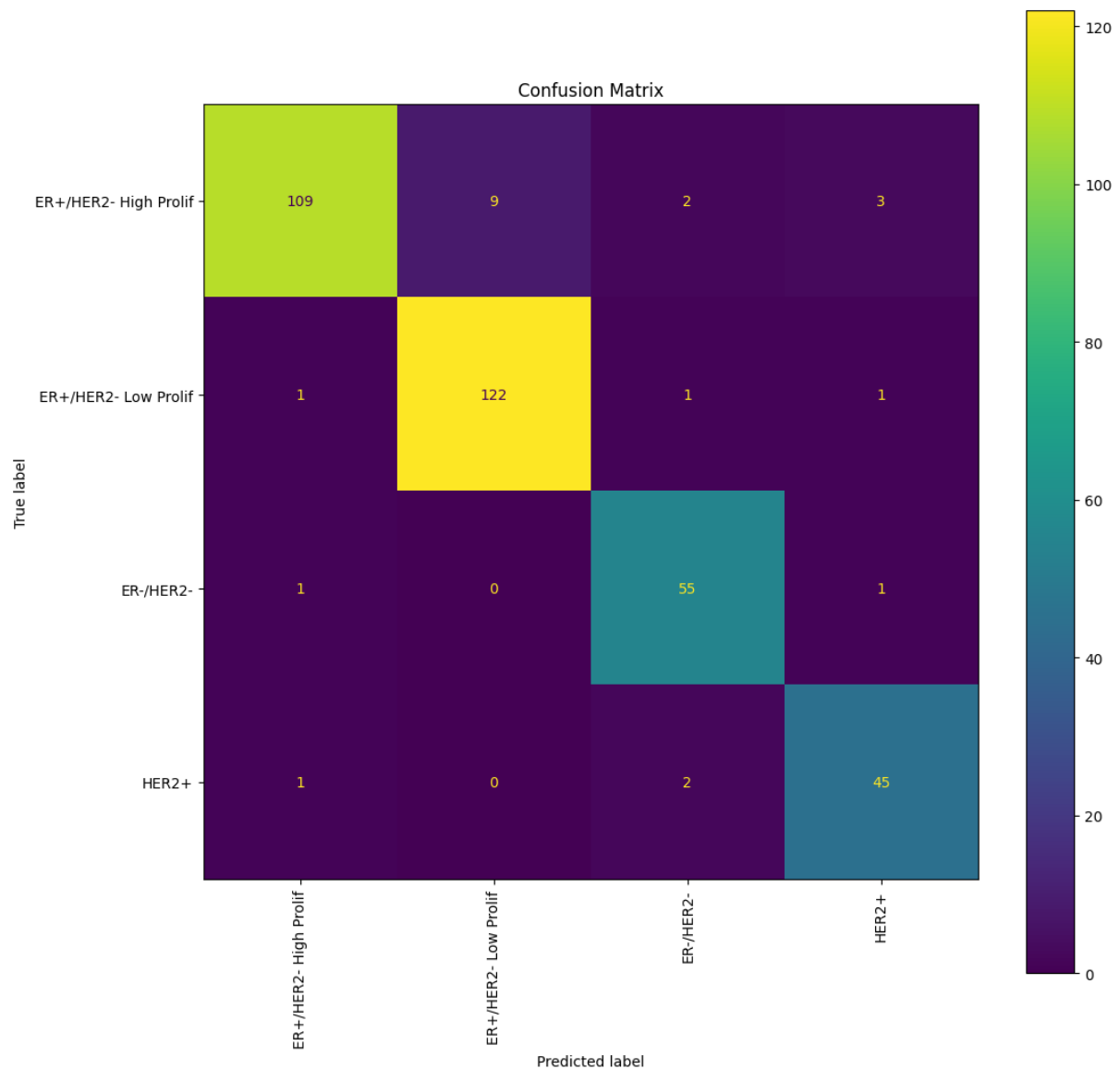


Figure 14: Confusion Matrix of Deep Neural Network (DNN) model. Vertical axis (True label): Represents the actual cancer subtypes for patients: "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif", "ER-/HER2-", and "HER2+". Horizontal axis (Predicted label): Represents the predicted cancer subtypes for patients: "ER+/HER2- High Prolif", "ER+/HER2- Low Prolif", "ER-/HER2-", and "HER2+". Numbers in diagonal cells (yellow and blue) represent correct predictions, while numbers outside this diagonal represent incorrect predictions.

Table 16: Score metrics of Deep Neural Network (DNN) model

	Precision	Recall	F1-score	Support
ER+/HER2- High Prolif	0.97	0.90	0.93	123
ER+/HER2- Low Prolif	0.94	0.96	0.95	125
ER-/HER2-	0.93	0.98	0.96	57
HER2+	0.94	0.98	0.96	48
Accuracy			0.95	353
Macro avg	0.94	0.96	0.95	353
Weighted avg	0.95	0.95	0.95	353

When observing the confusion matrix provided in figure 14 , it is noted that for the "ER+/HER2- High Prolif" subtype, 109 cases were correctly classified while 9 cases were incorrectly classified as "ER+/HER2- Low Prolif", 2 cases were incorrectly classified as "ER-/HER2-", and 3 cases were misassigned to the "HER2+" subtype. For the "ER+/HER2- Low Prolif" subtype, 122 cases were correctly categorized, but 1 case was misidentified as "ER+/HER2- High Prolif", and two other cases, one for "ER-/HER2-" and another for "HER2+", they were also incorrectly classified. Regarding the "ER-/HER2-" subtype, 55 samples were correctly classified and one case was incorrectly assigned to each of the other three subtypes. Finally, for the "HER2+" subtype, 45 cases were correctly identified, 1 case was incorrectly classified as "ER+/HER2- High Prolif", and 2 cases were incorrectly categorized as "ER-/HER2-". When evaluating the matrix as a whole, it becomes evident that the majority of samples were correctly classified,

When analyzing the metrics table 16 of the multi-layer neural network model for predicting breast cancer subtypes, it is seen that the precision, recall and f1-score values are generally high for all classes. The "ER+/HER2- High Prolif" subtype has a precision of 0.97, although its recall is slightly lower, recording 0.90, which results in an f1-score of 0.93. In contrast, the "ER+/HER2- Low Prolif" subtype exhibits a precision of 0.94 and a recall of 0.96, providing an f1-score of 0.95, slightly higher than the previous one. The "ER-/HER2-" and "HER2+" classes boast remarkable performance, with recalls of 0.98 and f1-scores of 0.96 for both. The accuracy for these two classes is 0.93 and 0.94, respectively. The model's overall precision, or accuracy, reaches an impressive value of 0.95 and the weighted average and macro are around the same value.

Overall, the DNN model performed exceptionally high in classifying subtypes, with an accuracy of 95%. DNN stands out as the most appropriate choice, since it was the best at capturing the biological relationships between the features, especially when dealing with complex data such as genetic expressions

and clinical information. This model not only achieved high precision but also presented the best balance between the critical metrics of precision and recall, with an average macro recall value of 0.96 [163, 164, 154, 165].

Recall is particularly important in the medical context, as it is more risky not to identify a positive case than to incorrectly identify a negative case as positive. Thus, the model that best balances precision and recall, possibly weighting more heavily toward recall, should be considered the most appropriate.

While ensemble methods like RF and AdaBoost did not fare as well, potential reasons include their intrinsic complexity and sensitivity to outliers. On the other hand, models like SVM and LR showcased high accuracy. However, their capabilities might not match the intricate complexities of medical data as efficiently as DNN. On the flip side, ANNs, which typically have a single layer, may not be as suited for such data. Despite their simplicity, ANNs have limitations in capturing complex non-linearities, compared to SVMs with non-linear kernels or DNNs, and may also be more sensitive to the presence of outliers [166]. However, SVM and LR exhibit advantages in specific scenarios due to their inherent simplicity, regularization features, and margin maximizing properties.

The prowess of DNN lies in its ability to decipher complex, non-linear relationships in data, crucial for medical applications. It optimizes multiple parameters, making it highly adaptable. Furthermore, its architectural depth allows for the recognition of high-level data features, vital for understanding intricate inter-relationships between variables [163]. The model's capacity to learn effective internal representations becomes invaluable for high-dimensional datasets like genetic expressions, simplifying data's complexity and aiding in classification. With noise and measurement error tolerance, coupled with a vast array of regularization techniques, the DNN proves robust against data variations, ensuring efficient performance on both training and unseen data [154, 165].

Given the criticality of recall in medical contexts, and the superior performance of DNN across all metrics, it stands out as the top choice for biomedical applications, underscoring the immense potential of ML in the medical domain.

In addition to these assessments, additional analyses were performed to find gene-to-gene correlations by cross-referencing the PAM50 gene list with two lists of 1500 and 1400 genes generated after the f-value feature selection (features that have been most implicated in cancer). The results of this analysis were saved in a two-text file. the resulting lists are named: list_cross1500 and list_cross1400. With the list of 1,500 characteristics crossed with the PAM50 list, only 2 genes were found in common, and have already been discussed in the previous section, despite the importance value of the 1,500 characteristics being very similar and all having low importance values. With the list of 1400 features belonging to the dataset

to predict the subtype, 32 genes in common were found, but here the range of feature importance values is from 158.2114 to 0.06139.

Furthermore, the significant correlation of 32 genes in common with the genes in the PAM50 panel, which is a reference in the field, strengthens the biological validity of the models. Below the 5 in common with the highest feature importance value. Among these genes, *KRT5* stands out, which codes for keratin 5, a protein involved in the structure and stability of epithelial cells. It is commonly used as a marker for basal-like breast cancer, a subtype that is typically more aggressive and has a worse prognosis [167]. Also *ESR1*, which is an estrogen receptor and plays a crucial role in modulating gene transcription and regulating the cell cycle. Its presence is generally a good indication for hormonal therapy in ER+ cancers [168]. Furthermore, *EGFR*, which is an epidermal growth factor receptor, is involved in several signaling pathways that control cell proliferation. It is frequently overexpressed in more aggressive breast cancer subtypes, such as HER2+ and basal-like [169]. *FOXC1* is associated with invasion and metastasis in breast cancer and is also considered a prognostic biomarker for the basal-like subtype [170]. Lastly, *BCL2* regulates apoptosis, a process of programmed cell death. Its imbalance can contribute to the survival of cancer cells [171].

However, there are genes that were not in the PAM50 panel and exhibited a higher feature importance. The lists with the name of all the genes in the feature selection are in the repository with the name: feature1500_list.txt and feature1400_list.txt .

Firstly, *ROPN1B*, that is still the subject of investigation, but initial studies suggest a possible involvement in relevant cellular processes [172]. Second, *CELSR2*, which is associated with cell adhesion, may be a factor in the migration of cancer cells [173]. Also *CSAD*, which is involved in the metabolism of the amino acid serine, has been linked to the growth of cancer cells [174]. *MPZL* plays a role in cell signaling pathways and may be related to cancer aggressiveness [175]. Finally, *MTFR2* plays a role in the formation of mitochondria, which can affect energy production in cancer cells [176].

Overall, it is possible to suggest that, while the models are highly effective in classifying cancer subtypes, corroborated by alignment with the PAM50 panel, they are less consistent in determining the presence or absence of the disease. This difference highlights the importance of balanced datasets and the need for further investigation of identified genes that are not in the panel as the genes were the same for both cases. Furthermore, by incorporating this analysis with existing biological knowledge about the genes involved, the research solidifies the emerging role of ML models in precision medicine, not only as diagnostic tools, but also as platforms for identifying biological pathways, relevant information and the development of targeted therapies.

Chapter 5

Conclusions and future work

This study aimed to develop an AI tool for the diagnosis and classification of breast cancer. Consequently, the effectiveness of several ML models to carry out this task were explored.

The study revealed that for classifying individuals with and without cancer, KNN demonstrated the best performance, particularly regarding recall, a crucial metric in medical contexts. DNN was the most suitable for classifying cancer subtypes, showing high accuracy and a remarkable balance between precision and recall. Of course, these results had the help of the f-value, which proved to be the best feature selection method compared to PCA and total.

One of the most significant challenges faced was the availability and quality of datasets. At the beginning of the research, it was difficult to find comprehensive datasets. The datasets that were available often were limited in size and lacked samples from individuals not diagnosed with cancer. This problem was exacerbated by the bureaucracy involved with acquiring medical data and the lack of standardized data formats. Inequality in sample sizes between classes and the sensitivity of some models to these issues, such as ANN, were also significant challenges.

To advance this field, it is crucial to work on expanding the dataset samples, both for cancer and non-cancer cases. It was also important to obtain clinical data from individuals without cancer. Initiating efforts to create a more accessible and standardized dataset would also be valuable. Furthermore, future research could focus on optimizing AdaBoost and DNN models to improve their metrics, especially recall, and on approaches to deal with class imbalance in a medical context. With more samples it is possible to improve the performance of the models as we will have more data for training and testing them.

The tool developed in this study proved to be effective and easy to use, providing methods for pre-processing, feature selection and hyperparameter optimization. In the future, it would be interesting to add more alternatives such as more normalization methods and grid search. Hyperparameter optimization was performed using Random Search due to its efficiency in exploring the parameter space in a more random and less costly way, although the use of Grid Search can be considered for future research for a more

exhaustive and accurate search.

This study contributed to the growing body of research in medical applications of ML by providing valuable insights into the relative effectiveness of different algorithms in critical medical tasks. The models developed in this study demonstrated promising metrics, making them viable candidates to assist physicians and hospitals with personalized medicine and challenging breast cancer diagnosis. In the future, it would also be interesting to explore the potential of neural networks in more depth. RNNs and CNNs were not used in this study due to the nature of the available data, which were gene expression data and not time series or images.

In summary, this study represents a significant step forward in applying ML to the field of oncology. With future adjustments and expansions, there is great potential for significant advances in this critical area of medicine. Finally, the high degree of importance attributed to genes that were not previously identified in the PAM50 panel suggests new avenues of research and possible targets for future therapies. The data provide a solid foundation for subsequent investigations and deeper understanding of the biology underlying breast cancer types and their respective subtypes.

Bibliography

- [1] Francis S. Collins, Michael Morgan, and Aristides Patrinos. The Human Genome Project: lessons from large-scale biology. *Science (New York, N.Y.)*, 300(5617):286–290, April 2003.
- [2] M V Olson. The human genome project. *Proceedings of the National Academy of Sciences*, 90(10):4338–4344, May 1993. Publisher: Proceedings of the National Academy of Sciences.
- [3] Yehudit Hasin, Marcus Seldin, and Aldons Lusi. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, May 2017.
- [4] Ebru Aydınoğlu Bayrak, Pınar Kırıcı, and Tolga Ensari. Comparison of Machine Learning Methods for Breast Cancer Diagnosis. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pages 1–3, April 2019.
- [5] Rakesh Kumar C. Rama Krishna, Maitreyee Dutta. *Proceedings of 2nd International Conference on Communication, Computing and Networking*. 2018.
- [6] Pankaj Kumar and Rupali Aggarwal. An overview of triple-negative breast cancer. *Archives of Gynecology and Obstetrics*, 293(2):247–269, February 2016.
- [7] Carey K. Anders, Cheng Fan, Joel S. Parker, Lisa A. Carey, Kimberly L. Blackwell, Nancy Klauber-DeMore, and Charles M. Perou. Breast Carcinomas Arising at a Young Age: Unique Biology or a Surrogate for Aggressive Intrinsic Subtypes? *Journal of Clinical Oncology*, 29(1):e18–e20, January 2011.
- [8] Rebecca Dent, Maureen Trudeau, Kathleen I. Pritchard, Wedad M. Hanna, Harriet K. Kahn, Carol A. Sawka, Lavina A. Lickley, Ellen Rawlinson, Ping Sun, and Steven A. Narod. Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence. *Clinical Cancer Research*, 13(15):4429–4434, August 2007.

- [9] Osama Hamzeh, Abedalrhman Alkhateeb, Julia Zheng, Srinath Kandalam, and Luis Rueda. Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data. *BMC Bioinformatics*, 21(2):78, March 2020.
- [10] Simona Serrati, Simona De Summa, Brunella Pilato, Daniela Petriella, Rosanna Lacalamita, Stefania Tommasi, and Rosamaria Pinto. Next-generation sequencing: advances and applications in cancer diagnosis. *OncoTargets and therapy*, 9:7355–7365, December 2016.
- [11] Khong-Loon Tiong, Kuo-Ching Chang, Kun-Tu Yeh, Ting-Yuan Liu, Jia-Hong Wu, Ping-Heng Hsieh, Shu-Hui Lin, Wei-Yun Lai, Yu-Chin Hsu, Jeou-Yuan Chen, Jan-Gowth Chang, and Grace S. Shieh. CSNK1E/CTNNB1 Are Synthetic Lethal To TP53 in Colorectal Cancer and Are Markers for Prognosis. *Neoplasia*, 16(5):441–450, May 2014.
- [12] Nadia Harbeck, Frédérique Penault-Llorca, Javier Cortes, Michael Gnant, Nehmat Houssami, Philip Poortmans, Kathryn Ruddy, Janice Tsang, and Fatima Cardoso. Breast cancer. *Nature Reviews Disease Primers*, 5(1):1–31, September 2019. Number: 1 Publisher: Nature Publishing Group.
- [13] Do-Youn Oh and Yung-Jue Bang. HER2-targeted therapies — a role beyond breast cancer. *Nature Reviews Clinical Oncology*, 17(1):33–48, January 2020. Number: 1 Publisher: Nature Publishing Group.
- [14] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21660>.
- [15] L. Irwig, P. Macaskill, and N. Houssami. Evidence relevant to the investigation of breast symptoms: the triple test. *The Breast*, 11(3):215–220, June 2002.
- [16] Bingbing Xie, Zifeng Yuan, Yadong Yang, Zhidan Sun, Shuigeng Zhou, and Xiangdong Fang. MOBCdb: a comprehensive database integrating multi-omics data on breast cancer for precision medicine. *Breast Cancer Research and Treatment*, 169(3):625–632, June 2018.
- [17] J. Larry Jameson and Dan L. Longo. Precision medicine—personalized, problematic, and promising. *The New England Journal of Medicine*, 372(23):2229–2234, June 2015.

- [18] Yu-Chiao Chiu, Hung-I Harry Chen, Tinghe Zhang, Songyao Zhang, Aparna Gorthi, Li-Ju Wang, Yufei Huang, and Yidong Chen. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Medical Genomics*, 12(1):18, January 2019.
- [19] Stephan C. Schuster. Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1):16–18, January 2008.
- [20] Sam Behjati and Patrick S Tarpey. What is next generation sequencing? *Archives of Disease in Childhood - Education and Practice*, 98(6):236–238, 2013.
- [21] Gavin R Oliver, Steven N Hart, and Eric W Klee. Bioinformatics for Clinical Next Generation Sequencing. *Clinical Chemistry*, 61(1):124–135, January 2015.
- [22] Ioannis Michalopoulos Athanasia Pavlopoulou, Demetrios A. Spandidos. Human cancer databases (Review), 2014.
- [23] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A. Smith, I. Richard Thompson, Sridhar Ramaswamy, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Cyril Benes, Ultan McDermott, and Mathew J. Garnett. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, January 2013.
- [24] Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. GenBank. *Nucleic Acids Research*, 41(D1):D36–D42, January 2013.
- [25] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lusk, R. Mani, T. Rayner, A. Sharma, E. Williams, U. Sarkans, and A. Brazma. ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research*, 35(suppl_1):D747–D750, January 2007.
- [26] Tanya Barrett and Ron Edgar. Mining Microarray Data at NCBI's Gene Expression Omnibus (GEO)*. In Minou Bina, editor, *Gene Mapping, Discovery, and Expression: Methods and Protocols*, Methods in Molecular Biology, pages 175–190. Humana Press, Totowa, NJ, 2006.
- [27] Anja Persson, Sophia Hober, and Mathias Uhlen. A human protein atlas based on antibody proteomics. *Current opinion in molecular therapeutics*, 8:185–90, July 2006.

- [28] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, January 2019.
- [29] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. PubChem Substance and Compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, January 2016.
- [30] Egon L. Willighagen, Andra Waagmeester, Ola Spjuth, Peter Ansell, Antony J. Williams, Valery Tkachenko, Janna Hastings, Bin Chen, and David J. Wild. The ChEMBL database as linked open data. *Journal of Cheminformatics*, 5(1):23, May 2013.
- [31] Lisa Dwane, Fiona M Behan, Emanuel Gonçalves, Howard Lightfoot, Wanjuan Yang, Dieudonne van der Meer, Rebecca Shepherd, Miguel Pignatelli, Francesco Iorio, and Mathew J Garnett. Project Score database: a resource for investigating cancer cell dependencies and prioritizing therapeutic targets. *Nucleic Acids Research*, 49(D1):D1365–D1372, January 2021.
- [32] Yann Joly, Edward S. Dove, Bartha M. Knoppers, Martin Bobrow, and Don Chalmers. Data Sharing in the Post-Genomic World: The Experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). *PLOS Computational Biology*, 8(7):e1002549, July 2012. Publisher: Public Library of Science.
- [33] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P. A. Futreal, M. R. Stratton, and R. Wooster. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer*, 91(2):355–358, July 2004. Number: 2 Publisher: Nature Publishing Group.
- [34] Eleonora Cappelli, Fabio Cumbo, Anna Bernasconi, Arif Canakoglu, Stefano Ceri, Marco Masseroli, and Emanuel Weitschek. OpenGDC: Unifying, Modeling, Integrating Cancer Genomic Data and Clinical Metadata. *Applied Sciences*, 10(18):6367, January 2020. Number: 18 Publisher: Multidisciplinary Digital Publishing Institute.
- [35] Bernard Pereira, Suet-Feung Chin, Oscar M. Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A. Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen-John Sammut, Dana W. Y. Tsui, Bin Liu, Sarah-Jane Dawson, Jean Abraham, Helen Northen, John F. Peden, Abhik Mukherjee, Gulisa Turashvili, Andrew R. Green, Steve McKinney, Arusha Oloumi, Sohrab Shah, Nitzan

- Rosenfeld, Leigh Murphy, David R. Bentley, Ian O. Ellis, Arnie Purushotham, Sarah E. Pinder, Anne-Lise Børresen-Dale, Helena M. Earl, Paul D. Pharoah, Mark T. Ross, Samuel Aparicio, and Carlos Caldas. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications*, 7(1):11479, May 2016. Number: 1 Publisher: Nature Publishing Group.
- [36] Mallory Ann Freeberg, Lauren A Fromont, Teresa D’Altri, Anna Foix Romero, Jorge Izquierdo Ciges, Aina Jene, Giselle Kerry, Mauricio Moldes, Roberto Ariosa, Silvia Bahena, Daniel Barrowdale, Marcos Casado Barbero, Dietmar Fernandez-Orth, Carles Garcia-Linares, Emilio Garcia-Rios, Frédéric Haziza, Bela Juhasz, Oscar Martinez Llobet, Gemma Milla, Anand Mohan, Manuel Rueda, Aravind Sankar, Dona Shaju, Ashutosh Shimpi, Babita Singh, Coline Thomas, Sabela de la Torre, Umuthan Uyan, Claudia Vasallo, Paul Flicek, Roderic Guigo, Arcadi Navarro, Helen Parkinson, Thomas Keane, and Jordi Rambla. The European Genome-phenome Archive in 2021. *Nucleic Acids Research*, 50(D1):D980–D987, January 2022.
- [37] Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, June 2012. Number: 7403 Publisher: Nature Publishing Group.
- [38] Ala’a El-Nabawy, Nashwa El-Bendary, and Nahla A. Belal. A feature-fusion framework of clinical, genomics, and histopathological data for METABRIC breast cancer subtype classification. *Applied Soft Computing*, 91:106238, June 2020.
- [39] Patrick Henry Winston. *Artificial intelligence* (3rd ed.), 1992. Archive Location: world.
- [40] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120, December 2019. Publisher: American Association for the Advancement of Science.
- [41] Earl B. Hunt. *Artificial Intelligence*. Academic Press, May 2014. Google-Books-ID: 9y2jBQAAQBAJ.

- [42] Pavel Hamet and Johanne Tremblay. Artificial intelligence in medicine. *Metabolism*, 69:S36–S40, April 2017.
- [43] Batta Mahesh. *Machine Learning Algorithms -A Review*. January 2019.
- [44] Issam El Naqa and Martin J. Murphy. What Is Machine Learning? In Issam El Naqa, Ruijiang Li, and Martin J. Murphy, editors, *Machine Learning in Radiation Oncology: Theory and Applications*, pages 3–11. Springer International Publishing, Cham, 2015.
- [45] Pramila P. Shinde and Seema Shah. A Review of Machine Learning and Deep Learning Applications. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–6, August 2018.
- [46] Mohamed Alloghani, Dhiya Al-Jumeily, Jamila Mustafina, Abir Hussain, and Ahmed J. Aljaaf. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In Michael W. Berry, Azlinah Mohamed, and Bee Wah Yap, editors, *Supervised and Unsupervised Learning for Data Science*, Unsupervised and Semi-Supervised Learning, pages 3–21. Springer International Publishing, Cham, 2020.
- [47] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4:237–285, May 1996.
- [48] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, November 2020.
- [49] Nannicini Samulowitz Diaz, Fokoue-Nkoutche. An effective algorithm for hyperparameter optimization of neural networks | IBM Journals & Magazine | IEEE Xplore, 2019.
- [50] Noemí DeCastro-García, Ángel Luis Muñoz Castañeda, David Escudero García, and Miguel V. Carriegos. Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a Machine Learning Algorithm. *Complexity*, 2019:e6278908, February 2019. Publisher: Hindawi.
- [51] Sheena Angra and Sachin Ahuja. Machine learning and its applications: A review. In *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, pages 57–60, March 2017.

- [52] Mohssen Mohammed, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. *Machine Learning Algorithms and Applications: Algorithms and Applications*. CRC Press, Boca Raton, August 2016.
- [53] Amanpreet Singh, Narina Thakur, and Aakanksha Sharma. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315, March 2016.
- [54] Vladimir Nasteski. An overview of the supervised machine learning methods. *HORIZONS.B*, 4:51–62, December 2017.
- [55] Tammy Jiang, Jaimie L. Gradus, and Anthony J. Rosellini. Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, 51(5):675–687, September 2020.
- [56] Chih-Fong Tsai and Ya-Ting Sung. Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches. *Knowledge-Based Systems*, 203:106097, September 2020.
- [57] Dejun Zhang, Lu Zou, Xionghui Zhou, and Fazhi He. Integrating Feature Selection and Feature Extraction Methods With Deep Learning to Predict Clinical Outcome of Breast Cancer. *IEEE Access*, 6:28936–28944, 2018. Conference Name: IEEE Access.
- [58] Huijuan Lu, Junying Chen, Ke Yan, Qun Jin, Yu Xue, and Zhigang Gao. A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing*, 256:56–62, September 2017.
- [59] Computational Methods of Feature Selection - Google Livros.
- [60] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, 2001.
- [61] An Introduction to Variable and Feature Selection.
- [62] A new feature selection method on classification of medical datasets: Kernel F-score feature selection - ScienceDirect.
- [63] Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97(457):77–87, March 2002. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1198/016214502753479248>.

- [64] Pradip Dhal and Chandrashekhar Azad. A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, 52(4):4543–4581, March 2022.
- [65] D Lavanya. ANALYSIS OF FEATURE SELECTION WITH CLASSIFICATION: BREAST CANCER DATASETS. 2(5), 2011.
- [66] Shokoufeh Aalaei, Hadi Shahraki, Alireza Rowhanimanesh, and Saeid Eslami. Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. *Iranian Journal of Basic Medical Sciences*, 19(5):476–482, May 2016.
- [67] Suhang Wang, Jiliang Tang, and Huan Liu. Feature Selection. pages 1–9. January 2016.
- [68] Jason Brownlee. How to Choose a Feature Selection Method For Machine Learning. 2020.
- [69] Christopher J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, June 1998.
- [70] Rustam Nadira. Classification of cancer data using support vector machines with features selection method based on global artificial bee colony: AIP Conference Proceedings: Vol 2023, No 1, 2018.
- [71] Ying Liu. Active Learning with Support Vector Machine Applied to Gene Expression Data for Cancer Classification. *Journal of Chemical Information and Computer Sciences*, 44(6):1936–1941, November 2004. Publisher: American Chemical Society.
- [72] Mehmet Fatih Akay. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36(2, Part 2):3240–3247, March 2009.
- [73] Antonio Mucherino, Petraq J. Papajorgji, and Panos M. Pardalos. k-Nearest Neighbor Classification. In Antonio Mucherino, Petraq J. Papajorgji, and Panos M. Pardalos, editors, *Data Mining in Agriculture*, Springer Optimization and Its Applications, pages 83–106. Springer, New York, NY, 2009.
- [74] Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, 4(11):218, June 2016.
- [75] Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, and Md. Kamrul Hasan. Prediction of breast cancer using support vector machine and K-Nearest neighbors. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 226–229, December 2017. ISSN: 2572-7621.

- [76] Qingbo Li, Wenjie Li, Jialin Zhang, and Zhi Xu. An improved k-nearest neighbour method to diagnose breast cancer. *Analyst*, 143(12):2807–2811, June 2018. Publisher: The Royal Society of Chemistry.
- [77] Shler Farhad Khorshid and Adnan Mohsin Abdulazeez. BREAST CANCER DIAGNOSIS BASED ON K-NEAREST NEIGHBORS: A REVIEW. *PalArch's Journal of Archaeology of Egypt / Egyptology*, 18(4):1927–1951, February 2021. Number: 4.
- [78] Thomas M. H. Hope. Chapter 4 - Linear regression. In Andrea Mechelli and Sandra Vieira, editors, *Machine Learning*, pages 67–81. Academic Press, January 2020.
- [79] Dastan Maulud and Adnan M Abdulazeez. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1(4):140–147, 2020.
- [80] S. Murugan, B. Muthu Kumar, and S. Amudha. Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest. In *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, pages 763–766, September 2017.
- [81] Chandrasegar Thirumalai and Rashad Manzoor. Cost optimization using normal linear regression method for breast cancer Type I skin. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, volume 2, pages 264–268, April 2017.
- [82] Mohammad Ali Mansournia, Angelika Geroldinger, Sander Greenland, and Georg Heinze. Separation in Logistic Regression: Causes, Consequences, and Control. *American Journal of Epidemiology*, 187(4):864–870, April 2018.
- [83] Connelly. Logistic Regression - ProQuest, 2020.
- [84] Anthony Anene Ude. *Classification of Breast Cancer Using Logistic Regression*. Thesis, June 2019. Accepted: 2019-08-08T13:40:58Z.
- [85] Ahmed F. Seddik and Doaa M. Shawky. Logistic regression model for breast cancer automatic diagnosis. In *2015 SAI Intelligent Systems Conference (IntelliSys)*, pages 150–154, November 2015.
- [86] Hsiao-Lin Hwa, Wen-Hong Kuo, Li-Yun Chang, Ming-Yang Wang, Tao-Hsin Tung, King-Jen Chang, and Fon-Jou Hsieh. Prediction of breast cancer and lymph node metastatic status with tumour markers

- using logistic regression models. *Journal of Evaluation in Clinical Practice*, 14(2):275–280, 2008. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2753.2007.00849.x>.
- [87] S. B. Kotsiantis. Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4):261–283, April 2013.
- [88] Anthony J. Myles, Robert N. Feudale, Yang Liu, Nathaniel A. Woody, and Steven D. Brown. An introduction to decision tree modeling. *Journal of Chemometrics*, 18(6):275–285, 2004. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/cem.873>.
- [89] Wei-Yin Loh. Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, 1(1):14–23, 2011. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.8>.
- [90] Wei-Yin Loh. Fifty Years of Classification and Regression Trees. *International Statistical Review*, 82(3):329–348, 2014. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12016>.
- [91] Ronak Sumbaly, N. Vishnusri, and s Jeyalatha. Diagnosis of Breast Cancer using Decision Tree Data Mining Technique. *International Journal of Computer Applications*, 98:16–24, July 2014.
- [92] P. Sathiyarayanan, S Pavithra., M SAI SARANYA., and M Makeswari. Identification of Breast Cancer Using The Decision Tree Algorithm. In *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, pages 1–6, March 2019.
- [93] Gajendra K. Vishwakarma, Pragya Kumari, and Atanu Bhattacharjee. Thresholding of prominent biomarkers of breast cancer on overall survival using classification and regression tree. *Cancer Biomarkers*, 34(2):319–328, January 2022. Publisher: IOS Press.
- [94] Mingzhe Chen, Ursula Challita, Walid Saad, Changchuan Yin, and Mérouane Debbah. Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial. *IEEE Communications Surveys & Tutorials*, 21(4):3039–3071, 2019. Conference Name: IEEE Communications Surveys & Tutorials.
- [95] Jinming Zou, Yi Han, and Sung-Sau So. Overview of Artificial Neural Networks. In David J. Livingstone, editor, *Artificial Neural Networks: Methods and Applications*, Methods in Molecular Biology™, pages 14–22. Humana Press, Totowa, NJ, 2009.
- [96] Steven Walczak. Artificial Neural Networks, 2019. ISBN: 9781522573685 Pages: 40-53 Publisher: IGI Global.

- [97] Krishna Mridha. Early Prediction of Breast Cancer by using Artificial Neural Network and Machine Learning Techniques. In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, pages 582–587, June 2021. ISSN: 2329-7182.
- [98] Filippo Amato, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara, Aleš Hampl, and Josef Havel. Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, 11(2):47–58, January 2013.
- [99] Robi Polikar. Ensemble Learning. In Cha Zhang and Yunqian Ma, editors, *Ensemble Machine Learning: Methods and Applications*, pages 1–34. Springer US, Boston, MA, 2012.
- [100] Peter Bühlmann. Bagging, Boosting and Ensemble Methods. In James E. Gentle, Wolfgang Karl Härdle, and Yuichi Mori, editors, *Handbook of Computational Statistics: Concepts and Methods*, Springer Handbooks of Computational Statistics, pages 985–1022. Springer, Berlin, Heidelberg, 2012.
- [101] Aakash Parmar, Rakesh Katariya, and Vatsal Patel. A Review on Random Forest: An Ensemble Classifier. In Jude Hemanth, Xavier Fernando, Pavel Lafata, and Zubair Baig, editors, *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, Lecture Notes on Data Engineering and Communications Technologies, pages 758–763, Cham, 2019. Springer International Publishing.
- [102] Mohamed Hosni, Ibtissam Abnane, Ali Idri, Juan M. Carrillo de Gea, and José Luis Fernández Alemán. Reviewing ensemble classification methods in breast cancer. *Computer Methods and Programs in Biomedicine*, 177:89–112, August 2019.
- [103] Moloud Abdar, Mariam Zomorodi-Moghadam, Xujuan Zhou, Raj Gururajan, Xiaohui Tao, Prabal D Barua, and Rashmi Gururajan. A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern Recognition Letters*, 132:123–131, April 2020.
- [104] Mohammad M. Ghiasi and Sohrab Zendeboudi. Application of decision tree-based ensemble learning in the classification of breast cancer. *Computers in Biology and Medicine*, 128:104089, January 2021.
- [105] Kamal Kant Hiran, Ritesh Kumar Jain, Dr Kamlesh Lakhwani, and Dr Ruchi Doshi. *Machine Learning: Master Supervised and Unsupervised Learning Algorithms with Real Examples (English Edition)*. BPB Publications, September 2021. Google-Books-ID: 4WVDEAAAQBAJ.

- [106] Douglas H. Fisher, Michael J. Pazzani, and Pat Langley. *Concept Formation: Knowledge and Experience in Unsupervised Learning*. Morgan Kaufmann, May 2014. Google-Books-ID: ATKnCQAAQBAJ.
- [107] Zoubin Ghahramani. Unsupervised Learning. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, Lecture Notes in Computer Science, pages 72–112. Springer, Berlin, Heidelberg, 2004.
- [108] Muhammad Usama, Junaid Qadir, Aunn Raza, Hunain Arif, Kok-lim Alvin Yau, Yehia Elkhatib, Amir Hussain, and Ala Al-Fuqaha. Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges. *IEEE Access*, 7:65579–65615, 2019. Conference Name: IEEE Access.
- [109] Jason Xu and Kenneth Lange. Power k-Means Clustering. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6921–6931. PMLR, May 2019. ISSN: 2640-3498.
- [110] Melissa Zhao, Yushi Tang, Hyunkyung Kim, and Kohei Hasegawa. Machine Learning With K-Means Dimensional Reduction for Predicting Survival Outcomes in Patients With Breast Cancer. *Cancer Informatics*, 17:1176935118810215, January 2018. Publisher: SAGE Publications Ltd STM.
- [111] Ade Jamal, Annisa Handayani, Ali Septiandri, Endang Ripmiatin, and Yunus Effendi. Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, page 192, December 2018.
- [112] Wenbo Zhu, Zachary T. Webb, Kaitian Mao, and José Romagnoli. A Deep Learning Approach for Process Data Visualization Using t-Distributed Stochastic Neighbor Embedding. *Industrial & Engineering Chemistry Research*, 58(22):9564–9575, June 2019. Publisher: American Chemical Society.
- [113] Nonita Sharma, Monika Mangla, Sachi Nandan Mohanty, and Suneeta Satpaty. A Stochastic Neighbor Embedding Approach for Cancer Prediction. In *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 599–603, March 2021.
- [114] Nonita Sharma, K. P. Sharma, Monika Mangla, and Rajneesh Rani. Breast cancer classification using snapshot ensemble deep learning model and t-distributed stochastic neighbor embedding. *Multimedia Tools and Applications*, 82(3):4011–4029, January 2023.

- [115] Ahmed Lasisi and Nii Attoh-Okine. Principal components analysis and track quality index: A machine learning approach. *Transportation Research Part C: Emerging Technologies*, 91:230–248, June 2018.
- [116] Tom Howley, Michael G. Madden, Marie-Louise O’Connell, and Alan G. Ryder. The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data. In Ann Macintosh, Richard Ellis, and Tony Allen, editors, *Applications and Innovations in Intelligent Systems XIII*, pages 209–222, London, 2006. Springer.
- [117] Hasmarina Hasan and Nooritawati Md Tahir. Feature selection of breast cancer based on Principal Component Analysis. In *2010 6th International Colloquium on Signal Processing & its Applications*, pages 1–4, May 2010.
- [118] Sarthak Sanjay Tilwankar and Bhupendra Singh Kirar. Breast Cancer Detection using Principal Component Analysis and Machine Learning Models. In *2021 First International Conference on Advances in Computing and Future Communication Technologies (ICACFCT)*, pages 80–84, December 2021.
- [119] Moses Charikar, Vaggos Chatziafratis, Rad Niazadeh, and Grigory Yaroslavltssev. Hierarchical Clustering for Euclidean Data. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 2721–2730. PMLR, April 2019. ISSN: 2640-3498.
- [120] Juan Nunez-Iglesias, Ryan Kennedy, Toufiq Parag, Jianbo Shi, and Dmitri B. Chklovskii. Machine Learning of Hierarchical Clustering to Segment 2D and 3D Images. *PLOS ONE*, 8(8):e71715, August 2013. Publisher: Public Library of Science.
- [121] Arjun P. Athreya, Alan J. Gaglio, Junmei Cairns, Krishna R. Kalari, Richard M. Weinshilboum, Liewei Wang, Zbigniew T. Kalbarczyk, and Ravishankar K. Iyer. Machine Learning Helps Identify New Drug Mechanisms in Triple-Negative Breast Cancer. *IEEE Transactions on NanoBioscience*, 17(3):251–259, July 2018. Conference Name: IEEE Transactions on NanoBioscience.
- [122] Alexander J. Titus, Carly A. Bobak, and Brock C. Christensen. A New Dimension of Breast Cancer Epigenetics - Applications of Variational Autoencoders with DNA Methylation:. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 140–145, Funchal, Madeira, Portugal, 2018. SCITEPRESS - Science and Technology Publications.
- [123] Huang Yi, Sun Shiyu, Duan Xiusheng, and Chen Zhigang. A study on Deep Neural Networks frame-

- work. In *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, pages 1519–1522, October 2016.
- [124] Shizhao Sun, Wei Chen, Liwei Wang, Xiaoguang Liu, and Tie-Yan Liu. On the Depth of Deep Neural Networks: A Theoretical View. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), March 2016. Number: 1.
- [125] Tao Luo, Zheng Ma, Zhi-Qin John Xu, and Yaoyu Zhang. Theory of the Frequency Principle for General Deep Neural Networks, July 2019. arXiv:1906.09235 [cs, math, stat].
- [126] Li Deng, Geoffrey Hinton, and Brian Kingsbury. New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8599–8603, May 2013. ISSN: 2379-190X.
- [127] Nivaashini Mathappan, R.s. Soundariya, Aravindhraj Natarajan, and Sathish Kumar Gopalan. Bio-medical analysis of breast cancer risk detection based on deep neural network. *International Journal of Medical Engineering and Informatics*, 12(6):529–541, January 2020. Publisher: Inderscience Publishers.
- [128] K Vijayakumar, Vinod J Kadam, and Sudhir Kumar Sharma. Breast cancer diagnosis using multiple activation deep neural network. *Concurrent Engineering*, 29(3):275–284, September 2021. Publisher: SAGE Publications Ltd STM.
- [129] S. Karthik, R. Srinivasa Perumal, and P. V. S. S. R. Chandra Mouli. Breast Cancer Classification Using Deep Neural Networks. In S. Margret Anuncia and Uffe Kock Wiil, editors, *Knowledge Computing and Its Applications: Knowledge Manipulation and Processing Techniques: Volume 1*, pages 227–241. Springer, Singapore, 2018.
- [130] Keiron O'Shea and Ryan Nash. An Introduction to Convolutional Neural Networks, December 2015. arXiv:1511.08458 [cs].
- [131] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6, August 2017.
- [132] Madhusmita Sahu and Rasmita Dash. A Survey on Deep Learning: Convolution Neural Network (CNN). In Debahuti Mishra, Rajkumar Buyya, Prasant Mohapatra, and Srikanta Patnaik, editors,

- Intelligent and Cloud Computing*, Smart Innovation, Systems and Technologies, pages 317–325, Singapore, 2021. Springer.
- [133] Maleika Heenaye-Mamode Khan, Nazmeen Boodoo-Jahangeer, Wasiimah Dullull, Shaista Nathire, Xiaohong Gao, G. R. Sinha, and Kapil Kumar Nagwanshi. Multi- class classification of breast cancer abnormalities using Deep Convolutional Neural Network (CNN). *PLOS ONE*, 16(8):e0256500, August 2021. Publisher: Public Library of Science.
 - [134] Saad Awadh Alanazi, M. M. Kamruzzaman, Md Nazirul Islam Sarker, Madallah Alruwaili, Yousef Alhwaiti, Nasser Alshammari, and Muhammad Hameed Siddiqi. Boosting Breast Cancer Detection Using Convolutional Neural Network. *Journal of Healthcare Engineering*, 2021:e5528622, April 2021. Publisher: Hindawi.
 - [135] Satya Prakash Yadav, Subiya Zaidi, Annu Mishra, and Vibhash Yadav. Survey on Machine Learning in Speech Emotion Recognition and Vision Systems Using a Recurrent Neural Network (RNN). *Archives of Computational Methods in Engineering*, 29(3):1753–1770, May 2022.
 - [136] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, 31(7):1235–1270, July 2019.
 - [137] Chuanlong Yin, Yuefei Zhu, Jinlong Fei, and Xinzheng He. A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks. *IEEE Access*, 5:21954–21961, 2017. Conference Name: IEEE Access.
 - [138] Rui Yan, Fei Ren, Zihao Wang, Lihua Wang, Yubo Ren, Yudong Liu, Xiaosong Rao, Chunhou Zheng, and Fa Zhang. A Hybrid Convolutional and Recurrent Deep Neural Network for Breast Cancer Pathological Image Classification. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 957–962, December 2018.
 - [139] Dehua Chen, Guangjun Qian, Cheng Shi, and Qiao Pan. Breast Cancer Malignancy Prediction Using Incremental Combination of Multiple Recurrent Neural Networks. In Derong Liu, Shengli Xie, Yuanqing Li, Dongbin Zhao, and El-Sayed M. El-Alfy, editors, *Neural Information Processing*, Lecture Notes in Computer Science, pages 43–52, Cham, 2017. Springer International Publishing.
 - [140] Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, Yaman Afadar, and Omar Elgendy. Breast cancer detection using artificial intelligence techniques: A systematic literature review. *Artificial Intelligence in Medicine*, 127:102276, May 2022.

- [141] Shadman Sakib, Nowrin Yasmin, Abyaz Kader Tanzeem, Fatema Shorna, Khan Md. Hasib, and Saadia Binte Alam. Breast Cancer Detection and Classification: A Comparative Analysis Using Machine Learning Algorithms. In V. Bindhu, João Manuel R. S. Tavares, and Ke-Lin Du, editors, *Proceedings of Third International Conference on Communication, Computing and Electronics Systems*, Lecture Notes in Electrical Engineering, pages 703–717, Singapore, 2022. Springer.
- [142] Waguih ElMaraghy Luis Rueda Alioune Ngom Ashraf Abou Tabl, Abedalrhman Alkhateeb. Frontiers | A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer, 2022.
- [143] Benzheng Wei, Zhongyi Han, Xueying He, and Yilong Yin. Deep learning model based breast cancer histopathological image classification. In *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 348–353, April 2017.
- [144] Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports*, 9(1):12495, August 2019. Number: 1 Publisher: Nature Publishing Group.
- [145] Luuk Balkenende, Jonas Teuwen, and Ritse M. Mann. Application of Deep Learning in Breast Cancer Imaging. *Seminars in Nuclear Medicine*, 52(5):584–596, September 2022.
- [146] Gunjan Chugh, Shailender Kumar, and Nanhay Singh. Survey on Machine Learning and Deep Learning Applications in Breast Cancer Diagnosis. *Cognitive Computation*, 13(6):1451–1470, November 2021.
- [147] Vikas Chaurasia and Saurabh Pal. Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer. *SN Computer Science*, 1(5):270, August 2020.
- [148] M. Sarkar and T. Y. Leong. Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. *Proceedings of the AMIA Symposium*, pages 759–763, 2000.
- [149] Anusha Bharat, N Pooja, and R Anishka Reddy. Using Machine Learning algorithms for breast cancer risk prediction and diagnosis. In *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)*, pages 1–4, October 2018. ISSN: 2381-4128.
- [150] Sarah M. Ayyad, Ahmed I. Saleh, and Labib M. Labib. Gene expression cancer classification using modified K-Nearest Neighbors technique. *Biosystems*, 176:41–51, February 2019.

- [151] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, and Lei Hua. Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *Journal of Medical Systems*, 36(4):2431–2448, August 2012.
- [152] D. Hammerstrom. Working with neural networks. *IEEE Spectrum*, 30(7):46–53, July 1993. Conference Name: IEEE Spectrum.
- [153] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The Limitations of Deep Learning in Adversarial Settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, March 2016.
- [154] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, November 2018.
- [155] Kristina Magnusson, Gabriela Gremel, Lisa Rydén, Victor Pontén, Mathias Uhlén, Anna Dimberg, Karin Jirstrom, and Fredrik Pontén. ANLN is a prognostic biomarker independent of Ki-67 and essential for cell cycle progression in primary breast cancer. *BMC Cancer*, 16(1):904, November 2016.
- [156] Shanmei Lv, Wenjie Xu, Yanyan Zhang, Jiling Zhang, and Xuejun Dong. NUF2 as an anticancer therapeutic target and prognostic factor in breast cancer. *International Journal of Oncology*, 57(6):1358–1367, December 2020. Publisher: Spandidos Publications.
- [157] Vessela N. Kristensen, Ole Christian Lingjærde, Hege G. Russnes, Hans Kristian M. Vollan, Arnoldo Frigessi, and Anne-Lise Børresen-Dale. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299–313, May 2014. Number: 5 Publisher: Nature Publishing Group.
- [158] Frontiers | Identification of Key Genes With Differential Correlations in Lung Adenocarcinoma.
- [159] Megan Beetch, Cayla Boycott, Sadaf Harandi-Zadeh, Tony Yang, Benjamin J. E. Martin, Thomas Dixon-McDougall, Kevin Ren, Allison Gacad, John H. Dupuis, Melissa Ullmer, Katarzyna Lubecka, Rickey Y. Yada, Carolyn J. Brown, LeAnn J. Howe, and Barbara Stefanska. Pterostilbene leads to DNMT3B-mediated DNA methylation and silencing of OCT1-targeted oncogenes in breast cancer cells. *The Journal of Nutritional Biochemistry*, 98:108815, December 2021.

- [160] Darawalee Wangsa, Rüdiger Braun, Christina H. Stuelten, Markus Brown, Kerry M. Bauer, Georg Emons, Leigh A. Weston, Yue Hu, Howard H. Yang, Maria Vila-Casadesús, Maxwell P. Lee, Philip Brauer, Lidia Warner, Madhvi Upender, Amanda B. Hummon, Jordi Camps, and Thomas Ried. Induced Chromosomal Aneuploidy Results in Global and Consistent Deregulation of the Transcriptome of Cancer Cells. *Neoplasia*, 21(7):721–729, July 2019.
- [161] Peng-Ping Li, Rong-Guo Li, Yu-Qing Huang, Jin-Pian Lu, Wei-Jun Zhang, and Zhen-Yu Wang. LncRNA OTUD6B-AS1 promotes paclitaxel resistance in triple negative breast cancer by regulation of miR-26a-5p/MTDH pathway-mediated autophagy and genomic instability. *Aging (Albany NY)*, 13(21):24171–24191, November 2021.
- [162] Evolutionary patterns of DNA base composition and correlation to polymorphisms in DNA repair systems | Nucleic Acids Research | Oxford Academic.
- [163] Wan Zhu, Longxiang Xie, Jianye Han, and Xiangqian Guo. The Application of Deep Learning in Cancer Prognosis Prediction. *Cancers*, 12(3):603, March 2020. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.
- [164] Jie Hao, Youngsoon Kim, Tejaswini Mallavarapu, Jung Hun Oh, and Mignon Kang. Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. *BMC Medical Genomics*, 12(10):189, December 2019.
- [165] A Deep Learning Approach for Process Data Visualization Using t-Distributed Stochastic Neighbor Embedding | Industrial & Engineering Chemistry Research.
- [166] Maad Mijwil. Artificial Neural Networks Advantages and Disadvantages. January 2018.
- [167] Cytokeratin 5 alters β -catenin dynamics in breast cancer cells | Oncogene.
- [168] Derek Dustin, Guowei Gu, and Suzanne A. W. Fuqua. ESR1 mutations in breast cancer. *Cancer*, 125(21):3714–3728, 2019. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/cncr.32345>.
- [169] Remah Ali and Michael K. Wendt. The paradoxical functions of EGFR during breast cancer progression. *Signal Transduction and Targeted Therapy*, 2(1):1–7, January 2017. Number: 1 Publisher: Nature Publishing Group.

- [170] Partha S. Ray, Jinhua Wang, Ying Qu, Myung-Shin Sim, Jaime Shamonki, Sanjay P. Bagaria, Xing Ye, Bingya Liu, David Elashoff, Dave S. Hoon, Michael A. Walter, John W. Martens, Andrea L. Richardson, Armando E. Giuliano, and Xiaojiang Cui. FOXC1 Is a Potential Prognostic Biomarker with Functional Significance in Basal-like Breast Cancer. *Cancer Research*, 70(10):3870–3876, May 2010.
- [171] Yong Hwa Eom, Hyung Suk Kim, Ahwon Lee, Byung Joo Song, and Byung Joo Chae. BCL2 as a Subtype-Specific Prognostic Marker for Breast Cancer. *Journal of Breast Cancer*, 19(3):252–260, September 2016. Publisher: Korean Breast Cancer Society.
- [172] Drivers and suppressors of triple-negative breast cancer | PNAS.
- [173] Liejun Jiang, Xiliu Zhang, Chenglin Xiang, Joseph Geradts, Qiang Wei, Yuanzi Liang, Huayi Huang, and Jun Fa Xu. Differential cellular localization of CELSR2 and ING4 and correlations with hormone receptor status in breast cancer. 2018. Accepted: 2022-05-12T08:29:01Z Publisher: Universidad de Murcia. Departamento de Biología Celular e Histología.
- [174] Gene: CSAD (ENSG00000139631) - Summary - Homo_sapiens - Ensembl genome browser 110.
- [175] Yunyao Ye, Jingyao Gu, Pei Liu, He Wang, Lihua Jiang, Tianyao Lei, Shanxun Yu, Gaohua Han, and Zhaoxia Wang. Long Non-Coding RNA SPRY4-IT1 Reverses Cisplatin Resistance by Downregulating MPZL-1 via Suppressing EMT in NSCLC. *OncoTargets and Therapy*, 13:2783–2793, April 2020. Publisher: Dove Medical Press _eprint: <https://www.tandfonline.com/doi/pdf/10.2147/OTT.S232769>.
- [176] Guanming Lu, Yuanhui Lai, Tiantian Wang, Weihao Lin, Jinlan Lu, Yanfei Ma, Yongcheng Chen, Haiqing Ma, Ruilei Liu, and Jie Li. Mitochondrial fission regulator 2 (MTFR2) promotes growth, migration, invasion and tumour progression in breast cancer cells. *Aging (Albany NY)*, 11(22):10203–10219, November 2019.