INTEGRATED BIOINFORMATICS AND MACHINE LEARNING APPROACH TO IDENTIFY CHEMOTHERAPY RESISTANCE BIOMARKERS AND DEVELOP PROGNOSTIC MODELS IN BREAST CANCER

ZHANG HAO

UNIVERSITI SAINS MALAYSIA

2025

INTEGRATED BIOINFORMATICS AND MACHINE LEARNING APPROACH TO IDENTIFY CHEMOTHERAPY RESISTANCE BIOMARKERS AND DEVELOP PROGNOSTIC MODELS IN BREAST CANCER

by

ZHANG HAO

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

September 2025

ACKNOWLEDGEMENT

I hereby extend my heartfelt gratitude to all those who have supported me in completing my PhD thesis.

I would like to express my deep appreciation to my main supervisors Prof. Dr. Badrul Hisham Yahaya. Your profound academic expertise, rigorous scholarly approach, and inclusive research spirit have guided me in the right direction. From the design of the research topic to the writing of the thesis, you have consistently led me with patience and wisdom to overcome obstacles. This experience will become a precious asset in my academic career. I sincerely thank my co-supervisor, Prof. Dr. Lin Juntang. Your broad academic perspective, interdisciplinary research approach, and selfless support of experimental resources have laid a solid foundation for the smooth progression of the project. Your enthusiasm for research and your spirit of generosity have allowed me to deeply appreciate the grandeur and warmth of scholarly demeanor. I also wish to thank Dr. Sun Yuliang. During the cellular and animal experimentation phases of the thesis, your rich expertise guided the experimental design, you patiently assisted in optimizing operational details, and you provided crucial advice during data analysis. Your professionalism and dedication have profoundly touched me. Furthermore, I am grateful for the collaborative efforts of my lab colleagues, the silent contributions of my family, and the encouragement and companionship of my friends.

The academic journey is long and fraught with challenges; only with a heart full of gratitude can one forge ahead. I would like to once again thank all those who have acted as sparks illuminating my research journey!

TABLE OF CONTENTS

ACKNOWLEDGEMENT ..... **ii**

TABLE OF CONTENTS ..... **iii**

LIST OF TABLES ..... **xi**

LIST OF FIGURES ..... **xii**

LIST OF SYMBOLS ..... **xvi**

LIST OF ABBREVIATIONS ..... **xvii**

LIST OF APPENDICES ..... **xxi**

ABSTRAK ..... **xxii**

ABSTRACT ..... **xxiv**

CHAPTER 1 INTRODUCTION **1**

1.1 Breast Cancer: Epidemiology and Burden 1

1.2 Chemotherapy Resistance in Breast Cancer 3

1.2.1 Molecular mechanisms of drug resistant 4

1.2.2 Biomarkers 6

1.2.3 Clinical significance 8

1.2.4 Research gap 8

1.3 Prognostic Models in Breast Cancer 10

1.4 Big Data 12

1.4.1 Foundations of big data 14

1.4.2 Big data-driven precision medicine: integration of multi-omics and methodologies 15

1.5 Machine Learning (ML) 21

1.5.1 Fundamental concepts and development overview 21

1.5.2 The synergy between big data and artificial intelligence (AI) 26

1.5.3 AI and precision medicine: the hope for innovation in healthcare 27

1.5.4 Ethics and governance of AI for health 29

1.6 Big Data and AI: Empowering Precision Treatment for Breast Cancer 31

1.6.1 Breast cancer: from data to insights and decision 31

1.6.2 Precision treatment for breast cancer: diversity of data types 32

1.6.3 The role and contributions of AI in modeling for precision treatment of breast cancer 34

1.7 Problem Statement 37

1.8 Research Objectives 39

1.8.1 General objective 39

1.8.2 Specific objectives 39

1.9 Hypothesis 39

1.10 Research Frameworks 40

CHAPTER 2 UNVEILING THE RESEARCH LANDSCAPE OF CHEMOTHERAPY RESISTANCE IN BREAST CANCER: A BIBLIOMETRIC ANALYSIS 44

2.1 Introduction 44

2.2 Objective of the Chapter 46

2.2.1 General objective 46

2.2.2 Specific objectives 46

2.3 Materials and Methods 47

2.3.1 Flowchart of the research in this chapter 47

2.3.2 Data sources 48

2.3.3 Search strategy 49

2.3.4 Eligibility criteria 50

2.3.4(a) Inclusion criteria 50

2.3.4(b) Exclusion criteria 50

2.3.5 Data organization 51

2.3.6 Data analysis 51

2.3.6(a) Application of COOC software 52
2.3.6(b) Application of CiteSpace software 53
2.3.6(c) Application of R software 54
2.4 Results 56
2.4.1 Changes in the trend of published articles 56
2.4.2 Discipline distribution and journal analysis 58
2.4.3 Countries/regions, institutions, authors, and their collaborative relationships 64
2.4.3(a) Publication volume by country/region 64
2.4.3(b) Publication volume by institution 69
2.4.3(c) Publication volume by author 69
2.4.4 Keywords analysis to identify research hotspots 71
2.4.5 Literature and references analysis 79
2.4.6 Focus in the field 87
2.5 Discussion 88
2.6 Conclusions of the Chapter 96
CHAPTER 3 DEVELOPMENT OF ML-DRIVEN PROGNOSTIC MODELS FOR CHEMOTHERAPY RESISTANCE IN BREAST CANCER AND RISK-STRATIFIED PRECISION TREATMENT STRATEGIES 98
3.1 Introduction 98
3.2 Objective of the Chapter 102
3.2.1 General objective 102
3.2.2 Specific objectives 102
3.3 Materials and Methods 103
3.3.1 Workflow of ML-based prognostic model development 103
3.3.2 Breast cancer data sources and preparation 105
3.3.2(a) Acquisition of training cohort data 106
3.3.2(b) Breast cancer cell line data 106

3.3.2(c) Collection of validation cohort data 106

3.3.2(d) Multi-omics data supplementation 107

3.3.3 Cell line data analysis 110

3.3.4 Functional enrichment analysis and protein-protein interaction (PPI) network construction 111

3.3.5 TCGA-BRCA cohort data processing 112

3.3.6 Development and verification of a breast cancer prognostic model based on LASSO and Cox regression 113

3.3.7 Meta-analysis of model robustness evaluation 116

3.3.8 Association analysis of Breast Cancer Chemoresistance Index (BCCI) with clinical characteristics 117

3.3.9 Analysis of BCCI with genomic variations and functional pathways 118

3.3.10 Association analysis of BCCI with cell death patterns 120

3.3.11 BCCI and TME, immune checkpoint (IC) analysis 121

3.3.11(a) TME scoring analysis 121

3.3.11(b) Evaluation of tumor-infiltrating immune cells (TIIC) 121

3.3.11(c) Analysis of BCCI and IC 122

3.3.12 Analysis of BCCI and prediction of responses to immunotherapy 122

3.3.12(a) The acquisition of response indicators for immunotherapy 122

3.3.12(b) Relationship between BCCI and immunotherapy response 124

3.3.12(c) Verification of the IMvigor210 cohort 124

3.3.13 BCCI and drug sensitivity analysis 126

3.3.14 Model comparison analysis 127

3.3.15 Identification of core genes in the model 131

3.3.16 Statistical analysis 131

3.4 Results 132

3.4.1 Identifying chemotherapy resistance genes 132

3.4.2 Enrichment analysis and PPI network construction 135

3.4.3 Identification of differentially expressed prognostic genes and molecular characterization analysis 137

3.4.4 ML-based prognostic model construction 140

3.4.5 Validation of the prognostic model 143

3.4.6 Meta-analysis of model evaluation 148

3.4.7 Analysis of BCCI combined with clinical features 151

3.4.8 Analysis of independent prognostic factors and nomogram construction 153

3.4.9 Somatic mutations, ITH, and model gene enrichment analysis ... 155

3.4.10 Analysis of cell death patterns 158

3.4.10(a) Distribution differences of cell death scores between high and low BCCI groups 158

3.4.10(b) Correlation analysis between BCCI and cell death scores 160

3.4.11 Immune characterization analysis 162

3.4.12 Immune therapy analysis 164

3.4.13 Drug sensitivity analysis 167

3.4.14 Comparison of multiple prognostic models 169

3.4.15 Identification of key genes in the model 171

3.5 Discussion 173

3.6 Conclusions of the Chapter 183

CHAPTER 4 SDC1 AS A KEY BIOMARKER IDENTIFIED THROUGH THE ML-BASED PROGNOSTIC MODEL: PAN-CANCER ANALYSIS, MECHANISTIC INSIGHTS INTO BREAST CANCER CHEMORESISTANCE, AND FUNCTIONAL VALIDATION 184

4.1 Introduction 184

4.2 Objective of the Chapter 188

4.2.1 General objective 188

4.2.2 Specific objectives 188

4.3 Materials and Methods 189

4.3.1 Flowchart of the research in this chapter 189

4.3.2 Data collection and processing 190

4.3.3 Analysis of the transcriptome expression patterns and prognostic value of SDC1 in pan-cancer 194

4.3.4 Protein expression patterns and subcellular structural localization of SDC1 in pan-cancer 194

4.3.5 The relationship between SDC1 and chemotherapy in breast cancer 195

4.3.6 Clinical characteristics analysis of SDC1 and chemotherapy resistance 196

4.3.7 Assessment of the prognostic value of SDC1 in the chemotherapy-resistant population 197

4.3.8 Functional and pathway enrichment analysis of SDC1 198

4.3.9 Analyze the relationship between SDC1 and TME 200

4.3.10 Expression characteristics of SDC1 in single-cell and spatial transcriptomics 202

4.3.11 Analysis of expression characteristics in different breast cancer cell lines 203

4.3.12 Preparation for experimental work 204

4.3.12(a) Cell culture and passage 204

4.3.12(b) Lentiviral short hairpin RNA (shRNA)-mediated knockdown of SDC1: protocol and validation 205

4.3.13 Quantitative real-time polymerase chain reaction (qPCR) to detect gene expression 207

4.3.14 The cell counting kit-8 (CCK-8) method was used to detect cell proliferation and viability 208

4.3.15 Colony formation assay analysis 209

4.3.16 Cell cycle detection 209

4.3.17 Animal experiments 209

4.3.17(a) Ethical approval 209

4.3.17(b) Nude mouse tumorigenesis experiment 210
4.3.17(c) Tumor tissue collection 210
4.3.17(d) Hematoxylin and eosin (H&E) staining 211
4.3.18 Drug enrichment and molecular docking analysis 211
4.3.19 Statistical analysis 212
4.4 Results 213
4.4.1 Pan-cancer characteristics and subcellular localization of SDC1 213
4.4.2 The relationship between SDC1 and chemotherapy as well as clinical characteristics 220
4.4.3 Independent prognostic value of SDC1 and nomogram construction 224
4.4.4 The correlation between SDC1 and chemotherapy resistance in breast cancer 226
4.4.5 SDC1 and TME Analysis 230
4.4.6 Single-cell and spatial transcriptomics analysis 234
4.4.7 Expression in breast cancer cell lines and qPCR analysis 238
4.4.8 CCK-8 detection analysis 240
4.4.9 Analysis of colony formation assay in cells 242
4.4.10 Cell cycle analysis 243
4.4.11 In *vivo* experimental analysis 244
4.4.12 Molecular docking analysis 247
4.5 Discussion 249
4.6 Conclusions of the Chapter 257
CHAPTER 5 CONCLUSION AND FUTURE RECOMMENDATIONS 259
5.1 Summary of the Results 259
5.2 Limitations of the Study and Future Directions 266
5.3 Conclusions 269

REFERENCES..... 272

APPENDICES

LIST OF PUBLICATIONS

LIST OF TABLES

Page

Table 1.1 Application of digitisation in healthcare sectors by focusing on cancers 19

Table 1.2 Applications of ML in the field of biomedicine (focusing on oncology) 25

Table 2.1 Parameter setting explanation 55

Table 2.2 Information on the top 15 discipline categories by frequency of occurrence 60

Table 2.3 Information on the top 10 journals by volume of publications and number of citations from reference sources 61

Table 2.4 Information on the top 10 countries/regions, institutions, and authors ranked by publication volume 66

Table 2.5 Information on the top 30 keywords by frequency of occurrence 73

Table 2.6 Information on the top 10 articles ranked by number of citations 80

Table 2.7 Information on the top 10 references ranked by number of citations 84

Table 3.1 Information on breast cancer data collected and organized in this chapter 108

Table 3.2 Basic information of the models included in the comparison 129

Table 4.1 Table 4.1: The breast cancer data downloaded and organized in this chapter 192

LIST OF FIGURES

Page

Figure 1.1 Diagram of the Relationship Between ML and AI and an Overview of ML Methods 23

Figure 1.2 The application and prospects of AI technology in precision medicine 29

Figure 1.3 The workflow of this study 43

Figure 2.1 Steps and processes of bibliometric analysis in this chapter..... 48

Figure 2.2 The analysis of publication volume in breast cancer chemoresistance 57

Figure 2.3 The analysis of the discipline category, co-citation of journals, and dual-map overlay in breast cancer chemoresistance. 63

Figure 2.4 The analysis of the publishing countries/regions for papers on breast cancer chemoresistance. 68

Figure 2.5 An analysis of the publishing institutions and authors in the field of breast cancer chemoresistance. 70

Figure 2.6 The keyword analysis in breast cancer chemoresistance (frequency count, co-occurrence network, annual attention trends). 75

Figure 2.7 Thematic evolution path of keywords. 76

Figure 2.8 The keyword analysis in breast cancer chemoresistance (clustering and burst analysis). 78

Figure 2.9 The analysis of references in breast cancer chemoresistance. 86

Figure 2.10 The current research focus on breast cancer chemoresistance. 88

Figure 3.1 ML-based prognostic model research process. 105

Figure 3.2 The selection process for chemotherapy resistance-related genes in breast cancer. 134

Figure 3.3 Enrichment analysis of chemotherapy resistance genes. 136

Figure 3.4 Identification of the top core proteins in the PPI network constructed based on the STRING database using Cytoscape software. 137

Figure 3.5 Screening of chemoresistance-related prognostic genes and molecular feature analysis based on the TCGA-BRCA cohort. 139

Figure 3.6 Prognostic model construction based on the TCGA-BRCA cohort using ML. 142

Figure 3.7 Validation results of the model in independent cohorts I. 145

Figure 3.8 Independent cohort validation model results II. 147

Figure 3.9 Meta-analysis of model evaluation. 150

Figure 3.10 Analysis of BCCI and clinical characteristics. 152

Figure 3.11 Analysis of independent prognostic factors and construction of the prognostic nomogram. 154

Figure 3.12 Somatic mutations, ITH, and gene enrichment analysis. 157

Figure 3.13 Analysis of differences in various modes of cell death between high and low BCCI groups. 159

Figure 3.14 Analysis of the correlation between BCCI and different modes of cell death. 161

Figure 3.15 Analysis of BCCI and immune characteristics. 163

Figure 3.16 The role of BCCI in predicting immunotherapy. 166

Figure 3.17 Analysis of BCCI in drug sensitivity and correlation. 168

Figure 3.18 Model comparative analysis. 170

Figure 3.19 Identification of core genes in the model. 172

Figure 4.1 Workflow of SDC1 biomarker investigation 190

Figure 4.2 SDC1 shRNA plasmid map and validation of knockdown efficiency 207

Figure 4.3 Pan-cancer analysis of SDC1 mRNA expression levels 215

Figure 4.4 Pan-cancer analysis of SDC1 protein expression levels 217

Figure 4.5 Pan-cancer prognostic analysis of SDC1 219

Figure 4.6 Analysis of SDC1 in relation to chemotherapy and prognosis in breast cancer.....221

Figure 4.7 Analysis of SDC1 in relation to the clinical characteristics of chemotherapy resistance in breast cancer 223

Figure 4.8 Analysis of Independent Prognostic Factors of SDC1 in Patients with Chemotherapy-Resistant Breast Cancer and Nomogram Construction 225

Figure 4.9 Enrichment analysis of functions and pathways involving SDC1 based on the TCGA-BRCA cohort 227

Figure 4.10 Meta-analysis of the correlation between SDC1 and the ECM, PI3K/AKT, and EMT signaling pathways based on chemotherapy cohorts.....229

Figure 4.11 The relationship between SDC1 and TME 231

Figure 4.12 The relationship between SDC1 and TME in chemotherapy breast cancer 233

Figure 4.13 Analysis of SDC1 expression in fibroblasts based on single-cell transcriptomics 235

Figure 4.14 Spatial transcriptomics analysis of SDC1 and fibroblasts 237

Figure 4.15 Expression and qPCR analysis of SDC1 and ASCL1 in breast cancer cell lines.....239

Figure 4.16 CCK-8 assay analysis of the effect of SDC1 on the proliferative activity of MCF-7 cells 241

Figure 4.17 CCK-8 assay analysis of the effect of SDC1 on the proliferative activity of MCF-7/ADR cells.....241

Figure 4.18 Analysis of the effect of SDC1 on the colony formation of MCF-7 and MCF-7/ADR cells 242

Figure 4.19 Analysis of the effect of SDC1 on the cell cycle of MCF-7/ADR cells 244

Figure 4.20 In *vivo* experimental analysis of the role of SDC1 246

Figure 4.21 Analysis of potential SDC1-targeted inhibitors screening.....248

Figure 5.1 Chapter framework of this study.....261

Figure 5.2 Summary of the main findings of this study 262

LIST OF SYMBOLS

β	Beta
CO_2	Carbon dioxide
g	gram
I^2	I-squared
λ	Lambda
×g	g-force
℃	Degree Celsius
µg	Microgram
ng	Nanogram
kg	Kilogram
µl	Microliter
ml	Milliliter
µm	Micrometer
min	Minutes
mm^3	Cubic millimeter
%	Percentage
™	Trademark

LIST OF ABBREVIATIONS

ABC	ATP-binding cassette
AI	Artificial Intelligence
ASCL1	Achaete-Scute Family bHLH Transcription Factor 1
AUC	Area under the curve
BCCI	Breast Cancer Chemoresistance Index
Bcl-2	B-cell lymphoma-2
Bcl-xL	B-cell lymphoma-extra large
BCRP	Breast cancer resistance protein
BCSCs	Breast cancer stem cells
BDA	Big data analytics
BP	Biological processes
BTK	Bruton's tyrosine kinase
CAD	Computer-aided diagnosis
CC	Cellular components
CCK-8	Cell counting kit-8
CCL5	C-C Motif Chemokine Ligand 5
CCLE	Cancer Cell Line Encyclopedia
CEGs	Co-expressed genes
CI	Confidence interval
C-index	Concordance index
circRNA	circular RNA
CNV	Copy number variation
COOC	Co-Occurrence
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CR	Complete Response
CSCs	Cancer stem cells
CTLA-4	Cytotoxic T-lymphocyte associated protein 4
DCA	Decision curve analysis
DDR	DNA damage response
DEGs	Differential expression genes
DFI	Disease-free interval

DMEM	Dulbecco's Modified Eagle Medium
DO	Disease Ontology
DPP4	Dipeptidyl Peptidase 4
DSigDB	Drug SIGnatures DataBase
DSS	Disease-specific survival
EB	Exabytes
ECM	Extracellular matrix
EDTA	Ethylenediaminetetraacetic acid
EGFR	Epidermal growth factor receptor
EMT	Epithelial-mesenchymal transition
FBS	Fetal bovine serum
FDR	False discovery rate
FFDM	Full-field digital mammography
FOXJ1	Forkhead Box J1
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
G0/G1	Gap 0/Gap 1
G2/M	Gap 2/Mitosis
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GBCI	Global Breast Cancer Initiative
GDC	Genomic Data Commons
GDSC	Genomics of Drug Sensitivity in Cancer
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSEA	Gene set enrichment analysis
H&E	Hematoxylin and eosin
HER2	Human epidermal growth factor receptor-2
HIF-1α	Hypoxia-Inducible Factor-1α
HPA	Human Protein Atlas
HR	Hazard ratios
IACUC	Institutional Animal Care and Use Committee
IC	Immune checkpoint
IC_{50}	Half-maximal inhibitory concentration
ICB	Immune checkpoint blockade
ICIs	Immune checkpoint inhibitors

IF	Impact Factor
IPS	Immunophenotype score
ITH	Intratumor heterogeneity
JCR	Journal Citation Report
KEGG	Kyoto Encyclopedia of Genes and Genomes
K-M	Kaplan-Meier
LASSO	Least Absolute Shrinkage and Selection Operator
LLMs	Large language models
lncRNA	Long noncoding RNA
MCF-7	Michigan Cancer Foundation-7
MCF-7/ADR	Michigan Cancer Foundation-7 / Adriamycin-resistant
MeSH	Medical Subject Headings
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
MF	Molecular functions
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MRP	Multidrug resistance-associated protein
MSI	Microsatellite instability
ncRNA	noncoding RNA
NCBI	National Center for Biotechnology Information
NF-κB	Nuclear Factor kappa B
NGS	Next-generation sequencing
OD	Optical density
OS	Overall survival
PBS	Phosphate-buffered saline
PCA	Principal component analysis
PD	Progressive Disease
PD-1	Programmed cell death protein 1
PD-L1	Programmed death-ligand 1
PFI	Progression-free interval
P-gp	P-glycoprotein
PI3K/AKT	Phosphoinositide 3-kinase/Protein Kinase B
PPI	Protein-protein interaction
PR	Partial Response

qPCR	Quantitative real-time polymerase chain reaction
ROC	Receiver operating characteristic
RSF	Random Survival Forest
S	Synthesis
SCI-E	Science Citation Index Expanded
SD	Stable Disease
SDC1	Syndecan-1
SDC1-KD	SDC1 knockdown
shRNA	Short hairpin RNA
sSDC1	Soluble SDC1
ssGSEA	Single-Sample Gene Set Enrichment Analysis
TB	Terabytes
TCGA	The Cancer Genome Atlas
TCGA-BRCA	The Cancer Genome Atlas Breast Invasive Carcinoma
TCIA	The Cancer Immunome Atlas
TFF1	Trefoil Factor 1
TGF-β	Transforming growth factor beta
TIDE	Tumor immune dysfunction and exclusion
TIIC	Tumor-infiltrating immune cells
TISCH	Tumor Immune Single-cell Hub
TMB	Tumor mutation burden
TME	Tumor microenvironment
TNBC	Triple-negative breast cancer
TNM	Tumor-Node-Metastasis
UMAP	Unified manifold approximation and projection
WGCNA	Weighted gene co-expression network analysis
WoSCC	Web of Science Core Collection
WT1	Wilms Tumor 1
|log2FC|	Absolute value of log2 fold change

LIST OF APPENDICES

Appendix A Supplementary Tables

Appendix B Supplementary Figures

Appendix C Ethic Approval Letters

PENDEKATAN BIOINFORMATIK DAN PEMBELAJARAN MESIN TERANGKUM UNTUK MENGENAL PASTI PENANDA RINTANGAN KEMOTERAPI DAN MEMBANGUNKAN MODEL PROGNOSTIK DALAM KANSER PAYUDARA

ABSTRAK

Rintangan terhadap kemoterapi kekal sebagai cabaran kritikal dalam rawatan kanser payudara, yang mengehadkan keberkesanan terapeutik dan menjejaskan kelangsungan hidup jangka panjang. Tesis ini mengguna pakai pendekatan bersepadu pelbagai peringkat untuk menangani isu ini, bermula dengan analisis bibliometrik bagi memetakan arah aliran penyelidikan global dan menyerlahkan keutamaan metodologi dalam kajian rintangan kemoterapi. Berdasarkan dapatan ini, set data multi-omik berskala besar yang mewakili data raya, termasuk profil transkriptomik daripada TCGA dan GEO, telah dianalisis secara sistematik menggunakan teknik bioinformatik seperti analisis ekspresi berbeza dan analisis rangkaian pengeksperesian bersama gen berpemberat (WGCNA). Seterusnya, strategi pemilihan ciri berasaskan pembelajaran mesin menggunakan regresi pengecutan mutlak terkecil (LASSO) telah digunakan untuk mengenal pasti gen yang paling prognostik, diikuti oleh analisis regresi Cox pelbagai pemboleh ubah bagi membina model ramalan yang kukuh. Daripada 850 calon gen yang berkaitan dengan rintangan kemoterapi, tujuh gen utama telah dipilih untuk membentuk Indeks Rintangan Kemoterapi Kanser Payudara (BCCI). Model ini mengelaskan pesakit berdasarkan hasil kelangsungan hidup dan telah disahkan dalam 12 kohort bebas, menunjukkan prestasi ramalan yang memberangsangkan. Pemprofilan tindak balas terapeutik seterusnya mendedahkan bahawa pesakit dengan skor BCCI tinggi lebih berkemungkinan mendapat manfaat daripada terapi tersasar,

manakala pesakit dengan skor BCCI rendah menunjukkan respons yang lebih baik terhadap kemoterapi dan imunoterapi. Dalam kalangan biopenanda yang dikenal pasti, Syndecan-1 (SDC1) telah disahkan secara fungsi sebagai perantara utama rintangan kemoterapi, yang menggalakkan perkembangan tumor dan rintangan terhadap ubatan. Ia menyumbang kepada rintangan dengan merombak persekitaran mikro tumor dan mengaktifkan tapak jalan PI3K/AKT. Pengekspresan SDC1 yang meningkat dalam tisu pesakit dikaitkan dengan prognosis yang buruk, sekali gus menyokong potensinya sebagai sasaran terapeutik. Secara keseluruhan, tesis ini membentangkan satu kerangka berasaskan data yang mengintegrasikan bibliometrik, bioinformatik dan pembelajaran mesin untuk menjelaskan asas rintangan molekul kemoterapi dalam kanser payudara. Pembangunan model BCCI dan penemuan SDC1 sebagai sasaran ubatan yang berpotensi memberikan pandangan baharu untuk penstrataan pesakit dan terapi jitu, seterusnya menyumbang kepada penambahbaikan pengurusan klinikal kanser payudara.

INTEGRATED BIOINFORMATICS AND MACHINE LEARNING APPROACH TO IDENTIFY CHEMOTHERAPY RESISTANCE BIOMARKERS AND DEVELOP PROGNOSTIC MODELS IN BREAST CANCER

ABSTRACT

Chemoresistance remains a critical challenge in breast cancer treatment, limiting therapeutic efficacy and compromising long-term survival. This thesis adopts a multi-stage, integrative approach to address this issue, beginning with a bibliometric analysis to map global research trends and highlight methodological priorities in chemoresistance studies. Building upon these insights, large-scale multi-omics datasets representative of big data, including transcriptomic profiles from TCGA and GEO, were systematically analysed using bioinformatics techniques such as differential expression analysis and weighted gene co-expression network analysis (WGCNA). Subsequently, a machine learning–based feature selection strategy using least absolute shrinkage and selection operator (LASSO) regression was applied to identify the most predictive genes, followed by multivariate Cox proportional hazards regression analysis to construct a robust prognostic model. From 850 candidate genes associated with chemoresistance, seven core genes were selected to build the Breast Cancer Chemoresistance Index (BCCI). This model stratifies patients by survival outcomes and was validated in 12 independent cohorts, demonstrating strong predictive performance. Therapeutic response profiling further revealed that high-BCCI patients are more likely to benefit from targeted therapies, while low-BCCI patients show greater responsiveness to chemotherapy and immunotherapy. Among the identified biomarkers, Syndecan-1 (SDC1) was functionally validated as a key

mediator of chemoresistance, promoting tumour progression and drug resistance. It contributes to chemoresistance by remodelling the tumour microenvironment and activating the PI3K/AKT pathway. Its elevated expression in patient tissues was strongly associated with poor prognosis, supporting its therapeutic potential. In summary, this thesis presents a data-driven framework integrating bibliometrics, bioinformatics, and machine learning to elucidate the molecular basis of chemoresistance in breast cancer. The development of the BCCI model and discovery of SDC1 as a druggable target provide novel insights for patient stratification and precision therapy, contributing to improved clinical management of breast cancer.

CHAPTER 1

INTRODUCTION

1.1 Breast Cancer: Epidemiology and Burden

The burden of breast cancer is increasing globally. In 2020, among the newly diagnosed 19.3 million cancer patients, 2.26 million cases were diagnosed as breast cancer, surpassing lung cancer as the most common type of cancer worldwide (Sung et al., 2021; Lei et al., 2021; Newman, 2023). Breast cancer not only ranks as the fourth leading cause of cancer deaths worldwide but also represents the primary cause of female mortality (Newman, 2023). Due to its high incidence and mortality rate, breast cancer has become a global challenge and poses a significant threat to women's health. In recent years, the mortality rate among Western populations, especially in young age groups, has declined due to advancements in breast cancer diagnosis and treatment (Autier et al., 2010; Allemani et al., 2015). However, breast cancer remains the leading cause of cancer-related deaths in women globally (Cardoso et al., 2019). In a study in Europe, it was estimated to have 66 women diagnosed with breast cancer every hour, and tragically, 18 of these cases result in death (Berdzuli, 2023). Without adequate control measures, it is estimated that the incidence of new breast cancer cases will rise by 26.7%, and the number of deaths will increase by 25.2% over the next 20 years (Berdzuli, 2023). In the United States, breast cancer is the second leading cause of cancer-related deaths after lung cancer (Giaquinto et al., 2022b). According to the American Cancer Society, the average risk of developing breast cancer in U.S. women is approximately 13%, meaning that 1 in every 8 women will be diagnosed with breast cancer, and approximately 1 in 39 women (approximately 3%) will die from breast cancer (Giaquinto et al., 2022b; Giaquinto et al., 2022a). Recent studies have highlighted the staggering impact of breast cancer across different countries. In the

United States, Giaquinto et al. reported that over 310,000 cases of invasive breast cancer were identified in 2024 alone, claiming the lives of roughly 42,250 women (Giaquinto et al., 2024). In China, the situation appeared even more severe—Han et al. noted that 357,200 women were newly diagnosed with breast cancer in 2022, and 75,000 deaths were attributed to the disease, reinforcing its role as a leading cause of cancer-related mortality among women (Han et al., 2024). The situation in Malaysia was similarly alarming, where breast cancer remained the most frequently diagnosed malignancy in women. Research by Yip et al. and Dahlui et al. revealed an age-standardized incidence rate of 46.2 per 100,000 women, accounting for nearly one-third of all female cancer cases in the country (Yip et al., 2006; Dahlui et al., 2011). According to data collected from 2005 to 2009, the survival rate for breast cancer patients was 43.5%, which is lower than levels reported in Western countries (Ibrahim et al., 2012). Statistics from 2022 indicate that the mortality rate was 19.3 deaths per 100,000 women, ranking second among Southeast Asian countries (Dee et al., 2024). Globally, new breast cancer cases will reach nearly 3 million by 2040 (Sedeta et al., 2023). In light of the rising global burden of breast cancer, there is an urgent need to enhance prevention strategies, optimize management approaches, and advance treatment modalities, making it a critical public health priority.

The World Health Organization has released the Global Breast Cancer Initiative (GBCI), a key measure to address the global breast cancer issue (Abila et al., 2022). Its primary objective is to assist governments and international organizations in formulating practical and feasible policies to improve breast cancer incidence and mortality rates among women. The core goal of this proposal is to reduce the global breast cancer mortality rate by 2.5% annually, with a vision to save 2.5 million lives by 2040 (Abila et al., 2022). The initiative aims to achieve a reduction in cancer

incidence through a series of measures that enhance awareness of breast cancer risk factors and related hazards, as well as screening levels. The GBCI, jointly established by the U.S. National Center for Chronic Disease Prevention and Health Promotion and the Centers for Disease Control and Prevention, consists of three core components: early health promotion screening, timely breast cancer diagnosis, and comprehensive breast cancer management (Wilkinson and Gathani, 2022). These pillars collectively form the strategic foundation for the initiative to achieve its ambitious goals. The GBCI is a new concept driven by data and prevention, which will bring revolutionary changes to breast cancer screening and treatment in the future (Priyadarshini et al., 2022). With the active participation of stakeholders and the achievement of the grand objectives of the GBCI, it becomes crucial to research how to utilize big data analytics (BDA) to improve the effectiveness and accuracy of breast cancer interventions.

1.2 Chemotherapy Resistance in Breast Cancer

The prevalence of breast cancer is gradually increasing globally, especially in developing countries. The burden of breast cancer continues to rise due to various factors, including changes in lifestyle, environmental changes, and an aging population (Bray et al., 2018). Over the past decades, significant medical advancements have substantially improved survival outcomes for patients with early-stage breast cancer. Currently, standard therapeutic approaches include surgical resection, radiation therapy, chemotherapy, endocrine therapy, and targeted treatments. In the comprehensive treatment of breast cancer, chemotherapy is viewed as a key component, with extensive applications in novel adjuvant therapies, adjuvant treatments, and palliative care for advanced breast cancer. Common chemotherapy drugs include anthracyclines (such as doxorubicin), taxanes (such as paclitaxel), and

platinum-based drugs (such as cisplatin) (Gray et al., 2019). However, despite the significant effectiveness of chemotherapy in the treatment of breast cancer, the emergence of drug resistance severely restricts its clinical application. Therefore, understanding the mechanisms of chemotherapy resistance and taking corresponding countermeasures is of great clinical significance. Chemotherapy resistance can be classified into two main categories: primary resistance and acquired resistance. Primary resistance refers to tumor cells being insensitive to chemotherapy drugs at the onset of treatment, while acquired resistance develops gradually during the course of treatment (Holohan et al., 2013). Whether it is primary or acquired resistance, both lead to a decline in the effectiveness of chemotherapy and even treatment failure. Consequently, a detailed exploration of the molecular mechanisms underlying chemotherapy resistance in breast cancer and the identification of effective strategies to reverse this resistance have become focal issues in current breast cancer research.

1.2.1 Molecular mechanisms of drug resistant

The formation of chemotherapy resistance in breast cancer involves a variety of complex molecular mechanisms, including the overexpression of drug efflux pumps, enhanced DNA repair capabilities, inhibition of apoptosis pathways, and changes in the tumor microenvironment. Drug efflux pumps are one of the key mechanisms by which tumor cells develop resistance. The ATP-binding cassette (ABC) transporter family is the most extensively studied group of drug efflux pumps, including P-glycoprotein (P-gp), multidrug resistance-associated protein (MRP), and breast cancer resistance protein (BCRP) (Gottesman et al., 2002). These transport proteins use ATP energy to pump chemotherapy drugs out of the cells, thereby reducing intracellular drug concentration and weakening the cytotoxic effects of these drugs. Studies have shown that the overexpression of P-gp in breast cancer cells is closely related to

resistance to anthracyclines and taxanes (Fletcher et al., 2010). Additionally, the expression of BCRP is associated with the resistance of breast cancer cells to topoisomerase inhibitors (Doyle and Ross, 2003). DNA damage is one of the primary mechanisms by which many chemotherapy drugs (such as anthracyclines and platinum-based drugs) exert their antitumor effects. However, tumor cells can effectively repair DNA damage caused by chemotherapy drugs by enhancing their DNA repair capabilities, leading to resistance. For instance, the activation of homologous recombination repair and non-homologous end joining pathways in breast cancer cells can repair DNA double-strand breaks and reduce the cytotoxicity of chemotherapy drugs (Lord and Ashworth, 2012). Moreover, upregulation of DNA damage response (DDR) proteins in breast cancer cells is also closely related to their resistance to chemotherapy drugs (O'Connor, 2015). Apoptosis is one of the primary pathways through which chemotherapy drugs induce tumor cell death. However, breast cancer cells can evade the cytotoxic effects of these drugs by inhibiting apoptosis pathways. For example, the overexpression of B-cell lymphoma-2 (Bcl-2) and B-cell lymphoma-extra large (Bcl-xL) can inhibit apoptosis through the mitochondrial pathway, enhancing the resistance of breast cancer cells to chemotherapy drugs (Czabotar et al., 2014). Additionally, the upregulation of members of the apoptosis-inhibitory protein family is also closely related to their resistance to chemotherapy drugs (Fulda and Vucic, 2012). Chemoresistance in breast cancer is profoundly influenced by the tumor microenvironment (TME), which serves as a dynamic and complex regulatory network. The TME is composed of not only malignant cells but also stromal components, immune infiltrates, vascular endothelial cells, and the extracellular matrix (ECM), these constituents engage in intricate interactions that collectively modulate tumor cell behavior and therapeutic response

(Hanahan and Weinberg, 2011). Research indicates that inflammatory factors and growth factors in the TME can activate various signaling pathways, such as Nuclear Factor kappa B (NF-κB) and Phosphoinositide 3-kinase/Protein Kinase B (PI3K/AKT), promoting the survival and proliferation of breast cancer cells, thereby enhancing their resistance to chemotherapy drugs (Balkwill and Mantovani, 2012). Furthermore, hypoxic conditions in the tumor microenvironment can also induce resistance in breast cancer cells. Under hypoxic conditions, the upregulation of Hypoxia-Inducible Factor-1α (HIF-1α) can promote the survival and invasion of breast cancer cells and enhance their resistance to chemotherapy drugs (Semenza, 2012).

1.2.2 Biomarkers

In recent years, the discovery and validation of biomarkers have gradually become the core focus in the field of research on chemotherapy resistance in breast cancer. Biomarkers not only help researchers gain a deeper understanding of the molecular mechanisms underlying chemotherapy resistance but also provide important evidence for clinicians in early prediction of treatment responses, assessment of efficacy, and formulation of personalized treatment plans. In recent years, the swift progress in molecular biology and multi-omics approaches has enabled the identification of a wide array of molecules involved in breast cancer chemoresistance. These molecular candidates are increasingly being incorporated into studies on clinical prognostic assessment and the development of targeted therapeutic strategies.

Notably, ABC transporters, particularly P-gp and BCRP, have been extensively investigated as key biomarkers implicated in chemoresistance. These transporters have significant clinical relevance due to their association with poor prognosis and potential as therapeutic targets for overcoming chemoresistance (Chen et al., 2024b). Emerging evidence suggests that elevated BCRP expression is strongly

linked to poor clinical outcomes in breast cancer patients following neoadjuvant chemotherapy (Tada et al., 2023).

Moreover, in-depth multi-omics studies have revealed a set of novel molecules closely related to chemotherapy resistance in breast cancer. For example, transcriptomic analysis has found that genes such as Cyclin D2, Tetratricopeptide Repeat Domain 4, and Sorting Nexin 15 are significantly upregulated in breast cancer tissues that are insensitive to neoadjuvant chemotherapy (Lichtenfels et al., 2023); at the same time, research has pointed out that overexpression of Capping Actin Protein, Gelsolin-like can promote tumor cell proliferation and drug resistance by activating the PI3K/AKT signaling pathway (Chi et al., 2019). In terms of epigenetic regulation, the dysfunction of the chromatin remodeling protein Alpha Thalassemia/Mental Retardation Syndrome X-linked is closely related to dual resistance to chemotherapy and endocrine therapy in breast cancer patients (Qian et al., 2023).

In addition to the aforementioned static genetic and protein biomarkers, TME-related factors have also increasingly attracted attention as dynamic biomarkers. For instance, under hypoxic conditions, the sustained high expression of HIF-1α is considered to be an important mechanism through which breast cancer cells enhance their adaptability to chemotherapy drugs (Mirzaei et al., 2023).

Overall, although research on biomarkers related to chemotherapy resistance in breast cancer has achieved certain results, it still faces challenges such as high complexity in molecular mechanisms and poor clinical translation efficiency. Future research will focus on how to accurately screen biomarkers with practical predictive value and intervention potential from a large number of candidate molecules, and advance their translation into clinical applications.

1.2.3 Clinical significance

Investigations into the mechanisms underlying chemoresistance in breast cancer are of great scientific and clinical importance. First, deciphering the molecular basis of chemoresistance offers critical theoretical support for identifying and developing innovative therapeutic targets. For example, inhibitors targeting drug efflux pumps, such as P-gp inhibitors, have demonstrated certain resistance-reversing effects in clinical trials (Kathawala et al., 2015). Second, investigating the molecular mechanisms of chemotherapy resistance in breast cancer can guide personalized treatment strategies. By analyzing the expression levels of resistance-related biomarkers in tumor tissues of breast cancer patients, chemotherapy drug sensitivity can be predicted, enabling the formulation of personalized treatment plans (Freitas et al., 2021). Furthermore, delving deeper into the molecular basis of chemoresistance in breast cancer could pave the way for groundbreaking drug development. A prime example lies in DNA repair pathway inhibitors—particularly poly(ADP-ribose) polymerase inhibitors—which have shown remarkable clinical promise for treating breast cancer, as highlighted by Lord and Ashworth's 2017 research (Lord and Ashworth, 2017). This approach not only sheds light on new treatment avenues but also offers hope for overcoming therapeutic roadblocks.

1.2.4 Research gap

Despite significant progress in the research on chemotherapy resistance in breast cancer in recent years, numerous challenges remain. Firstly, the molecular mechanisms underlying chemotherapy resistance in breast cancer are complex and diverse, involving multiple signaling pathways and molecular networks. Intervening in a single mechanism often fails to achieve the desired reversal of resistance. Therefore, future research needs to adopt a systems biology approach to

comprehensively analyze the molecular networks of chemotherapy resistance in breast cancer and identify key nodes for intervention (Kitano, 2002). Secondly, research on chemotherapy resistance in breast cancer requires more clinical samples and long-term follow-up data for support. Currently, most studies are based on cell lines and animal models, and their results may not align well with clinical realities. Thus, future research should prioritize the comprehensive collection and in-depth analysis of clinical specimens while integrating multi-omics approaches to systematically explore molecular biomarkers associated with chemoresistance in breast cancer (Collins and Varmus, 2015). Furthermore, the research on chemotherapy resistance in breast cancer should enhance international collaboration and data sharing. The molecular mechanisms of chemotherapy resistance vary among patients of different ethnicities and regions. Therefore, future research needs to foster international cooperation and establish multicenter, large-sample clinical research cohorts to promote the globalization of breast cancer chemotherapy resistance research (Cardoso et al., 2016).

In summary, chemoresistance remains a major obstacle in the clinical management of breast cancer, driven by complex and multifactorial molecular mechanisms. Elucidating these mechanisms is of considerable scientific and clinical significance. The integration of big data and machine learning (ML) technologies offers novel insights into this intricate problem; however, issues such as data heterogeneity, limited model interpretability, and translational gaps continue to pose significant challenges. Moving forward, a systems biology approach is essential to dissect the molecular networks underlying chemoresistance. Future efforts should also focus on enhancing the acquisition and analysis of clinical samples, fostering international collaboration and data sharing, and ultimately providing a robust

theoretical and experimental framework for the development of more effective therapeutic strategies.

1.3 Prognostic Models in Breast Cancer

Given the complexity and diversity of chemoresistance mechanisms and biomarkers, constructing predictive models that integrate multi-dimensional biological and clinical data is critical. In the context of precision medicine, constructing scientifically robust and effective prognostic models has increasingly become a focal area of research in breast cancer management. Traditional prognostic assessment methods rely primarily on clinicopathological parameters, such as tumor size, lymph node status, and molecular subtypes. However, these approaches have limited predictive capability regarding individualized treatment responses. Recent advancements in omics data integration and artificial intelligence (AI) have enabled researchers to develop biomarker-based prognostic models through ML algorithms, significantly contributing to the understanding of chemoresistance mechanisms and prediction of therapeutic outcomes.

Several studies have already explored integrating transcriptomic, proteomic, and biomarker expression data with clinical information to establish prognostic models for breast cancer (Yu et al., 2019; Yu et al., 2023; Kaur et al., 2024). For instance, Zhang et al. reviewed recent advances in ML–based prognostic modeling for breast cancer, emphasizing that the integration of multi-omics data is a key factor in improving predictive accuracy. Their review summarized multiple studies demonstrating that multi-omics–driven models outperform traditional prognostic tools in survival prediction and underscored the clinical applicability of such approaches for

developing robust, individualized treatment strategies across diverse patient cohorts (Zhang et al., 2024d).

Furthermore, ML-based prognostic models have demonstrated superior performance compared to traditional prognostic methods in larger clinical cohorts. Song et al. conducted a study involving 6,477 breast cancer patients and found that their Extreme Gradient Boosting predictive model achieved an AUC of 0.813 in predicting overall survival, substantially surpassing traditional systems such as the Nottingham Prognostic Index (Song et al., 2024b).

Recent research has also begun incorporating non-invasive biomarkers, including circulating tumor cells and circulating tumor DNA, into prognostic models, enabling dynamic prediction and monitoring of resistance progression in advanced breast cancer patients. Gerratana et al. demonstrated that integrating ML algorithms with liquid biopsy data significantly enhanced sensitivity and real-time capability for assessing disease progression risk, thereby facilitating early interventions for high-risk populations (Gerratana et al., 2024).

Although prognostic models have benefited from ongoing technological advancements, their widespread clinical adoption is still hindered by key challenges, including limited generalizability, poor interpretability, and data heterogeneity. To overcome these barriers, future research should focus on standardizing model development, performing multi-center external validations, and ensuring alignment with clinical workflows to enable truly personalized prediction and therapeutic decision-making.

1.4 Big Data

In recent years, with the rise of BDA and precision medicine, there has been a profound shift in the understanding of the biological characteristics and treatment response of breast cancer. By integrating BDA technology into the GBCI, researchers and clinical doctors can strengthen existing breast cancer prevention, diagnosis, and management strategies, gaining a deeper understanding of the molecular mechanisms driving the development and progression of breast cancer, thereby advancing the goal of precision oncology and personalized medicine. BDA has become a revolutionary tool in healthcare, providing unprecedented opportunities to leverage diverse data types to understand disease mechanisms, treatment responses, and patient outcomes (Kaur et al., 2018; Galetsi et al., 2019). However, despite the increasing research related to precision therapy for breast cancer, studies involving big data on breast cancer remain insufficient. In recent years, with the rapid advancement of precision medicine technologies, breast cancer research has entered a new era centered around data. Traditional research methods primarily rely on histological classification, single biomarkers, and clinical trials. Although certain progress has been made in molecular classification, prognosis assessment, and the optimization of treatment strategies, limitations still exist. For example, the high heterogeneity of breast cancer makes it difficult for traditional treatment models to meet the demands of precision medicine. Patients with the same molecular classification may exhibit different treatment responses or even develop resistance when receiving the same treatment regimen. Therefore, uncovering the molecular mechanisms of breast cancer based on large-scale data and further refining personalized treatment strategies has become a key challenge in the field of breast cancer research (Xu et al., 2019).

In the context of the digital age, the integration of BDA and AI has brought unprecedented new opportunities for precision medicine in breast cancer. BDA technologies can integrate and analyze multi-omics data, medical imaging information, electronic health records, and other clinical data sources to explore the occurrence, development, and resistance mechanisms of breast cancer from a holistic perspective. Meanwhile, the application of ML in AI technologies has demonstrated remarkable performance in disease classification, risk prediction, survival analysis, and personalized treatment recommendation. Through training on large-scale biomedical datasets, AI algorithms can automatically identify potential biomarkers, predict patient responses to specific treatments, and establish highly accurate predictive models for early screening, molecular classification, resistance prediction, and optimization of personalized care in breast cancer, thus promoting the intelligent and efficient development of precision oncology (Gandhi and Kumar, 2024). However, despite the increasing research on precise treatment for breast cancer, studies combining big data and AI are still limited. Therefore, it is necessary to integrate a large amount of clinical research data with advanced technologies to explore new strategic methods that enhance cancer diagnosis and efficacy evaluation, ultimately achieving the goal of precision tumor management. This study aims to systematically explore the applications of big data and AI in precision medicine for breast cancer, with a focus on analyzing their roles in predicting chemotherapy resistance, assessing survival prognosis, and formulating personalized treatment plans, thereby providing valuable insights and targeted guidance for related research, promoting the continuous advancement of precision oncology and the innovation of personalized patient care, and contributing positively to the rapid development of precision medicine in breast cancer.

1.4.1 Foundations of big data

The rapid development of information and communication technology has given rise to significant changes in the digital domain, forcing most industries to adapt to this digital trend and consider digitalization an indispensable requirement (Queiroz et al., 2021). The core of digital transformation lies in digitization, which begins with the digitalization process, i.e., the conversion of analogue signals (physical behaviours and observations) into digital form using digital technology (Legner et al., 2017). Subsequently, digitization involves utilizing and manipulating this digitalized information, generating impacts from organizational and societal perspectives. At the core of digitization lies BDA, which processes large datasets and transforms them into actionable knowledge for decision-making. Big data originates from massive datasets resulting from information exchange between multiple systems (Khanra et al., 2020) and possesses the following seven "V" characteristics (Kitchin, 2014; Gandomi and Haider, 2015; Debattista et al., 2015):

1. Volume (size): Big data involves managing and analyzing massive amounts of information, typically ranging from several terabytes (TB) to several exabytes (EB), requiring substantial processing and storage capabilities.

2. Variety (complexity): Big data is presented in various forms, including structured data (e.g., databases), semi-structured data (e.g., logs and XML), and unstructured data (e.g., social media posts and video content).

3. Velocity (speed): Big data is generally generated in real-time or near-real-time, necessitating instant analysis and processing to derive valuable insights.

4. Veracity (quality): The veracity of big data refers to the accuracy and reliability of data, ensuring the trustworthiness and quality of data sources.

5. Variability (flexibility): The content, scale, and speed of big data are constantly changing, requiring adaptability for effective analysis and processing.

6. Value (knowledge): The ultimate goal of big data is to extract useful information and insights from datasets to provide robust support and directional guidance for decision-making and innovation activities.

7. Valence (connectedness): Another characteristic of big data is the connectivity between datasets.

Due to the significant characteristics of large data volume and complexity, traditional processing methods have become extremely difficult. Recently, significant achievements have been made in analyzing and mining massive amounts of data by humans. The entire process of in-depth interpretation, mining, and analysis of these large-scale, high-dimensional, diverse, and rapidly generated data sets is collectively called BDA (Saggi and Jain, 2018).

1.4.2 Big data-driven precision medicine: integration of multi-omics and methodologies

Big data-driven precision medicine has emerged in the current medical science field. The key in this field is to achieve personalized medical treatment and smarter healthcare services by analyzing large amounts of data and adopting innovative technologies. The combination of multi-omics plays a decisive role by integrating rich omics data such as genomics, transcriptomics, proteomics, and metabolomics to reveal the molecular mechanisms of disease development, providing critical support for precision medicine (Olivier et al., 2019). Over the past decade, due to the rapid advancements in short-read sequencing technologies (such as the Illumina platform) and long-read sequencing technologies (such as PacBio and Oxford Nanopore platforms), there has been a deeper understanding of the complex interactions within

biological systems and the molecular and physical phenomena they entail (Bayega et al., 2018). Next-generation sequencing (NGS) technologies are being used to generate large datasets for interpreting human genetics and the regulatory mechanisms at the molecular level. This includes identifying single nucleotide polymorphisms through genome-wide association studies, exploring disease progression through transcriptome analysis and full transcriptome-wide associations, and discovering the interactions and impacts between food and diseases through microbiome-wide association studies. Since 2013, at least 14 countries have launched government-funded genomics medicine initiatives, with total investments reaching up to $4 billion. However, this amount is relatively small compared to the Precision Medicine Initiative in China. The initiative aims to sequence the genomes of 100 million individuals by 2030, with a funding of $9.2 billion (Stark et al., 2019).

The cost of NGS has dropped from millions to a few thousand dollars (Service, 2006), which will help comprehensively generate genetic data for all diseases, including cancer and rare diseases. It is encouraging that comprehensive digital catalogs documenting human genetic variations across diverse populations (Auton et al., 2015) and cancer-specific genomic alterations, such as The Cancer Genome Atlas (TCGA; https://www.cancer.gov/ccg/research/genome-sequencing/tcga) (Tomczak et al., 2015), have been established and made publicly accessible. Furthermore, significant breakthroughs have been made in collecting proteomic, metabolomic, and other data from millions of patients through techniques such as mass spectrometry, nuclear magnetic resonance, and imaging (Ma and Fernández, 2024). Combining this data with mathematical algorithms can lead to more effective classification based on specific phenotypes. Using big data has also offered unprecedented opportunities for comprehending pharmacology and toxicology. For instance, through integrated multi-

omics analysis, Lv et al. (Lv et al., 2024) demonstrated how cadmium exposure influences the survival and function of granulosa cells via multiple signalling pathways. This data was extensively stored and analyzed, laying a data foundation for future toxicity evaluations. This research approach, based on big data and integrated multi-omics, enables researchers to understand complex pathological mechanisms and the effects of drugs on cells at a broader molecular level, thereby providing a foundation for personalized treatment strategies in precision medicine. Therefore, the critical task for the future is to integrate this omics data and other relevant data and further perform genetic association and AI training for predictive analysis based on a multi-omics approach, thus revealing the molecular principles behind disease progression and treatment effects. In the future, these multi-omics technologies can be developed and transformed into user-friendly and reliable tools for routine clinical operations, thus achieving support for clinical decision-making and personalized treatment as the goal for all patients with chronic diseases, especially those diagnosed with cancer.

In addition, big data-driven precision medicine also relies on advanced data science methods. Anomaly detection is a critical challenge when processing large-scale medical data. The method proposed by Li et al. (Li et al., 2023a) demonstrates how to overcome these challenges through BDA, ensuring the reliability and accuracy of detection results. By processing, cleaning, transforming, exploring, dimensionality reduction, pattern searching, visualization, and reporting meaningful information and insights can be extracted from massive multi-omics data, thus optimizing the healthcare system (Dash et al., 2019). However, despite significant advancements in multi-omics integration technologies, some challenges still need to be addressed, such as poor data collection quality, high data heterogeneity, low data reliability, and lack of adequate data management and annotation. These issues will impact clinical

decision-making. Despite the challenges, many successful cases have shown that using advanced analytical technologies, such as ML and AI, is very effective in medical applications for cancer (especially breast cancer), as shown in Table 1.1. Additionally, programming languages such as R and Python are used to write scripts, algorithms, and software; big data tools such as Hadoop and Apache Spark, as well as high-performance computing clusters accessed through grid computing infrastructure, are significant contributors to the digitization of multidimensional medical data (Dash et al., 2019). Therefore, future research directions should include optimizing data integration and analysis algorithms, expanding the scale of samples, and establishing more rigorous quality control standards.

Table 1.1: Application of digitisation in healthcare sectors by focusing on cancers

Type	Application	Type of Cancer	Description	Reference
Early Detection	Detection of breast cancer using full-field digital mammography (FFDM)	Breast Cancer	FFDM consists of a dedicated electronic detector system that captures and displays the x-ray signals on a computer rather than film. This technology separates the process of image acquisition from that of image display, and because the signal for each pixel is digitally stored, the same image can be manipulated to show different brightness and contrast combinations, allowing radiologists to more easily see through denser tissues. Softcopy display of digital mammography allows adjustment of magnification, brightness, and contrast after the mammography examination, thus enabling a more detailed examination of questionable areas without necessarily requiring additional imaging.	(Chuan *et al.*, 2007)
	Screening high-risk women using Magnetic Resonance Imaging (MRI)	Breast Cancer	MRI is a non-invasive and non-ionizing radiation technology for early detection of breast cancer. Some studies have shown greater than 90% sensitivity in detecting lesions using contrast-enhanced MRI. Further improvements in specificity and optimal acquisition protocols, examination and interpretation standards are ongoing areas of investigation. Digitization has enabled the use of MRI for early detection of breast cancer without using ionizing radiation.	(Chuan *et al.*, 2007)
Diagnosis	Classification of benign-malignant patterns in digital mamograms	Breast Cancer	Computer-aided diagnosis (CAD) based on support vector machine recursive feature elimination, correlation bias reduction algorithm, and Q similarity-based learning algorithm is applied to real clinical mammograms and resulted in an accuracy of 98.16%, a sensitivity of 98.63%, specificity of	(Eltrass and Salama, 2020)

			97.80%, and computational time of 2.2 s for breast cancer diagnosis.	
	Classification of melanoma (fatal skin cancer) and nevus (non-cancer, benign)	Skin Cancer	An intelligent data-driven system with image processing technique utilises Gaussian filter, improved K-mean clustering and Support vector machine to classify skin cancer into melanoma and nevus with 96% accuracy.	(Khan *et al.*, 2019)
Prognosis	Machine learning-based data mining approach for the prediction of breast cancer survivability	Breast Cancer	With a large dataset of >200,000 cases, a decision tree with 10-fold cross-validation methods showed a prediction accuracy of 93.6% accuracy, providing relative prediction ability for cancer prognosis.	(Delen *et al.*, 2005)
Treatment	Deep-learning assisted auto-digitization method for interstitial needles	Gynecological Cancer	A deep-learning assisted auto-digitization method for interstitial needles was used in the treatment planning of 3D computed tomography image-based interstitial high dose-rate brachytherapy of gynecological cancer. The digitization process took ~5 min to complete, and the achieved accuracy and efficiency made the method clinically attractive.	(Jung *et al.*, 2019)

1.5 Machine Learning (ML)

1.5.1 Fundamental concepts and development overview

AI, as a key subfield of computer science, focuses on exploring how to equip computers with the ability to simulate or enhance human intelligence, enabling them to perform a variety of tasks that typically require human intelligence, including, but not limited to, learning, logical reasoning, problem-solving, language parsing, visual perception, and decision-making (Suryawanshi and Singh, 2024). AI technology has been widely applied in various fields, including military, healthcare, education, industrial manufacturing, and commercial services (Shrivastava, 2024). According to the definition by the American Association for AI, AI refers to "the ability of computers to imitate, execute, and ultimately extend human intelligence through self-learning and knowledge accumulation" (Suryawanshi and Singh, 2024). Compared to traditional computer systems, AI exhibits greater autonomy, intelligence, robustness, and automation. This definition encompasses not only the functions of computer hardware and algorithms but also emphasizes the importance of data in intelligent systems, enabling AI to demonstrate flexible adaptability and reasoning capabilities in different contexts (Salehi and Burgueño, 2018).

With the rapid development of information technology, ML has become an important branch of AI and is at the core of AI. ML enables computer systems to automatically extract patterns and rules from data, allowing them to learn from experience and improve task performance without explicit programming (Jordan and Mitchell, 2015). According to Mitchell, a widely cited definition states: "A computer program is said to learn from experience E with respect to some task T and performance measure P, if its performance on T, as measured by P, improves with

experience E" (Mitchell, 1997). This definition highlights three essential elements of ML—task, experience, and performance measure—reflecting its essence in deriving knowledge from data and making predictions.

The foundational idea of ML can be traced back to the fields of statistics and data analysis, using mathematical models to reveal underlying patterns and rules within data (Murphy, 2012). With the continuous advancement of computer technology, particularly the improvement of computational power and the emergence of massive datasets, ML has gradually shifted from theoretical research to practical applications, becoming one of the core driving forces of technological development in the 21st century. The research areas of ML are quite extensive, and its methods and techniques can generally be categorized into the following types based on the learning approach: supervised learning, unsupervised learning, and reinforcement learning. Figure 1.1 illustrates the relationship between ML and AI, as well as the methods of ML.

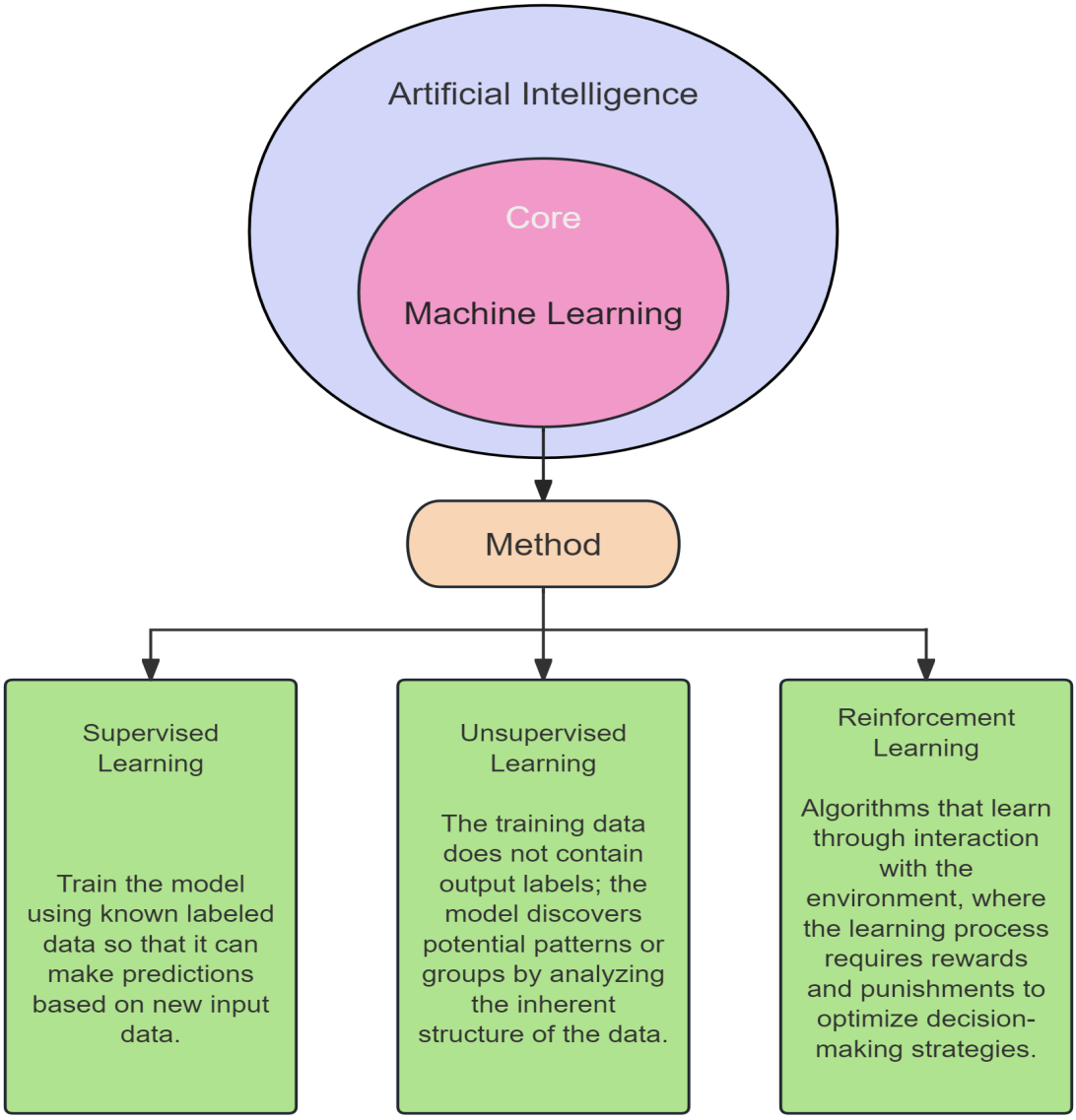

Figure 1.1 Diagram of the relationship between ML and AI and an overview of ML methods.

The core of ML lies in extracting valuable information from large amounts of data and learning and predicting through appropriate algorithms. With the rise of deep learning, many new technologies and methods have continuously emerged, significantly advancing the field of ML. ML has been widely applied across various industries, particularly in the field of biomedicine, where it is profoundly impacting disease diagnosis, personalized medicine, and drug development (Ozer et al., 2020). Biomedical research involves complex data, including genomic data, imaging data, and clinical data (Peng et al., 2021a). Traditional manual analysis methods often

struggle to effectively handle such high-dimensional and large-scale data, while ML can utilize nonlinear models to process these data types, uncovering underlying patterns and trends that facilitate advancements in medical research and transformative clinical practices.

Table 1.2 illustrates the wide application of ML in oncology. Through in-depth analysis of biomedical data, ML not only enhances diagnostic efficiency and optimizes treatment plans but also drives the development of drug discovery and personalized medicine (Huang et al., 2021). With ongoing technological advancements, ML will continue to play a crucial role in the biomedical field, steering medical science toward more intelligent and precise directions. In summary, ML holds vast application prospects and tremendous potential in the field of biomedicine.