# Evolutionary and Deep Mining Models for Effective Biomarker Discovery

Abeer Hamza Abd Alzubaidi

School of Science and Technology

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of

*Doctor of Philosophy*

September 2019

## Copyright Statement

This thesis is dedicated to my parents
With love and eternal appreciation.
I hope you’re proud of your little girl
I can see your smile from Heaven.

# Acknowledgements

I would like to express my deepest appreciation to my supervisory team. Im deeply indebted to Dr Jonathan Tepper for the dedicated support and guidance through the process of researching and writing this thesis. Dr Tepper continuously provided encouragement and was always willing to assist in any way he could throughout the research project. I am also grateful to Dr Tepper for insightful suggestions, which have contributed greatly to the improvement of the thesis. I would like to express my deepest gratitude to Prof Ahmad Lotfi for the regular advice and encouragement. The door to Prof Lotfi office was always open and he was prepared to sit and listen to me. My sincere thanks are also extended to Dr Benjamin Inden for his supervisory role and support throughout this research project. I take this opportunity to express my sincere appreciation to Dr Georgina Cosma, Prof David Brown, and Prof Graham Pockley for their support and encouragement throughout the process.

I gratefully acknowledge the funding received towards my PhD from the Ministry of Higher Education and Scientific Research in Iraq. I would like to thank my friends Dr Maria Bisele and Dr Edwin Abdurakman for the wonderful times we shared. Big thank you to my sister, Ms Zaineb Alzubaidi who was always there for me and gave me lots of support. The people with the greatest indirect contribution to this work are my husband and my children. I owe my deepest gratitude to my husband Captain Hamid Alhasnawi for being extremely supportive and understanding of my ambition. My children: Shahad, Ali, and Fatimah, you have inspired me to aim for extraordinary and made me more fulfilled. I appreciate all your patience and support, love you.

# Abstract

With the advent of high-throughput biology, large amounts of molecular data are available for purposeful analysis and evaluation. Extracting relevant knowledge from high-throughput biomedical datasets has become a common goal of current approaches to personalised cancer medicine and understanding cancer genotype and phenotype. However, the datasets are characterised by high dimensionality and relatively small sample sizes with small signal-to-noise ratios. Extracting and interpreting relevant knowledge from such complex datasets therefore remains a significant challenge for the fields of machine learning and data mining. This is evidenced by the limited success these methods have had in detecting robust and reliable biomarkers for cancers and other complicated diseases. This could also explain the lack of finding generic biomarkers among the identified published genes for identical diseases or clinical conditions.

This thesis proposes and evaluates the efficacy of two novel feature mining models established on the basis of the evolutionary computation and deep learning paradigms to position and solve biomarker discovery as an optimisation problem. Deep learning methods lack the transparency and interpretability found in the evolutionary paradigm. To overcome the inherent issue of poor explanatory power associated with the deep learning, this research also introduces a novel deep mining model that helps to deconstruct the internal state of such deep learning models to reveal key determinants underlying its latent representations to aid feature selection. As a result, salient biomarkers for breast cancer and the positivity of the Estrogen and Progesterone receptors are discovered robustly and validated reliably across a wide range of independently generated breast cancer data samples.

# Publications

The following publications have been published as a direct result of this thesis:

**Refereed Journal Papers**

Abeer Alzubaidi, Jonathan Tepper, Ahmad Lotfi. "A novel deep mining model for effective knowledge discovery from omics data". Artificial Intelligence in Medicine, page 101821, 2020.

Abeer Alzubaidi, Jonathan Tepper, Benjamin Inden, Ahmad Lotfi, "mRNA signatures for predicting oestrogen and progesterone receptor status in Breast Cancer: Findings from a novel deep mining approach", In Press.

**Refereed Conference Papers**

Abeer Alzubaidi, Georgina Cosma, David Brown, A. Graham Pockley, "Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information", 2016 International Conference on Interactive Technologies and Games (ITAG), pp. 70-76, 2016.

Abeer Alzubaidi, Georgina Cosma, David Brown, A. Graham Pockley, "A new hybrid global optimization approach for selecting clinical and biological features that are relevant to the effective diagnosis of ovarian cancer", Computational Intelligence (SSCI) 2016 IEEE Symposium Series on, pp. 1-8, 2016.

Abeer Alzubaidi, Georgina Cosma, "A multivariate feature selection framework for high dimensional biomedical data classification”, 2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1-8, 2017.

Abeer Alzubaidi, “Challenges in Developing Prediction Models for Multi-modal High-Throughput Biomedical Data”, Proceedings of SAI Intelligent Systems Conference, pp. 1056-1069, 2018.

# Contents

**Publications** v

**Contents** vii

**List of Figures** xi

**List of Tables** xxi

**1 Introduction** 1

1.1 Aims and Objectives . . . . . . . . . . . . . . . . . . . . . . . . . 6

1.2 Contribution of the Thesis . . . . . . . . . . . . . . . . . . . . . 7

1.3 Thesis Outline . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

**2 Literature Review** 14

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

2.2 Breast Cancer . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

2.3 Omics Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17

2.4 Statistical Techniques for Biomarker Discovery . . . . . . . . . . . 19

2.5 Machine Learning Methods for Biomarker Discovery . . . . . . . . 22

2.5.1 Support Vector Machine . . . . . . . . . . . . . . . . . . . 24

2.5.2 Bagging Decision Tree . . . . . . . . . . . . . . . . . . . . 26

2.6 Feature Mining Approaches for Biomarker Discovery . . . . . . . 29

2.6.1 Evaluation Measurements . . . . . . . . . . . . . . . . . . 30

2.6.2 Search Methods . . . . . . . . . . . . . . . . . . . . . . . . 32

2.7 Deep Learning Methods for Biomarker Discovery . . . . . . . . . 34

2.7.1 Unsupervised Pre-training Approach . . . . . . . . . . . . 36

2.7.2 Interpretation Methods for Deep Learning . . . . . . . . . . 37
2.8 Approaches for Validating Predictive Performance . . . . . . . . . 39
2.8.1 Hold Out Validation Approach . . . . . . . . . . . . . . . 40
2.8.2 Leave One Out Cross Validation Approach . . . . . . . . . 40
2.8.3 k-fold Cross Validation Approach . . . . . . . . . . . . . 41
2.9 Metrics for Estimating Predictive Performance . . . . . . . . . . 42
2.10 Internal and External Validation . . . . . . . . . . . . . . . . . . . 44
2.11 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . 45

**3 Datasets and Experimental Methodology 48**
3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . 48
3.2 Filtering Methods . . . . . . . . . . . . . . . . . . . . . . . . 49
3.3 Breast Invasive Carcinoma Datasets . . . . . . . . . . . . . . . . . 50
3.3.1 Breast Invasive Carcinoma (TCGA, Nature 2012) . . . . . 51
3.3.2 Breast Invasive Carcinoma (TCGA, Cell 2015) . . . . . . . 53
3.3.3 Breast Invasive Carcinoma (TCGA, Provisional) . . . . . . 54
3.4 The Integrated Breast Invasive Carcinoma Datasets . . . . . . . . 55
3.4.1 The Integrated Datasets with ER groups . . . . . . . . . . 56
3.4.1.1 NCP1 Dataset . . . . . . . . . . . . . . . . . . . 56
3.4.1.2 NCP2 Dataset . . . . . . . . . . . . . . . . . . . 56
3.4.1.3 NCP3 Dataset . . . . . . . . . . . . . . . . . . . 57
3.4.2 The Integrated Datasets with PR groups . . . . . . . . . . 58
3.4.2.1 NC Dataset . . . . . . . . . . . . . . . . . . . . . 58
3.4.2.2 CN Dataset . . . . . . . . . . . . . . . . . . . . . 59
3.4.2.3 NP Dataset . . . . . . . . . . . . . . . . . . . . . 59
3.4.2.4 PN Dataset . . . . . . . . . . . . . . . . . . . . . 60
3.4.2.5 CP Dataset . . . . . . . . . . . . . . . . . . . . . 61
3.4.2.6 PC Dataset . . . . . . . . . . . . . . . . . . . . . 62
3.5 METABRIC Breast Cancer Dataset . . . . . . . . . . . . . . . . . 62
3.6 Ovarian Cancer Dataset . . . . . . . . . . . . . . . . . . . . . 64
3.7 Experimental Methodology . . . . . . . . . . . . . . . . . . . . 64
3.7.1 Area Under the ROC Curve . . . . . . . . . . . . . . . . . 64
3.7.2 Stability and Generalisability . . . . . . . . . . . . . . . . 66

3.8 Discussion . . . 67

**4 Evolutionary Mining Model 68**

4.1 Introduction . . . 68

4.2 Genetic Algorithm . . . 69

4.3 Experimental Design of the Evolutionary Mining Model . . . 70

4.3.1 Univariate Approach . . . 70

4.3.2 Initial Population . . . 72

4.3.3 Multivariate Approach . . . 73

4.3.4 Ranking Scaling . . . 75

4.3.5 Selection . . . 75

4.3.6 Reproduction . . . 76

4.4 Results and Discussion . . . 79

4.4.1 Results and Discussion of Ovarian Cancer Dataset . . . 79

4.4.2 Results and Discussion of METABRIC Dataset . . . 81

4.4.3 Results and Discussion of Breast Invasive Carcinoma Datasets 83

4.4.3.1 ER Groups . . . 84

4.4.3.2 PR Groups . . . 86

4.4.4 Results and Discussion of Integrated Breast Invasive Carcinoma Datasets . . . 86

4.4.4.1 ER Groups . . . 87

4.4.4.2 PR Groups . . . 88

4.5 Discussion . . . 90

**5 Deep Mining Model 92**

5.1 Introduction . . . 92

5.2 Experimental Design of the Deep Mining Model . . . 94

5.2.1 Auto-Encoder . . . 95

5.2.2 Sparse Compressed Auto-Encoder . . . 96

5.2.3 Stacked Sparse Compressed Auto-Encoder . . . 98

5.2.4 A New Weight Interpretation Method . . . 100

5.3 Results and Discussion . . . 102

5.3.1 Results and Discussion of Ovarian Cancer Dataset . . . 103

5.3.2 Results and Discussion of METABRIC Dataset . . . . . . 106
5.3.3 Results and Discussion of Breast Invasive Carcinoma Datasets 111
5.3.3.1 ER Groups . . . . . . . . . . . . . . . . . . . . . 111
5.3.3.2 PR Groups . . . . . . . . . . . . . . . . . . . . . 116
5.3.4 Results and Discussion of Integrated Breast Invasive Carcinoma Datasets . . . . . . . . . . . . . . . . . . . . . . 120
5.3.4.1 ER Groups . . . . . . . . . . . . . . . . . . . . . 121
5.3.4.2 PR Groups . . . . . . . . . . . . . . . . . . . . . 125
5.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . 129

**6 Biomarkers and Bioinformatics 131**
6.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . 131
6.2 Discovered Biomarkers with HP Weight for ER . . . . . . . . . . 132
6.3 Discovered Biomarkers with HN Weight for ER . . . . . . . . . . 145
6.4 Discovered Biomarkers with HP Weight for PR . . . . . . . . . . 159
6.5 Discovered Biomarkers with HN Weight for PR . . . . . . . . . . 167
6.6 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . 175

**7 Conclusion and Future Work 176**
7.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . 176
7.2 Concluding Remarks . . . . . . . . . . . . . . . . . . . . . . 177
7.2.1 Filtering Methods . . . . . . . . . . . . . . . . . . . . . 177
7.2.2 Evolutionary Mining Model . . . . . . . . . . . . . . . . 178
7.2.3 Deep Mining Model . . . . . . . . . . . . . . . . . . . . 179
7.2.4 The weight Interpretation Method . . . . . . . . . . . . . 180
7.2.5 Discovered Biomarkers . . . . . . . . . . . . . . . . . . . 181
7.3 Directions for Future Work . . . . . . . . . . . . . . . . . . . 184

**Appendix A 186**

**Appendix B 189**

**References 208**

# List of Figures

2.1 The central dogma of molecular biology [121] and the types of omics data generated from each layer of dogma. . . . . . . . . . . 18

2.2 An example dataset described by two genes of a linear separator that maximises the margin between positive samples and negative samples, the red dots on the right side represent the positive group; and the blue dots on the left side represent the negative group. There are three support vectors. One point of the positive group on the right dashed line, and two negative samples on the left dashed line. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25

2.3 SVM soft-margin allows some data points to be misclassified or within the margin through slack variables $\xi_i$. . . . . . . . . . . . . 26

2.4 An illustration of a decision tree constructed using a gene expression dataset of four genes. . . . . . . . . . . . . . . . . . . . . . . . 27

2.5 An illustration of k-fold cross validation approach. . . . . . . . . . 41

3.1 The description of (Nature 2012) dataset showing the unification of clinical data and mRNA expression data. . . . . . . . . . . . . 52

3.2 The description of (Cell 2015) dataset showing the unification of clinical data and mRNA expression data. . . . . . . . . . . . . . . 53

3.3 The description of (Provisional) dataset showing the unification of clinical data and mRNA expression data. . . . . . . . . . . . . . . 55

3.4 The description of NCP1 dataset showing the unification of (Nature 2012), (Cell 2015), and (Provisional) datasets. . . . . . . . . 57

3.5 The description of NCP2 dataset showing the unification of (Nature 2012), (Cell 2015), and (Provisional) datasets. . . . . . . . . 57

3.6 The description of NCP3 dataset showing the unification of (Nature 2012), (Cell 2015), and (Provisional) datasets. . . . . . . . . 58
3.7 The description of NC dataset showing the unification of (Nature 2012) and (Cell 2015) datasets. . . . . . . . . . . . . . . . . . . 59
3.8 The description of CN dataset showing the unification of (Cell 2015) and (Nature 2012) datasets. . . . . . . . . . . . . . . . . . 60
3.9 The description of NP dataset showing the unification of (Nature 2012) and (Provisional) datasets. . . . . . . . . . . . . . . . . . 60
3.10 The description of PN dataset showing the unification of (Provisional) and (Nature 2012) datasets. . . . . . . . . . . . . . . . . . 61
3.11 The description of CP dataset showing the unification of (Cell 2015) and (Provisional) datasets. . . . . . . . . . . . . . . . . . 61
3.12 The description of PC dataset showing the unification of (Provisional) and (Cell 2015) datasets. . . . . . . . . . . . . . . . . . . . 62
3.13 The description of (METABRIC) dataset showing the unification of clinical data and mRNA expression data. . . . . . . . . . . . 63
3.14 An example of Receiver Operating Characteristics (ROC) curve. 65

4.1 Steps of the experimental design of the evolutionary mining model. 71
4.2 An illustration of the fitness score of each individual. . . . . . . 74
4.3 Histogram of parent individuals. . . . . . . . . . . . . . . . . . . . . 76
4.4 An illustration of the average distance between individuals at each generation. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 77
4.5 An illustration of the algorithm's evolution over generations. . . . 78
4.6 Scatter plots matrix of the stable predictors (index) of ovarian cancer dataset. . . . . . . . . . . . . . . . . . . . . . . . . . . . 80
4.7 Scatter plots matrix of the stable predictors of METABRIC dataset with ER groups. . . . . . . . . . . . . . . . . . . . . . . . . . . . 82
4.8 Scatter plots matrix of the stable predictors of METABRIC dataset with PR groups. . . . . . . . . . . . . . . . . . . . . . . . . . . . 83
4.9 Scatter plots matrices of the generic biomarkers of the breast invasive carcinoma datasets with ER and PR groups. . . . . . . . 85

4.10 Scatter plots matrices of the generic biomarkers of the integrated datasets with ER groups. . . . . . . . . . . . . . . . . . . . . . . . 88

4.11 Scatter plots matrices of the generic biomarkers of the integrated datasets with PR groups. . . . . . . . . . . . . . . . . . . . . . . . 89

5.1 An illustration of the validation performance of the SSCAE. . . . 99

5.2 Histogram of z-scores of the weight vector. . . . . . . . . . . . . . 101

5.3 Scatter plots matrix of the stable predictors (index) with HP weight of ovarian cancer dataset. . . . . . . . . . . . . . . . . . . . . . . 104

5.4 Scatter plots matrix of the stable predictors (index) with HN weight of ovarian cancer dataset. . . . . . . . . . . . . . . . . . . . . . . 105

5.5 Scatter plots matrix of the stable predictors with HP weight of METABRIC dataset with ER groups. . . . . . . . . . . . . . . . . . 107

5.6 Scatter plots matrix of the stable predictors with HN weight of METABRIC dataset with ER groups. . . . . . . . . . . . . . . . . . 108

5.7 Scatter plots matrix of the stable predictors with HP weight of METABRIC dataset with PR groups. . . . . . . . . . . . . . . . . . 109

5.8 Scatter plots matrix of the stable predictors with HN weight of METABRIC dataset with PR groups. . . . . . . . . . . . . . . . . . 110

5.9 Scatter plots matrices of the generic biomarkers with HP weight of the breast invasive carcinoma datasets with ER groups. . . . . 113

5.10 Scatter plots matrices of the generic biomarkers with HN weight of the breast invasive carcinoma datasets with ER groups. . . . . 114

5.11 Scatter plots matrices of the generic biomarkers with HP weight of the breast invasive carcinoma datasets with PR groups. . . . . 118

5.12 Scatter plots matrices of the generic biomarkers with HN weight of the breast invasive carcinoma datasets with PR groups. . . . . 119

5.13 Scatter plots matrices of the generic biomarkers with HP weight of the integrated datasets with ER groups. . . . . . . . . . . . . . 123

5.14 Scatter plots matrices of the generic biomarkers with HN weight of the integrated datasets with ER groups. . . . . . . . . . . . . . 124

5.15 Scatter plots matrices of the generic biomarkers with HP weight of the integrated datasets with PR groups. . . . . . . . . . . . . . 126

5.16 Scatter plots matrices of the generic biomarkers with HN weight of the integrated datasets with PR groups. . . . . . . . . . . . . . 127

6.1 Scatter plot of ESR1 and GFRA1 of (Nature 2012) dataset with ER groups, illustrating that the observation 306 from the ER+ group has high expression levels of (ESR1, GFRA1), which are $(4.7657, 4.512)$ in comparison to the observation 235 from the ER- group, which has low expression levels of (ESR1, GFRA1), which are $(-4.5636, -4.3411)$. . . . . . . . . . . . . . . . . . . . . . . . . 133

6.2 Scatter plot of AGR3 and SIAH2 of (Nature 2012) dataset with ER groups, illustrating that the observation 406 from the ER+ group has high expression levels of (AGR3, SIAH2), which are $(5.5085, 2.9504)$ in comparison to the observation 233 from the ER- group, which has low expression levels of (AGR3, SIAH2), which are $(-3.9565, -1.3913)$. . . . . . . . . . . . . . . . . . . . . . . . . 135

6.3 Scatter plot of C6orf97 and SLC39A6 of (Nature 2012) dataset with ER groups, illustrating that the observation 142 from the ER+ group has high expression levels of (C6orf97, SLC39A6), which are $(3.1637, 4.6489)$ in comparison to the observation 187 from the ER- group, which has low expression levels of (C6orf97, SLC39A6), which are $(-3.5103, -2.0668)$. . . . . . . . . . . . . 137

6.4 Scatter plot of ANXA9 and NAT1 of (Nature 2012) dataset with ER groups, illustrating that the observation 142 from the ER+ group has high expression levels of (ANXA9, NAT1), which are $(3.698, 6.6264)$ in comparison to the observation 13 from the ER- group, which has low expression levels of (ANXA9, NAT1), which are $(-3.108, -1.3075)$. . . . . . . . . . . . . . . . . . . . . . . . . 139

6.5 Scatter plot of CA12 and SCUBE2 of (Nature 2012) dataset with ER groups, illustrating that the observation 200 from the ER+ group has high expression levels of (CA12, SCUBE2), which are $(3.8954, 5.2376)$ in comparison to the observation 373 from the ER- group, which has low expression levels of (CA12, SCUBE2), which are (-3.15, -3.1562). . . . . . . . . . . . . . . . . . . . . . . . . 140

6.6 Scatter plot of EVL and PCP2 of (Nature 2012) dataset with ER groups, illustrating that the observation 89 from the ER+ group has high expression levels of (EVL, PCP2), which are $(3.3006, 2.5823)$ in comparison to the observation 105 from the ER- group, which has low expression levels of (EVL, PCP2), which are $(-2.0036, -0.67983)$. 142

6.7 Scatter plot of FSIP1 and GATA3 of (Nature 2012) dataset with ER groups, illustrating that the observation 12 from the ER+ group has high expression levels of (FSIP1, GATA3), which are $(7.2055, 2.6464)$ in comparison to the observation 175 from the ER- group, which has low expression levels of (FSIP1, GATA3), which are $(-0.949, -4.9429)$. . . . . . . . . . . . . . . . . . . . . . . . . 143

6.8 Scatter plot of IGFALS and LRRC56 of (Nature 2012) dataset with ER groups, illustrating that the observation 448 from the ER+ group has high expression levels of (IGFALS, LRRC56), which are $(5.4277, 2.4235)$ in comparison to the observation 512 from the ER- group, which has low expression levels of (IGFALS, LRRC56), which are $(-1.2985, -1.8847)$. . . . . . . . . . . . . . . . . . . . . . 145

6.9 Scatter plot of VGLL1 and PPP1R14C of (Nature 2012) dataset with ER groups, illustrating that the observation 93 from the ER+ group has low expression levels of (VGLL1, PPP1R14C), which are $(-0.54387, -4.3153)$ in comparison to the observation 81 from the ER- group, which has high expression levels of (VGLL1, PPP1R14C), which are $(5.0809, 2.7428)$. . . . . . . . . . . . . . . 146

6.10 Scatter plot of PROM1 and PSAT1 of (Nature 2012) dataset with ER groups, illustrating that the observation 336 from the ER+ group has low expression levels of (PROM1, PSAT1), which are $(-4.1265, -6.553)$ in comparison to the observation 478 from the ER- group, which has high expression levels of (PROM1, PSAT1), which are $(5.5998, -0.1715)$. . . . . . . . . . . . . . . . . . . . . . 148

6.11 Scatter plot of B3GNT5 and SOX11 of (Nature 2012) dataset with ER groups, illustrating that the observation 303 from the ER+ group has low expression levels of (B3GNT5, SOX11), which are $(-2.6707, -2.9344)$ in comparison to the observation 378 from the ER- group, which has high expression levels of (B3GNT5, SOX11), which are $(3.9505, 4.103)$. . . . . . . . . . . . . . . . . . . . . . . . 150

6.12 Scatter plot of KRT16 and TMEM40 of (Nature 2012) datasets with ER groups, illustrating that the observation 255 from the ER+ group has low expression levels of (KRT16, TMEM40), which are $(-2.2758, 0.11)$ in comparison to the observation 373 from the ER- group, which has high expression levels of (KRT16, TMEM40), which are $(5.3142, 4.379)$. . . . . . . . . . . . . . . . . . . . . . . . 152

6.13 Scatter plot of BBOX1 and C1orf106 of (Nature 2012) dataset with ER groups, illustrating that the observation 495 from the ER+ group has low expression levels of (BBOX1, C1orf106), which are $(-0.38025, -2.7811)$ in comparison to the observation 243 from the ER- group, which has high expression levels of (BBOX1, C1orf106), which are $(8.0045, 3.2498)$. . . . . . . . . . . . . . . . . . . . . . . 153

6.14 Scatter plot of DKK1 and KRT81 of (Nature 2012) dataset with ER groups, illustrating that the observation 287 from the ER+ group has low expression levels of (DKK1, KRT81), which are $(-7.716, -0.467)$ in comparison to the observation 513 from the ER- group, which has high expression levels of (DKK1, KRT81), which are $(1.7045, 3.7175)$. . . . . . . . . . . . . . . . . . . . . . . 155

6.15 Scatter plot of RARRES1 and S100A8 of (Nature 2012) dataset with ER groups, illustrating that the observation 288 from the ER+ group has low expression levels of (RARRES1, S100A8), which are $(-2.294, -0.23025)$ in comparison to the observation 433 from the ER- group, which has high expression levels of (RARRES1, S100A8), which are $(6.7431, 6.5045)$. . . . . . . . . . . . . 156

6.16 Scatter plot of S100A9 and TRPV6 of (Nature 2012) dataset with ER groups, illustrating that the observation 231 from the ER+ group has low expression levels of (S100A9, TRPV6), which are $(-1.4145, -0.8875)$ in comparison to the observation 412 from the ER- group, which has high expression levels of (S100A9, TRPV6), which are $(6.1717, 5.1405)$. . . . . . . . . . . . . . . . . . . . . . . . 158

6.17 Scatter plot of HRASLS and PPP1R1A of (Nature 2012) dataset with ER groups, illustrating that the observation 159 from the ER+ group has low expression levels of (HRASLS, PPP1R1A), which are $(-4.3462, -2.4207)$ in comparison to the observation 432 from the ER- group, which has high expression levels of (HRASLS, PPP1R1A), which are $(3.4473, 3.6841)$. . . . . . . . . . . . . . . . 159

6.18 Scatter plot of AGR3 and GFRA1 of (Nature 2012) dataset with PR groups, illustrating that the observation 241 from the PR+ group has high expression levels of (AGR3, GFRA1), which are $(5.8145, 3.688)$ in comparison to the observation 463 from the PR- group, which has low expression levels of (AGR3, GFRA1), which are $(-4.4672, -4.1019)$. . . . . . . . . . . . . . . . . . . . . . . . . 160

6.19 Scatter plot of SCUBE2 and SIAH2 of (Nature 2012) dataset with PR groups, illustrating that the observation 171 from the PR+ group has high expression levels of (SCUBE2, SIAH2), which are $(4.3738, 2.3989)$ in comparison to the observation 232 from the PR- group, which has low expression levels of (SCUBE2, SIAH2), which are $(-2.457, -1.3913)$. . . . . . . . . . . . . . . . . . . . . . . . . 162

6.20 Scatter plot of FGD3 and SUSD3 of (Nature 2012) dataset with PR groups, illustrating that the observation 318 from the PR+ group has high expression levels of (FGD3, SUSD3), which are $(4.0443, 5.2162)$ in comparison to the observation 185 from the PR- group, which has low expression levels of (FGD3, SUSD3), which are $(-1.4253, -2.309)$. . . . . . . . . . . . . . . . . . . . . . . . . 163

6.21 Scatter plot of GRPR and PGLYRP2 of (Nature 2012) dataset with PR groups, illustrating that the observation 239 from the PR+ group has high expression levels of (GRPR, PGLYRP2), which are $(6.0467, 4.3773)$ in comparison to the observation 295 from the PR- group, which has low expression levels of (GRPR, PGLYRP2), which are $(-0.72033, -1.16177)$. . . . . . . . . . . . . 165

6.22 Scatter plot of GREB1 and PGR of (Nature 2012) dataset with PR groups, illustrating that the observation 100 from the PR+ group has high expression levels of (GREB1, PGR), which are $(1.262, 5.4402)$ in comparison to the observation 380 from the PR- group, which has low expression levels of (GREB1, PGR), which are $(-4.3128, -4.4838)$. . . . . . . . . . . . . . . . . . . . . . . . . 166

6.23 Scatter plot of ATP6V0A4 and LAD1 of (Nature 2012) dataset with PR groups, illustrating that the observation 169 from the PR+ group has low expression levels of (ATP6V0A4, LAD1), which are $(-4.1855, -1.5495)$ in comparison to the observation 364 from the PR- group, which has high expression levels of (ATP6V0A4, LAD1), which are $(4.459, 3.3618)$. . . . . . . . . . . . . . . . . . . . 168

6.24 Scatter plot of C9orf58 and NXPH1 of (Nature 2012) dataset with PR groups, illustrating that the observation 23 from the PR+ group has low expression levels of (C9orf58, NXPH1), which are $(-3.3423, -1.46)$ in comparison to the observation 126 from the PR- group, which has high expression levels of (C9orf58, NXPH1), which are $(2.4937, 6.3407)$. . . . . . . . . . . . . . . . . . . . . . . . 170

6.25 Scatter plot of CLCA2 and FGFR4 of (Nature 2012) dataset with PR groups, illustrating that the observation 252 from the PR+ group has low expression levels of (CLCA2, FGFR4), which are $(-3.5682, -2.5823)$ in comparison to the observation 32 from the PR- group, which has high expression levels of (CLCA2, FGFR4), which are $(6.3023, 1.8037)$. . . . . . . . . . . . . . . . . . . . . . . . 171

6.26 Scatter plot of PPP1R1A and TRPV6 of (Nature 2012) dataset with PR groups, illustrating that the observation 396 from the PR+ group has low expression levels of (PPP1R1A, TRPV6), which are $(-2.6309, -0.37375)$ in comparison to the observation 215 from the PR- group, which has high expression levels of (PPP1R1A, TRPV6), which are $(4.18067, 5.1658)$. . . . . . . . . . . . . . . . . . 173

6.27 Scatter plot of C1orf115 and TSPAN8 of (Nature 2012) dataset with PR groups, illustrating that the observation 286 from the PR+ group has low expression levels of (C1orf115, TSPAN8), which are $(-2.3804, -5.9947)$ in comparison to the observation 138 from the PR- group, which has high expression levels of (C1orf115, TSPAN8), which are $(2.9592, 3.1603)$. . . . . . . . . . . . . . . . . . 174

1 The performance of the SSCAE at the final iteration of ovarian cancer dataset and METABRIC dataset with ER and PR groups. 189

2 The performance of the SSCAE at the final iteration of the breast invasive carcinoma datasets with ER groups. . . . . . . . . . . . . 190

3 The performance of the SSCAE at the final iteration of the breast invasive carcinoma datasets with PR groups. . . . . . . . . . . . . 191

4 The performance of the SSCAE at the final iteration of the integrated datasets with ER groups. . . . . . . . . . . . . . . . . . . . . 192

5 The performance of the SSCAE at the final iteration of the integrated datasets with PR groups. . . . . . . . . . . . . . . . . . . . . 193

6 The performance of the SSCAE at the final iteration of the integrated datasets with PR groups. . . . . . . . . . . . . . . . . . . . . 194

7 The performance of the SVM and BDT models at the final iteration of ovarian cancer dataset. . . . . . . . . . . . . . . . . . . . . . . . 195

8 The performance of the SVM and BDT models at the final iteration of METABRIC dataset with ER groups. . . . . . . . . . . . . . . . 196

9 The performance of the SVM and BDT models at the final iteration of METABRIC dataset with PR groups. . . . . . . . . . . . . . . . 197

10 The performance of the SVM model at the final iteration of breast invasive carcinoma dataset with ER groups. . . . . . . . . . . . . 198

11 The performance of the BDT model at the final iteration of breast invasive carcinoma dataset with ER groups. . . . . . . . . . . . . 199
12 The performance of the SVM model at the final iteration of breast invasive carcinoma dataset with PR groups. . . . . . . . . . . . . 200
13 The performance of the BDT model at the final iteration of breast invasive carcinoma dataset with PR groups. . . . . . . . . . . . . 201
14 The performance of the SVM model at the final iteration of the integrated dataset with ER groups. . . . . . . . . . . . . . . . . . 202
15 The performance of the BDT model at the final iteration of the integrated dataset with ER groups. . . . . . . . . . . . . . . . . . 203
16 The performance of the SVM model at the final iteration for the integrated dataset with PR groups. . . . . . . . . . . . . . . . . . 204
17 The performance of the BDT model at the final iteration of the integrated dataset with PR groups. . . . . . . . . . . . . . . . . . 205
18 The performance of the SVM model at the final iteration of the integrated dataset with PR groups. . . . . . . . . . . . . . . . . . 206
19 The performance of the BDT model at the final iteration of the integrated dataset with PR groups. . . . . . . . . . . . . . . . . . 207

# List of Tables

4.1 The performance of the SVM and BDT models built on the stable predictors of METABRIC datasets with ER and PR groups. . . . 81

4.2 The performance of the SVM and BDT models built on the generic biomarkers of the breast invasive carcinoma datasets with ER and PR groups. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84

4.3 The performance of the SVM and BDT models built on the generic biomarkers of the integrated datasets with ER groups. . . . . . . 87

4.4 The performance of the SVM and BDT models built on the generic biomarkers of the integrated datasets with PR groups. . . . . . . 90

5.1 The performance of the SVM and BDT models built on the stable predictors of ovarian cancer dataset. . . . . . . . . . . . . . . . . . . 106

5.2 The performance of the SVM and BDT models built on the stable predictors of METABRIC dataset with ER and PR groups. . . . . 110

5.3 The performance of the SSCAE of the breast invasive carcinoma datasets with ER groups. . . . . . . . . . . . . . . . . . . . . . . . 112

5.4 The performance of the SVM and BDT models built on the generic biomarkers of the breast invasive carcinoma datasets with ER groups. 115

5.5 The performance of the SSCAE of the breast invasive carcinoma datasets with PR groups. . . . . . . . . . . . . . . . . . . . . . . . 116

5.6 The performance of the SVM and BDT models built on the generic biomarkers of the breast invasive carcinoma datasets with PR groups. 117

5.7 The performance of the SSCAE of the integrated datasets with ER groups. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 121

5.8 The performance of the SVM and BDT models built on the generic biomarkers of the integrated datasets with ER groups. . . . . . . 122
5.9 The performance of the SSAE of the integrated datasets with PR groups. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125
5.10 The performance of the SVM and BDT models built on the generic biomarkers of the integrated datasets with PR groups. . . . . . . 128

1 The sizes of the training-validation sets of ovarian cancer dataset and METABRIC dataset with ER and PR groups. . . . . . . . . . 186
2 The average MSE of each SCAE of ovarian cancer dataset and METABRIC datasets with ER and PR groups. . . . . . . . . . . . 186
3 The sizes of the training-validation sets of the breast invasive carcinoma datasets with ER groups. . . . . . . . . . . . . . . . . . . 186
4 The sizes of the training-validation sets of the breast invasive carcinoma datasets with PR groups. . . . . . . . . . . . . . . . . . . 187
5 The average MSE of each SCAE of the breast invasive carcinoma datasets with ER groups. . . . . . . . . . . . . . . . . . . . . . . . 187
6 The average MSE of each SCAE of the breast invasive carcinoma datasets with PR groups. . . . . . . . . . . . . . . . . . . . . . . . 187
7 The sizes of the training-validation sets of the integrated datasets with ER groups. . . . . . . . . . . . . . . . . . . . . . . . . . . 187
8 The sizes of the training-validation sets of the integrated datasets with PR groups. . . . . . . . . . . . . . . . . . . . . . . . . . . 188
9 The average MSE of each SCAE of the integrated datasets with ER groups. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 188
10 The average MSE of each SCAE of the integrated datasets with PR groups. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 188

# Nomenclature

**Acronyms**

| | |
|---|---|
| ACO | Ant Colony Optimisation |
| AE | Auto-Encoder |
| AI | Artificial Intelligence |
| AUC | Area Under the Curve |
| BDT | Bagging Decision Tree |
| CV | Cross Validation |
| DL | Deep Learning |
| DM | Deep Mining |
| EA | Evolutionary Algorithm |
| EC | Evolutionary Computation |
| ER | Estrogen Receptor |
| ER+ | ER-positive |
| FCBF | Fast Correlation-Based Filter |
| FN | False Negative |
| FP | False Positive |
| FPR | False Positive Rate |
| GA | Genetic Algorithm |
| HDSSS | High Dimensional Small Sample Size |
| HN | High Negative |
| HP | High Positive |
| IW | Input Weight |

| | |
|---|---|
| LDA | Linear Discriminant Analysis |
| LOOCV | Leave One Out Cross Validation |
| LW | Layer Weight |
| ML | Machine Learning |
| MRI | Magnetic Resonance Imaging |
| mRMR | minimal-Redundancy-Maximal-Relevance |
| mRNA | messenger RNA |
| MSE | Mean Square Error |
| MSKCC | Memorial Sloan Kettering Cancer Center |
| NCI | National Cancer Institute |
| NHGRI | National Human Genome Research Institute |
| NIH | National Institutes of Health |
| NIPS | Neural Information Processing Systems |
| PPV | Positive Predictive Value |
| PR | Progesterone Receptor |
| PR+ | PR-positive |
| PSO | Particle Swarm Optimisation |
| ROC | Receiver Operating Characteristics |
| SCAE | Sparse Compressed Auto-Encoder |
| SNPs | Single Nucleotide Polymorphisms |
| SOMs | Self-Organising Maps |
| SSCAE | Stacked Sparse Compressed Auto-encoder |
| SVM | Support Vector Machine |
| SVM-REF | Support Vector Machine - Recursive Feature Elimination |
| TCGA | The Cancer Genome Atlas |
| TN | True Negative |
| TNR | True Negative Rate |
| TP | True Positive |
| TPR | True Positive Rate |
| WHO | World Health Organisation |

# Chapter 1

# Introduction

The term *feature mining* refers to emerging statistical data analysis and computational intelligence techniques with the goal of knowledge discovery based on a better understanding of the data. Feature mining can also refer to the process of endowing explanatory capability within the statistical and computational paradigms used. Accordingly, feature mining can be described as the discovery of the underlying structure of the data. The knowledge domain addressed by the research discussed in this thesis is that of clinically relevant 'biomarkers' for cancers of interest. A biomarker is formally defined as *"a biological characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention"* [116]. According to World Health Organisation (WHO), International Programme on Chemical Safety[1], a biomarker is defined as *"Any substance, structure or process that can be measured in the body or its products and can influence or predict the incidence of outcome or disease"*. Based on both definitions, a biomarker can be described as a quantifiable biological indicator for detecting diseases, monitoring its progression, and estimating susceptibility to treatment therapy. Clinical tests based on biomarkers have been applied in medical practice for decades for diseases diagnosis and prognosis and drug discovery [14].

Advances in molecular science and the recent availability of microarray data have led to an exponential growth in volume, variety, and complexity of biologi-

[1] http://www.inchem.org/documents/ehc/ehc/ehc222.htm

cal information. The completion of the first sequenced human genome [200, 300] is one of the main triggers of evolution in biology. Microarray technology allows for thousands of genes from a given cell or tissue sample to be examined simultaneously heralding a new era of research in relatively nascent fields such as computational biology and bioinformatics. These changes and others constitute what is called high throughput or high dimensional biology that produces omics data, which is discussed in details in Chapter 2. The availability of omics data repositories such as The Cancer Genome Atlas (TCGA) [314] brings tremendous opportunities for clinicians, bioinformaticians, statisticians and data scientists to benefit from this abundance of cancer data to build a wide range of more accurate models of the mechanisms underlying cancer and related diseases. Furthermore, analysing omics data over several research studies can allow for more innovative discoveries and findings, and this is further illustrated in Chapter 3, along with the impact of the wealth of such data on the research community.

Biomarker identification from omics data has become a key goal to approach precision medicine. Precision medicine aims to exploit this explosion of molecular data together with individual patient characteristics to personalise medical treatment [56]. Therefore, the next frontier in the move towards personalised cancer medicine is to develop sophisticated knowledge discovery models that can detect biomarkers underlying the variations of control (i.e. individuals without disease) and cancer (i.e. individuals with the disease) groups. The extraction of relevant knowledge from omics data can contribute to answering serious etiologic questions about cancer and developing effective procedures to prevent, detect, manage, and treat this heterogeneous complicated disease. Omics data is characterised by high dimensionality, complexity, relatively small sample sizes and the amount of noise. Omics datasets typically contain tens of thousands of molecules (e.g. genes). The problem with high dimensional data was coined firstly by Richard Bellman as 'the curse of dimensionality' [19]. The curse of dimensionality term refers to various phenomena that arise when dealing with data that comprise hundreds or thousands of variables [92].

Having tens of thousands of variables means that the number of possible input configurations is exponential. However, not all of this information is relevant because the feature spaces of such data comprise large amounts of irrelevant and noisy features, including genes with unreliable measurements that can be considered indistinguishable from noise. Moreover, when the dimensionality of a dataset increases, the number of possible variable value combinations exponentially increases, and therefore the available samples become sparse. In addition to such 'curse of dimensionality' issues, the number of genes of omics data vastly exceeds the number of observations, thus, the available biological samples become even more sparse, making the process of discovering a robust subset of relevant molecular markers a very challenging task. As a result, the problem of omics data analysis is more likely to be that the relevant variations underlying the data is not adequately exploited due to an insufficient number of biological samples (i.e. a couple of hundreds), which in turn have a low signal-to-noise ratio as well as their response groups are more likely to have considerable disparate sizes.

High dimensional complex data generated by omics technology has significantly challenged traditional statistical techniques and machine learning methods due to a range of subsequent issues, such as the curse of dimensionality, overfitting, bias-variance trade-off, model robustness, interpretability, and computational cost. Machine learning models applied to this data will have to mitigate against the high risk of becoming too sensitive to the variations in the data used for model fitting and less sensitive to variation in the unseen data during model evaluation, so that the models will have to minimise 'overfitting' the data. Consequently, achieving the trade-off between these bias-variance quantities is becoming more challenging where situations of overfitting (low bias and high variance) or underfitting (high bias and low variance) being easily achieved whilst good generalisation (low bias and low variance) remaining notoriously elusive. Furthermore, a knowledge discovery model that focuses on detecting an informative subset of candidate biomarkers from such small datasets could be very sensitive to which observations are included in the data modelling phase of data mining raising the issue of model robustness, where different outcomes could be obtained due to the little variations in the data. In order to mitigate against these limitations and boost the level of accuracy, the complexity of models has been increased, where increasing the complexity of a model is more likely decreasing its explainability; due to the trade-off between model complexity and interpretability. Herein, it is relevant to emphasise the importance of adding some explanatory capability to the model used by health practitioners and decision-making professionals for prediction relevant to precision medicine. Moreover, handling high-throughput omics data could be computationally intensive, and the potential for the process to become intractable is increased dramatically if the utilised model is slow to fit.

Classical statistical techniques based on univariate and multivariate approaches have been extensively exploited as analytical tools in biology and medicine to detect statistically significant changes in the behaviour of gene/protein expressions among different biological conditions. In other words, biomarker discovery at the molecular level depends on the principle that the discrimination between healthy (control) and diseased patient groups of samples can be determined by the differential expression levels, intensity values, or activity of genes, proteins, and other molecules. For example, intensity values of highly predictive proteins for cancer patients differ significantly from samples in the control group. Therefore, genes or proteins that exhibit significantly the greatest variations across different conditions can be considered as potential biomarkers for a disease or clinically relevant outcome. Accordingly, the comparison between control and cancer groups was the traditional approach to recognise any statistically significant variations, which could lead to discovering any potential biomarkers. However, biological samples of microarray or mass spectrometry data are usually defined with thousands or tens of thousands of variables. From a statistical perspective, inferring useful knowledge from such data using those traditional methods is difficult because they cannot exploit enough of the relevant variations underlying the data. This is particularly true when analysing biological datasets with statistical models that make inherent specific assumptions about the data, such as linearity, normality, and homogeneity of variances that do not necessarily resemble the true function, leading to poor estimation. The detailed evaluation of utilising traditional statistical techniques for knowledge discovery from omics data is critically discussed in Chapter 2.

The research interest has therefore transferred to machine learning algorithms that allow the discovery of interesting complex patterns, which are often missed by the traditional statistical techniques. Since the advent of the big data revolution and the increasingly ubiquitous availability of terabyte data storage and giga- and tera-flop compute power, machine learning methods have become an invaluable tool in computational biology and its cognate disciplines. Machine learning methods have been incorporated in diverse problem domains in healthcare area, leading to many successful applications, ranging from cancer diagnosis and prognosis, medical imaging, to predictive modelling and decision support. Due to the fact that the performance of machine learning methods depends on the data, high throughput complex data generated by omics technology has significantly challenged these learning models. The curse of dimensionality issues combined with the challenge of relatively small sample sizes made it no longer applicable for machine learning algorithms to be employed alone for omics data analysis because the number of representative samples required to exploit enough of the relevant variations underlying the data and achieve an acceptable level of accuracy is growing exponentially. The detailed and critical discussions of employing machine learning methods for knowledge extraction from omics data are presented in Chapter 2.

This has motivated the development of more sophisticated feature mining models to support knowledge discovery for prediction purposes, which has become a core process in the construction of high dimensional classification models. Feature mining aims to detect interesting complexity from the unknown structure of omics data that could not be discovered by traditional statistical techniques or machine learning methods alone. Consequently, a variety of different methodologies and techniques from the fields of statistical data analysis and computational intelligence are integrated in the hope of achieving better performance than using approaches from one field alone. Detailed discussions of various feature mining paradigms are considered in Chapter 2, particularly for high dimensional problems. However, omics data has the additional challenge of small sample sizes such that the number of features is much greater than the number of samples, putting even more pressure on such feature mining models for extracting robust and reliable molecular markers. This is evidenced by the limited success these methods have had in detecting robust and reliable biomarkers for diseases, such as cancers. It can also explain why the discovery of meaningful biomarkers from such datasets remains a major challenge in personalised cancer medicine, and also could illustrate the lack of finding generic biomarkers among the identified published genes for identical clinical conditions.

As a result, the problem of biomarker discovery from High Dimensional Small Sample Size (HDSSS) omics data is complicated and requires more sophisticated approaches that can address these challenges. The significance of choosing the right methodology for each step of an effective feature mining model applied to omics data is emphasised in the research presented in this thesis. The aim of the knowledge discovery models can be achieved by understanding the key research challenges, using the proper techniques, not the available and popular ones, and careful attention to performance estimation in order to report significant and reliable findings.

## 1.1 Aims and Objectives

The overarching aim of this research is to develop effective feature mining models that robustly aid the extraction of knowledge from HDSSS omics data in a way that is transparent and supports the endeavour of precision medicine.

In order to accomplish this aim, the following objectives will be met:

- Identify and characterise suitable and reliable high quality HDSSS omics datasets for cancers of interest (e.g. TCGA datasets).
- Empirically establish effective data pre-processing methods that maximise the ability of the feature mining models to identify salient biomarkers.
- Critically evaluate state-of-the-art evolutionary computation and deep neural network methods for biomarker discovery.
- Develop novel feature mining models and related innovations which mitigate against the limitations reported in the research literature, whilst maximising their strengths for biomarker discovery.
- Determine and mitigate against the sensitivity of the feature mining models to imbalanced group datasets.

- Investigate and establish most appropriate model selection methods (including objective functions) to effect the simplest model with the highest level of generalisation for each model class.

- Explore and implement a technique for interpreting salient features identified within the deep feature learning model.

- Investigate and implement appropriate validation and evaluation metrics for estimating the generalisation and robustness performance of the feature mining models.

- Identify appropriate validation criteria to verify the validity of the biomarkers discovered by the feature mining models with the specific criteria of predictivity, stability, and generalisability.

## 1.2 Contribution of the Thesis

The first fundamental issue addressed by this research was the reliable extraction of important biomarker information from HDSSS omics data. The research explored a number of disparate paths within the cognate disciplines of computational intelligence and computational biology. The first path was conducted by investigating the direction of solving the biomarker discovery as an optimisation problem. Biological data generated by omics technology has thousands of variables and to identify relevant genes to the response groups or conditions, an extremely large number of evaluations is required. Therefore, feature mining approaches that can guarantee to find the optimal subset of features are computationally expensive and infeasible in most practical cases. Optimisation methods attempt to identify the best possible subset of features from the exponential search space of omics data with the least amount of effort. Therefore, *the primary contribution of this thesis* is to develop an ensemble evolutionary mining model based on a hybrid selection approach to navigate through large genomic and proteomic data and detect an ensemble subset of stable predictors. Different paradigms of feature selection based on the optimisation method have been investigated to find the most appropriate measurement for a HDSSS problem. Consequently, the ensemble hybrid selection approach is integrated with the parallel adaptive search of the evolutionary method so that the curse of dimensionality issues can be handled and the robustness of the selected subsets of candidate predictors can be enhanced.

What has driven us to the second direction of this research is that a feature learning model that can discover relevant knowledge automatically from large-scale data, without the need for hand designed features that require domain expertise or ad-hoc specific methodologies or techniques is highly desirable. Therefore, the second path of our research is investigating the usefulness of state-of-the-art Deep Learning to mitigate against the mentioned limitations on the basis of automatically capturing enough of the meaningful abstractions latent with the available biological samples. Deep learning methods provide superior performance over traditional learning approaches by handling the curse of dimensionality, improving the generalisability, and making meaningful use of the data in a wide range of problem domains such as computer vision, natural language processing, and speech recognition. Recently, in the healthcare area, deep learning methods have brought about breakthroughs in medical imaging such as CheXpert [146], a large dataset that contains $224,316$ chest radiographs of $65,240$ patients for chest radiograph interpretation.

In many of these problem domains, a large number of samples are typically available to train a deep learning model where the signal-to-noise ratio is quite high. The key challenge is to capture the generic factors of variations that underlie the unknown structure of the data in a way that can significantly enhance the generalisation to unseen observations. This is, however, not the case in bioinformatics research where high throughput omics datasets are characterised by a small number of biological samples (i.e. hundreds of patient samples), which in turn have a low signal-to-noise ratio. Therefore, for omics data analysis, the problem is more likely to be that the relevant variations underlying the data can not be adequately captured due to an insufficient number of biological samples. As a result, it may seem somewhat counterintuitive to use deep learning methods for HDSSS datasets due to the fact that these learning models typically require substantial data to constrain their parameters and learn a useful hypothesis. Applications of deep neural network methods for knowledge discovery from HDSSS omics data remain scarce. This necessities introducing new deep learning-inspired paradigms that can approximate enough of the relevant variations represented by those biological samples. Therefore, *the second contribution of this thesis* is introducing a new deep feature learning model that can capture enough of the complexity of interest represented by the available biological samples. More specifically, the proposed deep learning model is introduced based on a set of non-linear sparse Auto-Encoders that are deliberately constructed in an under-complete manner to force the network to discover enough of the interesting complexity underlying the biological samples. The ability of using a stacked set of neural auto-encoders alleviates the issue of vanishing gradients and therefore provides a robust deep learning model to automatically identify the complex featural representations necessary to capture the important variations within the original dataset. The proposed deep feature learning model is utilised to discover and interpret important signals from omics data that aid prediction relevant to precision medicine.

The proposed deep feature learning model applies multiple levels of projections to the input features to abstract the problem and capture high-level dependencies for achieving high-level of generalisability. This would be a powerful learning model for high dimensional classification problems. However, for the problem of biomarker identification, it is hard to interpret which subsets of genes were dominant within the internal representations and responsible for deriving such predictions. Therefore, a fundamental issue with the deep learning paradigm is the lack of explanatory power, and their inability to unambiguously state which input features are responsible for its behaviour. To overcome the inherent issue of poor explanatory power associated with the deep learning paradigm, we endeavour in a new direction of research that focuses on deconstructing the internal mechanism of such deep learning models based on a new weight interpretation method. The learning process of the deep learning relies mainly on sensibly fitting the weight configurations to define the model's input-output function. This reflects the fact that the weight is the main indicator of variable's importance, in which the weight of each variable reflects its contribution through the network, so that the signal with a larger positive or negative weight has a greater impact. *Therefore,the third contribution of this thesis* is proposing a new technique called deep mining to sculpt inside the deep feature learning model and open the so-called black box of the network for biomarker identification. A model that is able to state which phenotypes are key determinants is a crucial element of prediction systems used by health practitioners and decision-making professionals. It is therefore important we are able to provide some explanatory capability to our deep learning model. Our novel deep mining model provides yet another arrow within the quiver of bioinformaticians for discovering and evaluating new biomarkers that may help further the endeavour of producing more effective and personalised medicine.

The application of the proposed feature mining models to the utilised omics datasets has led to *the fourth contribution of this research*, which is discovering relevant, robust and reproducible biomarkers for breast cancer and the positivity of Estrogen and Progesterone receptors. The detected biomarkers are validated reliably across a wide range of independently generated breast cancer data samples that are collected from completely different studies. The fundamental concepts of omics data, breast cancer, and understanding the role Estrogen Receptor and Progesterone Receptor play in this heterogeneous complex disease are detailing covered and discussed in Chapter 2.

In this thesis, the principle has been emphasised that the discovered molecular markers should meet the following criteria to act as true biomarkers, which are *Predictivity*, *Stability*, and *Generalisability*. Predictivity is introduced to examine the capability of the discovered biomarkers to separate patients in the cancer group from those in the control group with a good level of certainty. The lack of overlap among the published genes or proteins for identical diseases or clinical outcomes is essentially caused by the lack of robustness or stability of the selected genes across samples. Therefore, stability is utilised to investigate how the variations in the training data can affect the feature preferences of the proposed feature mining models, and to fight the sparsity of data points in a high-dimensional space. *"If the same features are selected in multiple independent iterations, they more likely are reliable biomarkers"* [100].

Generalisability is employed to test the potential of the proposed feature mining models to detect generic biomarkers from multiple independent datasets that are collected from completely different studies so that the highest evidence can be provided. The research study [287] has hypothesised that *"External validation using data from a completely different study provides the highest irrefutable evidence that a tool validates"*. From a large body of research that focuses on biomarkers discovery, few studies have adopted another independent dataset for validation purposes despite the availability of the data generated by TCGA program with high standard samples. That could explain why the number of clinically validated biomarkers is very few, despite the numerous proposals in the literature.

## 1.3 Thesis Outline

This thesis is structured as follows:

**Chapter** 2: Literature Review
This chapter provides the groundwork for informing how best to achieve the stated objectives by providing a summary of the key concepts and research directions in the area of cancer biomarker discovery from omics data. It starts with an introduction to breast cancer and the fundamental types of data generated from omics technologies. Subsequently, this chapter provides a critical discussion of the current approaches to biomarker discovery found in the literature. Chapter 2 investigates the appropriateness of different experimental methodologies, validation and evaluation metrics for verifying the outcomes of the feature mining models constructed using HDSSS data.

**Chapter** 3: Datasets and Experimental Methodology
This chapter explains the datasets used to perform omics data modelling and analysis and the experimental methodologies and evaluation metrics applied to estimate the robustness of the discovered biomarkers. The chapter starts by explaining the data pre-processing methods that are utilised for filtering out genomic datasets from genes with unreliable measurements. Then, the sources of high quality HDSSS omics datasets for cancers of interest are illustrated with an emphasis on gaining the maximum benefit from these publically available datasets. In this project, 18 datasets have been utilised to examine the potential of the presented feature mining models to discover robust and generic knowledge.

The cancer datasets, evaluation metrics, and validation approaches used to analyse them are discussed in detail in this chapter.

**Chapter** 4: Evolutionary Mining Model
This chapter covers the design, implementation, and application of the ensemble evolutionary mining model proposed for biomarker identification from omics data. It starts with an introduction to why and how to solve the problem of biomarker discovery using optimisation methods. Subsequently, an overview to one of the most powerful optimisation methods, the Genetic Algorithm, is introduced. Thereafter, Chapter 4 discusses the experimental design of the evolutionary mining model, which integrates the Genetic Algorithm and the ensemble hybrid selection approach. Feasible choices for each step of the experimental design are investigated and justified. The performance of the proposed evolutionary mining model is evaluated using the datasets and the experimental methodology mentioned in Chapter 3.

**Chapter** 5: Deep Mining Model
This chapter covers the design, implementation, and application of the deep mining model proposed for biomarker discovery from omics data. It provides an introduction to the fundamental components necessitated to develop an effective deep feature learning model that can exploit the unknown structure of omics data effectively. Consequently, the design steps of the deep mining model based on an unsupervised data-orientated approach is introduced to discover and interpret important signals from proteomic and genomic data. Furthermore, Chapter 5 discusses a new weight interpretation technique that is proposed to add explanatory power to our deep learning model, helping to alleviate one of the most challenging problems associated with the deep learning paradigm. The proposed deep mining model is evaluated using the datasets and experimental methodology introduced in Chapter 3.

**Chapter** 6: Biomarkers and Bioinformatics
The generic biomarkers for breast cancer discovered by our feature mining models have been validated in terms of predictivity, stability, and generalisability

in Chapters 4 and 5. In this chapter, the clinical relevance of the discovered biomarkers will be evaluated with respect to current bioinformatics research into breast cancer. It is important to emphasise that, at the time of writing, there is no research that has found or examined the combination of these biomarkers or some of them simultaneously. Furthermore, the association between each biomarker and the hormone receptors recognised in this PhD work is discussed to identify the type of existent relationship.

**Chapter** 7: Conclusion and Future Work
This thesis concludes with a discussion of the crucial challenges underlying the problem of inferring knowledge from HDSSS omics data, a summary of the contributions made to help alleviate these challenges and finally, potential future directions for this research and cancer biomarker discovery.

# Chapter 2

# Literature Review

## 2.1 Introduction

Current computational models and tools for detecting breast cancer and understanding the role Estrogen Receptor and Progesterone Receptor play in this heterogeneous disease are detailing reviewed in this chapter. A short overview of breast cancer, available omics data and the central dogma of molecular biology are first considered together with the challenges these pose for any data mining or computational model that may be used for biomarker discovery. Current state-of-the-art approaches for knowledge extraction are then subsequently reviewed along with strengths, limitations and challenges. An emphasis is made in this chapter on critical underlying issues of validating and evaluating the empirical results of biomarker discovery models proposed for HDSSS omics data. Increasing the awareness of the key research challenges allows for more efficacious solutions by understanding the required computational and statistical resources.

## 2.2 Breast Cancer

Breast cancer is the most common neoplasm in women and the second leading cause of cancer-related mortality in females worldwide [18]. Mammography is the standard tool that has been used for detecting breast cancer [114]. However, several issues have been raised about this procedure including the risk of false positives, over diagnosis of indolent disease, and lowering the sensitivity of recognising tumours in women with dense breast tissue [40, 224, 316]. Magnetic Resonance Imaging (MRI) offers a powerful alternative and provides excellent imaging even around dense breast tissue [26]. However, a high risk of obtaining false positives could lead to needless, stressful and expensive procedures [137]. Therefore, there is a critical necessity for measurement of molecular markers that could estimate the potential occurrence of a disease, and providing the probability of specific outcomes to the clinician for treatment stratification. Recognition of breast cancer at early stages can bring better prognosis with a 5-year survival rate of up to (90%), however, when breast cancer spreads to distant organs, this survival rate declines drastically to (20%) [90]. Detection at the early stages and monitoring breast cancer remain major challenges for healthcare professionals. Moreover, the aetiology of breast cancer is still ambiguous, where breast cancer can differ significantly in regards to clinical, pathological, and biological properties.

Breast cancer begins when healthy cells change and grow out of control, forming a mass called a tumour. A tumour can be *malignant* or *benign*. A benign tumour means a tumour can grow but will not spread. A malignant tumour can grow and spread to other parts of the body. A malignant tumour has an abnormally high level of *Estrogen Receptor* and *Progesterone Receptor* in the nucleus. According to the website of National Cancer Institute (NCI)[1], Estrogen Receptor is *"a protein found inside the cells of the female reproductive tissue, some other types of tissue, and some cancer cells. The hormone estrogen will bind to the receptors inside the cells and may cause the cells to grow. Also called ER"*. The NCI's website[2] defines Progesterone Receptor as follows: *"A protein found inside the cells of the female reproductive tissue, some other types of tissue, and some cancer cells. The hormone progesterone will bind to the receptors inside the cells and may cause the cells to grow. Also called PR"*. Testing the tumour for Estrogen Receptor and Progesterone Receptor is a standard part of the initial evaluation of breast cancer diagnosis and treatment planning. The analysis of Estrogen and

[1] https://www.cancer.gov/publications/dictionaries/cancer-terms/def/estrogen-receptor

[2] https://www.cancer.gov/publications/dictionaries/cancer-terms/def/progesterone-receptor

Progesterone Receptors by Immunohistochemistry (IHC) is considered currently the most commonly used method to test the tumour for both hormone receptors in cancer cells from a sample of tissue, which may come from a biopsy [97].

If breast cancer cells have high ER, the cancer is described as ER-positive (ER+), and if breast cancer cells have high PR, the disease is specified as PR-positive (PR+) cancer. ER and PR expressions have been utilised as robust indicators for the evaluation of breast cancer. All newly diagnosed invasive breast cancer patients and breast cancer recurrences should be examined for both ER and PR according to the recommendations of the American Society of Clinical Oncology and the College of American Pathologists [123]. According to cancer research UK, $\sim$ 37000 out of 50000 new cases are distinguished by the presence of ER. However, it has been shown that the expression of ER and PR receptors changes during the development of breast cancer and in response to systemic therapies [199].

For patients with ER+, particular treatments that block the activity of ER are recommended. ER activation plays a significant role in different biological processes like cell development and cell death [160]. The mechanism of blocking ER activity relies essentially on changing ER function in such a way that ER is becoming unable to regulate gene expression [259]. According to Carroll [46] *"Oestrogen Receptor (ER) is a transcription factor that regulates gene expression events that culminate in cell division"*. Several expression profiling studies have illustrated that the expression of hormone receptors is linked with diverse genetic variations [231, 267, 268]. That means several mutated genes can affect the development and progression of breast cancer and contribute to its heterogeneity [29]. As a result, investigating molecular characteristics of the tumours that could act as risk factors of breast cancer is considered a serious aetiologic question [104]. This research project aims to identify mRNA markers from gene expression data that underlie the biological processes of ER and PR receptors.

With the advent of omics technologies, various biological molecules like genes, transcripts, proteins, metabolites, and other species have been provided. The next section provides an introduction to the central dogma of molecular biology and the fundamental types of data generated from omics technologies.

## 2.3 Omics Data

In 1957, a symposium of the Society for Experimental Biology in London presented one of the fundamental ideas of molecular biology, which is called the **central dogma**, and then it was published by Francis Crick in 1958 [61]. The concept of the central dogma of molecular biology specifies the transfer of genetic information within the biological system. This sequential process is shown in Figure 2.1 and involves the following processes: Replication, Transcription, Reverse Transcription, and Translation. Replication (DNA to DNA): is the process of copying all of a cell's DNA. Transcription (DNA to RNA) is the process, in which the DNA is transcribed to RNA, which carried the needed information to protein. Reverse Transcription (RNA to DNA): in this process, the RNA is reserved transcribed to DNA. Translation (RNA to protein): is the process in which the RNA is decoded to make a protein. Crick states that *"once (sequential) information has passed into protein it cannot get out again*" [121].

Different measurements provided by current technologies can be performed on and beyond distinct layers of the dogma to produce the so-called omics data, as shown in Figure 2.1. The fundamental aim of omics technologies is detecting genomics, transcriptomics, proteomics and metabolomics in a specific biological sample. Further to the role of omics technology in providing a great insight to the physiological system, they play a significant role in developing diagnosis and prognosis systems, investigating biomarkers at the molecular level, advance pharmacogenomics studies and expand our knowledge about the aetiology of complex diseases.

Omics fields can be grouped as follows [152]:

- Genomics is the systematic study of an organism's genome. The genome can be defined as the complete set of genetic information (DNA sequence) of a cell or organism. Conventional methods have analysed genes independently, whilst recent microarray technology measures genetic variants between individuals and the expression of thousands of genes simultaneously in order to reveal if any abnormality is associated with a trait [138]. The most popular differences in genetic information between humans are Single Nucleotide Polymorphisms (SNPs), where a SNP is a variation at a

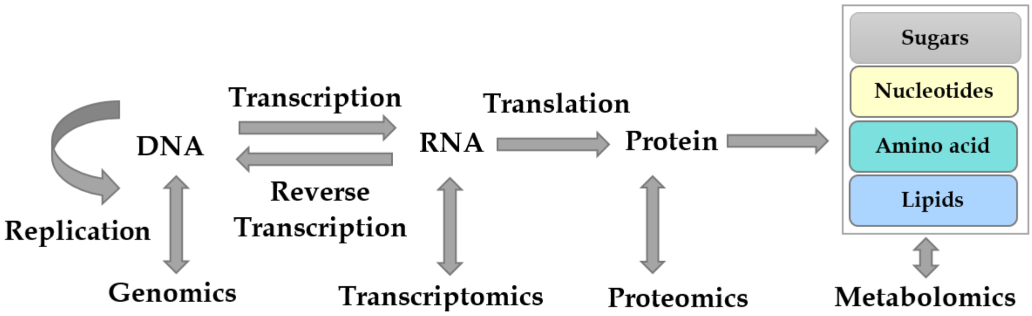


Figure 2.1: The central dogma of molecular biology [121] and the types of omics data generated from each layer of dogma.

single DNA site [83]. Therefore, SNPs have been explored for detecting diseases with a genetic determination and in pharmacogenomics for assessing the efficacy of drug therapies.

- Transcriptomics is the study of the mRNA within a cell or organism. The transcriptome is the total mRNA transcripts that reflect the gene activity within the cell. Microarrays have been utilised in several areas of bioinformatics, and it is used in transcriptome to measure mRNA and summarises the actively expressed genes.

- Proteomics is the large-scale study of proteins, including their structure and function, within a cell or organism [289]. The proteome is the set of all expressed proteins in a cell or organism. Proteomics is another interesting area of research after genomics because it can provide more comprehension to the complex biological procedures due to its direct role in cell physiology. The proteome is considered a reflection of genomic and environmental factors. Therefore, it may hold a promising piece of knowledge, which can address different biological questions of interest [276]. However, a large number of proteins is produced.

- Metabolomics is the study of global metabolite profiles in a cell or organism [113]. The metabolome is the outcome of integrating the transcriptome and the proteome [296]. Thus, changes in the metabolome are related to changes in this product. The metabolome involves the smallest domain size comparing with other omics data. Among different metabolite molecules, which are illustrated in Figure 2.1, the research interest has been focused recently on lipidomics due to their significant role in several diseases such as obesity, atherosclerosis, stroke, hypertension and diabetes [124].

High-throughput technologies allow thousands of variables to be examined simultaneously in a biological sample within a single experiment. Thus, it has the potential to detect key molecules that can answer the biological questions of interest so that new treatment strategies and drugs can be provided. The potential research directions for biomarker discovery using the state-of-the-art approaches proposed in the literature will be discussed in the next sections.

## 2.4 Statistical Techniques for Biomarker Discovery

This section offers a brief introduction to traditional statistical methods applied in disease biomarker discovery studies. Conventional statistical techniques used to be the standard methods for the analysis of biomedical data such as hypothesis testing, correlation, regression, and clustering analysis. Statistical methods based on the univariate approach (e.g. [193]) assess the optimality of each variable independently from the others assuming there is no interaction between them. The univariate analysis produces a list of features, sorted according to their discriminative power in separating the samples of different response groups. However, omics data analysis based on univariate tests can increase the risk of obtaining 'spurious' markers by misclassifying genes as differentially expressed when they are not. When a large number of genes is available, the risk of obtaining false positives is increased due to the challenge of multiple comparisons [228, 270]. Although several procedures have been introduced in the literature to tackle the multiple comparisons problem such as the Benjamini-Hochberg [25, 242], and the Bonferroni correction [263] as well as procedures in pattern mining established by [312, 313], some issues have been raised about them [14]. In this research, finding robust biomarkers for cancers of interest is a discovery-based approach, and more information about hypothesis-based style can be found in [197].

On the other hand, the conceptual simplicity and the lower demands of univariate statistical techniques still attract researchers to utilise them as a pre-processing step to reduce the dimensionality of the data in preparation for more complex multivariate modelling or learning (e.g. [310]). Among a wide range of univariate statistical methods, $t-$test is widely utilised as a pre-processing step [183, 262]. In genomic data, it has been shown that there is a non-trivial proportion of genes that have unequal group variances [72]. Thus, it is important to consider that the unequal variance $t-$test is more appropriate to find discriminative features than other hypothesis testing methods.

High-throughput omics data are multivariate, where the biological outcome is distributed in several biomarkers that need to be assessed simultaneously rather than independently. Statistical techniques based on multivariate approach (e.g. [181]) consider the effect of variables jointly rather than individually. Many successes in biology and medicine have been achieved using these conventional statistical methods (e.g. [202,304]). However, extracting and interpreting relevant knowledge from high dimensional and complex omics data remain significant challenges for these classical models. The main drawback of these models is that they make specific assumptions about the data such as linearity, normality and homogeneity of variances that do not necessarily resemble the true function, leading to poor estimation [204]. This is evidenced by the limited success these methods have had in discovering robust and reliable molecular markers for diseases such as cancers. Therefore, high dimensional data analysis has become an active area of statistical research [75].

The trade-off between model complexity and the possibility of overfitting has re-acknowledged conventional multivariate linear methods. Moreover, the simplicity of the theoretical concepts of linear models like these suggested by Hotelling [139] and Fisher [96] still attracts us nowadays, to be employed as a powerful methodology that can understand the underlying structure of the data and summarise it in simpler ways. *"simple methods typically yield performances almost as good as more sophisticated methods"* [125]. However, it still seems hopeless to adopt these simple models alone for handling high dimensional problems rather they can be useful on top of other sophisticated methods.

Many of the early microarray analysis studies have utilised clustering techniques for the aim of biomarker identification. In the machine learning domain, a clustering approach is referred to as Unsupervised Learning and with other techniques like Discriminant Analysis, Statistical Learning is formed. It is significant here to differentiate the mechanism of Principle Component Analysis (PCA) [246] that looks to find a low-dimensional representation of the observations that explains a good fraction of the variance from the clustering analysis that looks to find homogeneous subgroups among the observations. For PCA, in addition to its linearity, it is impossible to estimate the amount of information that is preserved in a space defined by the first few principle components as well as there is no universally agreed method for reliably recovering key determinants (i.e. genes) from the principal components. For clustering analysis, diverse techniques have been introduced in the literature, such as hierarchical clustering (e.g [299]), k-means clustering [203], and Self-Organising Maps (SOMs) [168]. Research has shown that none of these clustering analysis methods have consistently outperformed the others, therefore, diverse clustering methods are typically applied to complex molecular data for producing ensemble outcomes.

Clustering analysis partitions the data space into smaller distinct clusters or regions so that the observations within each cluster are quite similar to each other and dissimilar to observations in other clusters. For example, assigning samples to similar cancer subtypes. As a result, the hypothesis of cluster analysis depends on the similarity notion that measures a distance between patterns, which has diverse forms (e.g. Euclidean, Manhattan). The new groups may not be related to the status of these samples so that interactions can be uncovered in the data. In the biomarker identification context from genomic or proteomics data, clustering methods attempt to find genes or proteins that exhibit similar expression patterns (e.g. [107]). However, the major drawback of using clustering analysis for high dimensional data is that the number of distinct regions grows linearly with the number of parameters. While with a deep neural network, it has been shown that the number of distinct regions can grow exponentially with the number of parameters using sparse representations. More information about the difference between clustering and *multiclustering* can be found in the research paper [23].

## 2.5 Machine Learning Methods for Biomarker Discovery

The term of Machine Learning (ML) refers to the capability of an algorithm to learn from data. Many different definitions have been introduced to specify a machine learning. A brief definition has been provided by Mitchell in 1997 [211]: *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"*. According to this formal definition, learning refers to the ability to execute the task, while the task can be defined as processing a set of examples, and an example is a collection of features and desired responses or groupings. The training process relies on learning an underlying and previously unknown structure of the data from the training examples so that the learned model can assign an unobserved example to its target. Where learning typically involves a search procedure over the parameter or rule space to identify a range of values or settings that minimises the cost function. Let's assume $X$ is a $d$-dimensional vector, where $X \in \mathrm{R}^d$, and $Y$ is the response group, where $Y \in \mathbb{R}$, which takes two numerical values {0,1} or {-1,1}. The classification function $f$ represents the systematic information that $X$ provides about $Y$, where $f : \mathrm{R}^d \rightarrow \{0, 1\}$.

Since the advent of high-throughput biomedical data, ML methods have become an invaluable tool in computational biology and its cognate disciplines. Although the characteristics of most ML methods are well understood, the performance of these learning models depends on the data. High dimensional complex data generated by omics technology has significantly challenged ML models due to various phenomena ranging from the curse of dimensionality, overfitting, bias-variance trade-off, model robustness, interpretability, and computational cost. Simple classification models cannot be developed using all the available features. Even if other learning models can be constructed, the large dimension spaces of omics data contain many irrelevant and noisy variables that do not contribute to reduce the misclassification rate, rather degrade the prediction performance to the level of random guessing. As a result, the curse of dimensionality issues of omics data made it no longer applicable to use ML methods alone because the number of examples needed to represent the number of the variations in the data and achieve an acceptable level of classification accuracy is growing exponentially. Therefore, feature mining has become a critical pre-processing step before the data is applied to the ML model. Reducing the number of features contribute to reducing the number of samples required to achieve a good-level of generalisation. Therefore, for high dimensional problems, choosing the appropriate methodology for the feature mining stage seems to be an essential precursor to the machine learning stage, since ML models will be trained using reduced feature spaces. However, dealing with small sample sizes datasets is another critical issue that needs to be properly handled for true estimation.

Complex classification models with highly fitted decision boundaries discriminate the training observations optimally. That means, training error rate (the percentage of training samples misclassified by the learned model) consistently decreases with the increase of model complexity. However, they might not be able to assign the testing samples to their response groups correctly due to the trade-off between model complexity and the possibility of overfitting. That means, the learner becomes infeasibly flexible such that it is becoming too sensitive to the variance found in the training set and as a consequence becoming less sensitive to any additional variation found in testing data. Therefore, classification models with complex boundaries are likely to overfit to training data causing poor generalisation ability on testing data. For example, a quadratic curve might fit the data points perfectly, however it might not generalise well. While linear classifiers with simple hyperplane decision boundaries tend to suffer less from overfitting and generalise well. That means, the underlying variations in the data can be better allowed with a simple straight line contributing to reducing the risk of overfitting.

Therefore, in this thesis, powerful but not too adaptable classification models are utilised in this research to assess the predictive power of the selected biomarkers, which are Support Vector Machine and Bagging Decision Tree. These learning techniques are selected due to their empirical power and success in the same or similar domains. The aim of proposing the feature mining models is to derive cancer markers whose behaviour differs significantly across the biological conditions, thus the utilised learning models can be employed to develop reliable prediction systems. Consequently, the accuracy of these classification model built on the dataset containing only the informative genes is listed as one of the main criteria to assess the quality of the discovered biomarkers.

### 2.5.1 Support Vector Machine

Support Vector Machine (SVM) [35] is one of the most robust classification models that is well-proven across a wide range of settings, especially for high dimensional biomedical data [20,217]. In a classification problem, SVM finds its boundaries in the $d-$dimension space that can distinguish observations of differentiated groups. In a $d-$dimension space, the boundary is called a hyperplane, where the hyperplane can be defined as a flat affine subspace of dimension $d-1$. Having a set of $n$ examples $(x_i, y_i), i = 1, 2, ..., n$, where $\mathbf{x}_i \in \mathrm{R}^d$, and $y_i \in \mathbb{R}$, which takes two numerical values {1,-1}. A hyperplane can be defined by the equation:

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{2.1}$$

where $\mathbf{w}$ is a $d-$dimensional coefficient vector and the bias term $b$ is the offset of the hyperplane from the origin. The decision rule to classify new observations is:

$$\begin{cases} y_i = -1 & \text{if } \mathbf{w}^T\mathbf{x}_i + b < 0 \\ y_i = +1 & \text{if } \mathbf{w}^T\mathbf{x}_i + b > 0. \end{cases} \tag{2.2}$$

This could be valid for multiple hyperplanes, thus which of the possible separating hyperplanes should be chosen. SVM chooses the best separating hyperplane (i.e. the decision boundary), which maximises the margin of separation (i.e. the maximum safety distance between the boundary and the training points that are closest to the boundary) by solving the following optimisation task:

$$\min_{w} = \frac{\|\mathbf{w}\|^2}{2} \tag{2.3}$$

$$Subject\,to:\; y_i(\mathbf{w}^T.\mathbf{x}_i) + b \geq 1, i = 1, 2, ..., n \tag{2.4}$$

the size of the margin is $\frac{1}{\|\mathbf{w}\|}$, thus minimising $\|\mathbf{w}\|$ leads to maximise the margin,

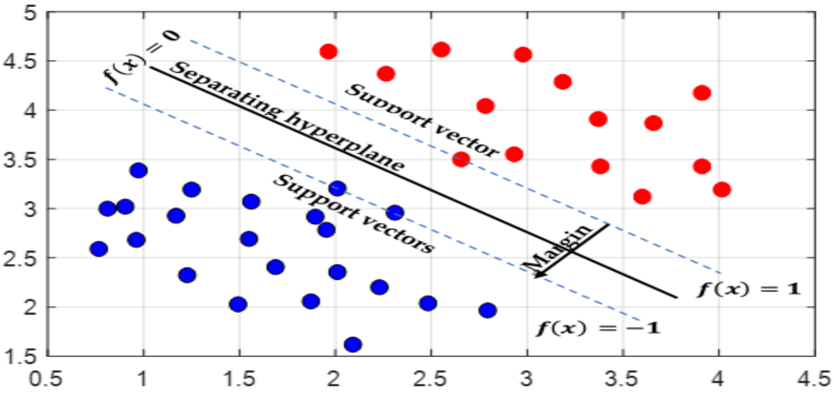


Figure 2.2: An example dataset described by two genes of a linear separator that maximises the margin between positive samples and negative samples, the red dots on the right side represent the positive group; and the blue dots on the left side represent the negative group. There are three support vectors. One point of the positive group on the right dashed line, and two negative samples on the left dashed line.

where each data point lies in the correct side of the separating hyper-plane as shown in Figure 2.2. The training data points that are closest to the decision boundary are called the support vectors.

For a learning model with best separating hyperplane, the risk of overfitting could be increased due to its high sensitivity to the change in training points. Therefore, a hyperplane that allows some training observations to be misclassified could achieve a better job in classifying testing observations. In the maximal margin classifier, every point must be on the correct side of the hyperplane and the margin. A compromise between maximising the margin and minimising the cost of misclassification is required. Therefore, the support vector machine or soft margin classifier allows some points to be on the incorrect side of the margin or even the hyperplane. The soft-margin relaxes the constraints of Equation 2.3 by imposing a penalty on the length of the margin for every point that is on the wrong side of the decision boundary and as follows:

$$\min_{w,\xi} = \frac{\|\mathbf{w}\|^2}{2} + C\sum_{i=1}^{n}\xi_i \tag{2.5}$$

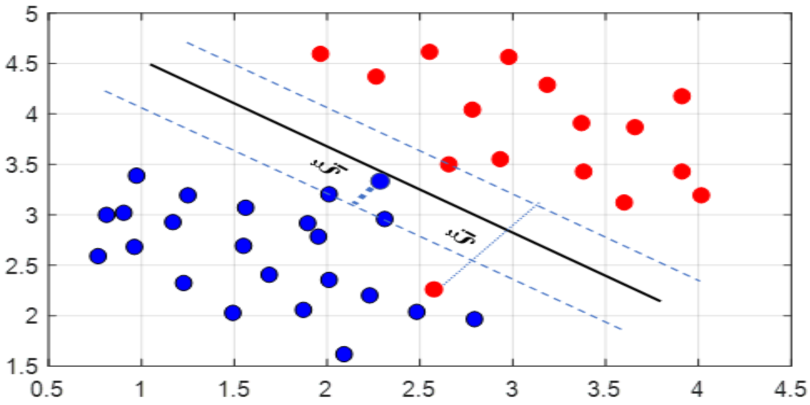


Figure 2.3: SVM soft-margin allows some data points to be misclassified or within the margin through slack variables $\xi_i$.

$$Subject\,to:\ y_i(\mathbf{w}^T.\mathbf{x}_i) + b \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{2.6}$$

The slack variables $\xi_i$ are measures of the margin violations for the training data points, such as $\xi_i < 1$ for the training points that are correctly classified, $\xi_i = 1$ for the training points on the separating hyperplane, and $\xi_i > 1$ for the training points that are misclassified (Figure 2.3). That means that the data point that locates strictly on the correct side of the margin does not impact the model, and only the support vectors can affect SVM classifier. The fact that only the support vector points can affect the decision rule of SVM classifier makes it more robust classifier, due to its low sensitivity to the behaviour of training points that locate far from the hyperplane. Parameter $C$ determines a penalty for misclassification - (the trade-off between maximising the margin and minimising the number of misclassified training points). Therefore, it is important to reduce the risk of overfitting and enhance generalisation performance.

### 2.5.2 Bagging Decision Tree

Decision trees are widely utilised classification techniques due to their simplicity and comprehensibility. Decision trees represent the relationships between features hierarchically such that the relationships and the values of each feature contribute

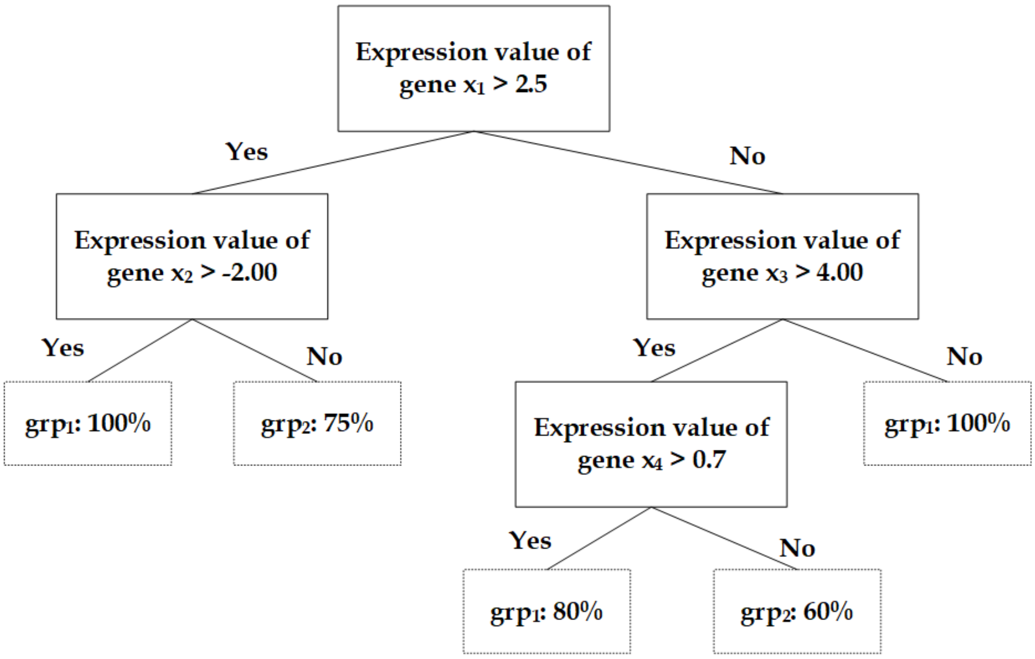

Expression value of gene x1 > 2.5
Yes
No
Expression value of gene x2 > -2.00
Expression value of gene x3 > 4.00
Yes
No
Yes
No
grp1: 100%
grp2: 75%
Expression value of gene x4 > 0.7
grp1: 100%
Yes
No
grp1: 80%
grp2: 60%

Figure 2.4: An illustration of a decision tree constructed using a gene expression dataset of four genes.

to constructing a classification model that can be used to assign new observations to their response groups correctly. To illustrate the hierarchical structure of decision trees, let's discuss how to construct decision trees and how to use that construction to estimate the response groups of a new observation. Figure 2.4 presents a decision tree model constructed from a gene expression dataset that consists of four genes: $x_1, x_2, x_3, x_4$, and two response groups: $grp_1$ and $grp_2$. The nodes that are represented by solid-line rectangles in Figure 2.4 represent the selected genes and their expression values, which should be good cut-points to best assign the samples into their response groups. The leaves that are represented by dashed-line rectangles in Figure 2.4 involve the percentage of observations that are classified to their response groups based on different gene expression values that are obtained by navigating the tree from the top (i.e. the root node) to the bottom (i.e. the leave).

The tree starts where all the training examples belong to a single region and a series of splitting rules are developed further down of the model. Therefore, a new observation will eventually be dropped down the tree into a decision node (i.e leave) where no further evaluations are required. The splitting criterion relies on maximising the information about the response groups, so that how well the new split can discriminate observations of $grp_1$ from $grp_2$. Diverse tree induction techniques have been developed for performing the splitting rule and selecting candidate predictors such as homogeneity or purity of the group distribution associated with each node. The entropy [251] and the Gini impurity index [42] are well-known measurements of node's purity. Therefore, the Gini impurity index is utilised in this research to measure the total variance over the $K$ response groups, and it can be defined by the equation:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{2.7}$$

where $\hat{p}_{mk}$ is the proportion of training examples in the $m$th region that belong to $k$th response group. If all $\hat{p}_{mk}$ are close to 0 or 1, the value of the Gini impurity index will be small. Small Gini index means that a node predominantly has training examples that belong to the same response group. If the node is pure, the index will be equal to 0, and 0.5 when the response groups of the training sample of a node are equally represented. However, the sensitivity of the decision tree to the training examples leads to generate variant classification models causing overfitting and leading to poor generalisation performance when applied to new observations. Therefore, aggregating multiple trees like bagging or boosting can considerably improve the predictive accuracy and robustness of the learning model.

Bagging is widely utilised in the context of decision trees since they notoriously suffer from high variance. Bagging is a portmanteau of *"**B**ootstrap **agg**regat**ing**"*. As mentioned previously, bootstrapping is a resampling technique that generates multiple training sets from the original dataset by repeatedly sampling with replacement, and it is used to measure the uncertainty of the statistical properties of learning models. To account for variance in performance estimation of a learning model, training sets generated as bootstrap samples from the original dataset are used to construct multiple classification models and then averaging the obtained results. Therefore, the bootstrapped training sets are used to construct Bagging Decision Trees (BDT), then the obtained predictions of individual trees are averaged. In the testing stage, the response group of a new observation that is predicted by each tree model is recorded, and the most commonly predicted response group among all the predictions of the tree models is assigned to the new observation.

## 2.6 Feature Mining Approaches for Biomarker Discovery

In 1986, George Box [39] stated that *"a large proportion of process variation is explained by a small proportion of the process variables"*. In the context of biomarker discovery, detecting a small group of robust molecular markers underlying the variations of different groups of samples would be the most cost-effective procedure in developing reliable and explainable diagnosis and prognosis models. Therefore, a large body of research in the literature has adopted feature mining techniques, which can map a high number of features into smaller useful representations, which can be utilised thereafter for the development of various ML systems. In general, the feature mining approach can be classified into two major techniques, which are *feature extraction* and *feature selection*. Feature extraction involves a linear or nonlinear transformation of the original data from large feature space to a relatively lower dimensional space by minimising information loss (e.g. Principle Component Analysis [246], Auto-Encoders [24]). The extracted features are combinations of the original ones, different and more likely smaller, thus a disadvantage of this technique is that it is difficult to determine which subsets of the original features constituted the new transformed representation.

Feature selection, on the other hand, involves searching the search space of a dataset to find the best possible subset of features with respect to evaluation measures. The identified features are a compact and informative subset of the original ones. Unlike feature extraction, no new latent representations are formed. This property of not altering the original features has led to the widespread study of the feature selection approach because it enhances the comprehensibility of the obtained results. The research of feature selection remains an active area in ML domain (e.g. [94,142,266,272]). ML methods coupled with feature selection techniques have been widely utilised in bioinformatics research [2, 82, 144, 201, 250]. The feature selection approach has contributed to improving the predictive performance and robustness of classification models and decreasing its computational cost. The importance of feature selection can also be found in adding an explanation to the problem at the hand. In our research problem, the utilised genomic and proteomic datasets contain thousands of genes or proteins whose relevance to the response groups is not recognised by domain experts. Discovering a small subset of robust biomarkers helps biologists and variants thereof to investigate the relation of these molecular markers to the disease or clinically relevant outcomes.

The success of the feature selection approach mainly depends on considering two aspects: effective search methods to navigate the search space of the data and find the best possible subset of candidate predictors; and evaluation measurements to assess the quality of the features and guide the search process. Therefore, the biomarker discovery problem is discussed in the next sections based on two forms: how to search exponential dimension spaces of omics data and how to assess the optimality of features.

### 2.6.1 Evaluation Measurements

An effective selection method needs an effective evaluation criterion to detect relevant features. Different selection paradigms have been proposed, which are mainly *filter*, *wrapper*, and *embedded*. These paradigms can be grouped into classifier-independent approaches - (filter methods) and classifier-dependent approaches - (wrapper and embedded methods). The embedded methods [175] incorporate the feature selection process as part of the training process to reduce the computational time required for reclassifying different subsets, which is undertaken in wrapper methods.

Filters assign importance scores to features based on statistics of the data, without dependency on any particular classifier [120]. Univariate filters are the most common filter methods such as Welchs t-test [315], mutual information [301]. Saeys et al. [250] have highlighted the practical features of those filter techniques by claiming that *"even when the subset of features is not optimal, they may be preferable due to their computational and statistical scalability"*. Multivariate filter techniques such as minimal-Redundancy-Maximal-Relevance (mRMR) [229], and Fast Correlation-Based Filter (FCBF) [332] take the interdependence between features into consideration, thus better outcomes could be obtained. However, they tend to be computationally more expensive and statistically less scalable than univariate filter methods.

Wrapper methods simultaneously analyse groups of features based on an accuracy criterion involving a classifier and is therefore a classifier-dependent approach [167]. Wrappers consider the underlying dependencies among features and may perform better than filters. However, for such high dimensional data, the high risk of computational cost and overfitting restricts the use of wrapper methods to estimate the goodness of combinations of predictors for a given classification task. In this project, the merits of filters and wrappers are integrated into a hybrid evaluation measurement to estimate the optimality of the candidate subsets of features and will be optimised to discover the best possible combination in these large feature spaces - (as discussed in Chapter 4). The hybrid selection approach is considered another class of feature selection proposed to handle the curse of dimensionality (e.g. [28, 119, 192]).

A review study of the feature selection techniques applied to bioinformatics, including microarray and mass spectrometry data, has highlighted the multivariate selection algorithms as one of the most promising future lines of research for the bioinformatics community [250]. However, multivariate selection methods are more likely to identify several subsets of candidate predictors with similar classification accuracies making it difficult to ascertain the optimum subset. The feature preferences based on multivariate selection methods could be very sensitive to data sampling, thus it can be considered less robust and more unstable than filters. This is especially valid for multivariate selection methods, which search through a dataset with high dimensional feature space and a small number of cases. Therefore, the development of bespoke ensemble feature selection approaches is considered the second line of future research, particularly for knowledge discovery from HDSSS datasets [126]. Ensemble feature selection is similar to ensemble learning in that it relies mainly on performing multiple selectors, and the outcomes of these independent selectors are integrated into ensemble results (e.g. [100, 173]).

Abeel et al. [2] introduced an ensemble feature selection method based on Support Vector Machine - Recursive Feature Elimination (SVM-REF) and bootstrapping method to address the challenges of sampling variation for biomarker identification from high-throughput microarray datasets. A number of different selectors are used, and the outputs of these separate selectors are aggregated and returned as the final (ensemble) result. Several ensemble feature selection methods introduced in the literature [329] adopted the bootstrap procedure to address the challenges of sampling variation when using HDSSS data. The bootstrap method is a resampling technique that generates data sets by repeatedly sampling with replacement from the original data. However, each bootstrap set has significant overlap with the original data, which could lead to an optimistically biased estimation of the performance. Therefore, in this project, an ensemble selection approach based on repeated cross-validation procedure is employed to enhance the robustness of the finally selected subsets of predictors where no overlap can be found between the validation partition $k$ and the $k-1$ training sets, which is crucial factor for estimating the feature preferences as well as prediction performance reliably.

### 2.6.2 Search Methods

Biomarker discovery can be viewed as a feature selection problem. Suppose we have a dataset $D$ of $n$ observations $\{x_i, y_i\}$, where $x_i$ is a $d-$dimensional feature vector, and $y_i$ is the target class. A biomarker discovery problem is to find a $nFeat-$dimensional vector of key genes, where $nFeat < d$, whose expressions assign new observations into their response groups $y_i$ with a good level of certainty. However, due to the high dimensionality of omics data, the search space, $S$, grows exponentially when the size of $d$ increases due to the relationship $S = 2^d$. Therefore, finding the optimal subset of genes from these large-scale datasets can be computationally expensive for traditional search algorithms and infeasible in most situations because it requires an extremely large number of evaluations. Therefore, solving high dimensional selection problems has shifted towards more suitable optimisation algorithms such as Evolutionary Computation (EC) algorithms [69], and EC term refers to all biologically inspired techniques.

During 1950 − 1960, several computer scientists and engineers separately studied the notion that evolution can be utilised as an optimisation tool for engineering problems [111]. The notion relies mainly on the iterative evolution of a population of candidate solutions to a given problem towards the optimal solution using *genetic operators*. It retraces the natures path to find the best possible solution in as little search time as possible. Therefore, EC methods have been successfully applied for solving a variety of optimisation and feature mining problems [101]. Genetic Algorithms (GA) [134, 135], Ant Colony Optimisation (ACO) [76, 77], Particle Swarm Optimisation (PSO) [84, 163] are all classic examples of EC algorithms. The collection of algorithms that utilise procedures inspired by natural systems like survival and reproduction of the fittest are known as Evolutionary Algorithms (EA).

Several studies have worked on comparing various EC algorithms in different aspects and for diverse kinds of problems including feature selection (e.g. [85, 324]). However, any effort to summarise or compare EC methods may depend on choosing which areas to be covered. The parallel adaptive search of a GA has been adopted in a wide range of potential areas in order to obtain solutions to high dimensional, complex, and nonlinear problems. The GA method has been employed to solve feature selection problems (e.g. [43, 143, 154, 223]). A study by Siedlecki and Sklansky [261] revealed that the GA had a solid capability to reduce the time needed for finding the best possible set of features from large datasets compared to other traditional algorithms. Subsequently, several researchers have shown the advantages of using GA as a search algorithm for feature selection [172, 241, 327].

In the literature, various evaluation measurements have been combined with the GA to solve feature selection problems (e.g. [50, 63, 282, 319]). A review of EC methods for feature selection in classification problems [70] has discussed integrating the GA with diverse evaluation paradigms and emphasised the hybrid selection approach as promising for large datasets. In this thesis, an ensemble evolutionary mining model based on a hybrid selection approach, which combines the merits of univariate and multivariate statistical techniques is introduced to navigate through the high dimensional spaces of omics data and identify a subset of informative and robust predictors.

## 2.7 Deep Learning Methods for Biomarker Discovery

As discussed in the previous section, feature mining has become a core process in the construction of prediction systems for HDSSS omics data. The performance of such prediction models relies heavily on the features on which they are given to associate with particular outcomes. Therefore, it is highly desirable to develop a feature learning model directly from the raw high dimensional data so that high-level abstract features can be automatically captured and used for prediction purposes. This automatic feature learning can advance the move toward Artificial Intelligence (AI) where high-level abstractions can be automatically discovered and used in a similar way to that of the human brain. *"An AI must fundamentally understand the world around us, and we argue that this can only be achieved if it can learn to identify and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data" - Bengio et al.* [22].

In the neural network literature, the emphasis has been made on the composition of multiple levels of nonlinearity and the transformation of the input signals from low-level representations into high-level abstractions [130, 248]. This type of deep feature learning allows us to mitigate against the curse of dimensionality and enhance the generalisability for high dimensional complex recognition problems. Deep learning (DL) can be defined as deep feature learning methods that consist of multiple layers of non-linear functions that are connected in a hierarchical fashion, where the output of units in one layer feed as input into the next or preceding layers so that complex functions can be constructed using the well-known stochastic gradient descent algorithm, back-propagation [177].

Deep feature learning models have been incorporated in diverse areas of Bioinformatics (e.g. [7,11,161,165,340]). Furthermore, such deep neural network models have been applied across different problem domains in healthcare area like clinical imaging (e.g. [44, 51, 195, 331]), electronic health record (e.g. [209, 233, 291]), and wearable sensor (e.g. [122, 253, 341]). Moreover, DL models have produced superior performance over traditional methods in a wide range of domains such as computer vision (e.g. [170, 277]), natural language processing (e.g. [58, 275]), speech recognition (e.g. [1, 73, 129]), and remote sensing (e.g. [309, 343]).

These deep feature learning models typically require substantial data to constrain their parameters and learn a useful hypothesis. Therefore, in many of these problem domains, a large number of samples are typically available to train a deep network model. Furthermore, the datasets are characterised by high signal-to-noise ratios, thus the deep feature learning models attempt to discover high-level abstract features that can recover the data and boost the generalisability. The problem of deep feature learning from HDSSS omics data is a significant challenge as there are relatively few patients, compared to the huge number of features stored about them. That means the number of variations underlying the high dimensional genomic or proteomic data is not adequately exploited due to an insufficient number of biological samples, which in turn have a low signal-to-noise ratio. Therefore, new DL-inspired paradigms are required to sufficiently model the meaningful complexity represented by those biological samples.

The most popular form of deep learning is the supervised approach. When the desired outcomes are known, the learning process relies on fitting the model to reduce the distance between the desired outcomes and the actual outputs and thus to adjust the internal parameters to shorten that distance according to some cost function (e.g. sum of the squared errors or log likelihood). Supervised learning procedures do not typically allow for self-taught learning where the model is free to identify and exploit more subtle patterns in high dimensional spaces [303]. Therefore, an unsupervised pre-training approach is utilised in this work to be an essential characteristic of the deep feature learning model proposed to exploit the unknown structure of HDSSS omics data for the goal of discovering useful knowledge. Moreover, research has shown that for a small dataset, the unsupervised pre-training approach that will be discussed in the next section produces better generalisation error [87].

### 2.7.1 Unsupervised Pre-training Approach

A novel unsupervised pre-training approach was presented by a group of renowned researchers in 2006 to advance the traditional method of training DL models: Restricted Boltzmann Machines (RBMs) [131] by Geoffrey Hinton, Auto-encoder variants [24] by Yoshua Bengio, Sparse coding variants [237] by Yann LeCun. *"A fast learning algorithm for deep belief nets"* paper, which was published in 2006 by Geoffrey Hinton et al. [131], introduced RBMs based on very interesting notions. Mainly, a deep neural network model can be learned based on the unsupervised pre-training approach hidden layer by hidden layer 'sequentially', where within each layer, the net attempts to discover a useful representation of its input, which may be a previous hidden layer of activations. This greedy recursive approach to transforming the data starting from the input layer, to form a hidden layer, which is then provided as input to a process to form another hidden layer provides a powerful means to create high-level abstract representations from detailed low-level representations. Moreover, previously learned knowledge by the greedy layer-wise approach can be passed as input to a supervised classifier model, such as an SVM or perceptron. That means the learning task can be conducted using a semi-supervised approach, with the goal of learning to discover a good representation of $X$ that shapes the input distribution $P(X)$, which is also relevant in part to discover the target $P(Y|X)$. Therefore, the discovered representations by the DL models can be shared between tasks. The identification of salient invariant features that make sense for several tasks is a highly desirable property.

The successful training approach presented by Hinton et al. [131], followed by Yoshua Bengio et al. [24] introducing Auto-Encoder based approaches to pre-training. The Auto-encoder approach to pre-training the weights to the hidden layers of a deep net is predicated on using backpropagation to learn the identity function of each layer (except the output layer) of the network, one layer at a time. For example, starting with the input layer, a feedforward network with one hidden layer is used to reproduce the input layer values on the output layer (a task known as the identity function). The resulting hidden layer is then used as input to another feedforward network, which is trained to learn the identity function to form a new hidden layer and so on. These hidden layers can then be stacked in-between the original input layer and the output layer to form the deep net. The weights between the input and hidden layer of the respective auto-encoders are the weights used within the deep network and are now considered to be the pre-trained weights. The output layer can now be added and a global-fine tuning stage applied, based on supervised criterion using standard backpropagation [132]. Several studies have demonstrated the potential of the Auto-Encoder to discover intrinsic structure in high dimensional spaces and obtain better classification performances [150, 220, 260, 302]. Ranzato et al. at NIPS2006 (i.e. Neural Information Processing Systems) [237] presented a Sparse Coding Variants by adding the sparsity penalty on the hidden layer in order to boost the free energy of all units.

In this thesis, a new deep feature learning model is proposed, based on a set of non-linear sparse Auto-Encoders constructed on the under-complete representations to discover and interpret important signals from omics data that can be utilised to develop approaches to personalised and precision medicine. Learning sparse compressed representations of increasing complexity from HDSSS omics data forces the neural network to discover a small fraction of the possible factors that can recover a large proportion of variations underlying the data. These types of *expressive* representations capture high-level abstract features, which are characterised of being invariant to most of the irrelevant projections while collectively perceiving the information that approximates the input distribution.

### 2.7.2 Interpretation Methods for Deep Learning

Deep feature learning models perform multiple levels of transformations to the input features in order to abstract the problem. That means, the discovered representations are combinations of the original features, different and more likely smaller. Thus, identifying which features constituted the latent representations and were responsible for deriving such predictions is very challenging. However, a model that is able to state which phenotypes are key determinants is a crucial element of prediction systems used by health practitioners and decision-making professionals. It is therefore important to provide some explanatory capability to such deep learning models.

In the literature, few attempts have investigated going beyond the prediction to understand the machinery of the deep feature learning model and interpret its outcomes. Tan et al. in [284] and later in [283] have examined the significance of each neuron by computing its activity value in a single layer Auto-Encoder and for each sample. Such models are considered shallow Auto-Encoder models as they typically only contain one hidden layer in-between the input and output layer. For example, the shallow network in [284] contains 100 hidden neurons and [283] auto-encoder contains 50 hidden neurons in order to allow the manual interpretation of these nodes, which cannot be generalised to the deep models with higher capacities.

More complex hierarchical representations can be formed by recursively autoencoding the hidden layer of the original shallow autoencoder - this is known as stacking the autoencoder. Danaee et al. [66] map back the lower dimensional representations of the Stacked Denoising Auto-Encoder (SDAE) to the original data to detect what they called the Deeply Connected Genes (DCGs). The interpretation method of SDAE results in a $500 \times G$ matrix, where $G$ is the number of genes in the gene expression data and 500 seems to be the code dimension. The authors state that genes with the largest weights in the detected matrix are the DCGs. However, it is not clear how they defined the DCGs especially when each gene has 500 values and there is no evidence whether they have considered the largest weights in the positive or the negative direction. On the other hand, for biomarker discovery models, the predictivity, stability, and generalisability should be considered equally in order to report practically significant findings.

Therefore, in this research project, a new weight interpretation method is proposed to add explanatory power to the proposed deep feature learning model and identify a reduced set of highly predictive and robust biomarkers that are generic across independently generated datasets. The presented weight interpretation method will shed light on the innovative way to provide the explainability to such DL methods, and it can be utilised as a promising tool to discover unexplored knowledge in different domains.

## 2.8 Approaches for Validating Predictive Performance

The main objective of identifying robust biomarkers is to develop a reliable prediction system that generalises well when applied to new cases. Generalisability of a classification model can be defined as its ability to correctly estimate the response groups of unobserved sample cases (that were not included in the training data). Practically, if a classifier is fit in a setting far from its true function, estimation bias error is encountered. Variance error is reported when a classifier differs over the variation in the data. Variance refers to the amount of change that a classification model would have if it is estimated using different training sets. The classification model is learned using training data, and different training sets may fit different classifiers. However, in the typical case, the estimation of the classifier should not differ significantly when it is learned using variant training sets. If the classification model suffers from high variance, the little variation in the data could lead to the large variation in the classifier.

As a result, high bias and low variance refer to underfitting, where overfitting refers to low bias and high variance. The relative rate or the trade-off between bias-variance quantities should be achieved to some extent so that the expected test error is minimised when a classification model has low bias and low variance simultaneously. Therefore, the bias-variance trade-off term refers to the relationship between bias, variance and test error. It is called trade-off due to the fact that low biased and high variance model or high biased and low variance model can be obtained easily and the real challenge is to find a learning model that can improve on this trade-off.

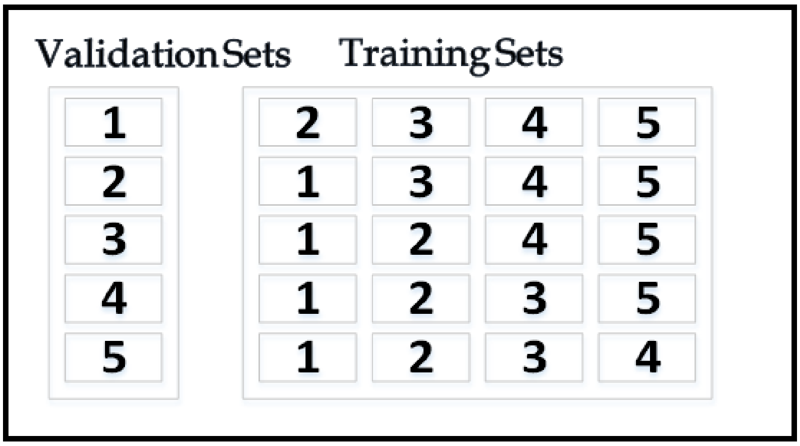
To quantify the performance of the classification model, the training error rate, which is the proportion of misclassified training samples, is usually utilised. However, the interest is in testing or generalisation error rate that results from classifying test samples that were not seen during the learning process because the training error rate can differ from the generalisation error rate. Therefore, the simplest popular approach is to partition the original data using Cross Validation (CV) approaches into training and validation sets. The training data is used to develop the learning model, while validation data is used to validate its predictive performance. The CV procedure has been utilised in many machine learning tasks like classification and regression to estimate the generalisation error rate. However, there is an increased risk of obtaining high variability in the estimation error with a small dataset. Therefore, the choice of the suitable validation method for such datasets should be considered carefully in order to report reliable estimations. A discussion on the appropriateness of various CV approaches is provided, along with the advantages and limitations of each approach.

### 2.8.1 Hold Out Validation Approach

This approach includes holding out a set of available observations, where the model is fit on the training samples and validated on the held out samples. The validation error of hold out approach, that is assessed in terms of misclassification rate, estimates the generalisation error. However, the validation error can vary according to which samples are involved in the training set and which samples are involved in the validation set. Moreover, since the model is trained using only the samples that constitute the training set, the learning model usually performs worse when fitting on fewer examples. Leaving more examples for the validation set leads to increase the estimation bias error, while having more training examples could lead to degrading the estimation process. As a result, the validation estimation of hold out approach is impractical with small sample size data.

### 2.8.2 Leave One Out Cross Validation Approach

The Leave One Out Cross Validation (LOOCV) approach involves temporarily leaving out one sample from a dataset of size $n$, and then training the classification model on the remaining $n-1$. Therefore, the LOOCV procedure tends to overcome the limitation of wasting the data by holdout validation approach, especially for small datasets. Since only the excluded observation was not included in the learning procedure, the biased error can be minimised. However, validating the performance of the classification model based on a single observation causes highly variable estimation. The LOOCV procedure is iterated $n$ times (i.e. $n$ is equal to the total number of observations), in each iteration a different case is taken out. Therefore, LOOCV is computationally expensive, since the model



| Validation Sets | Training Sets | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| 2 | 1 | 3 | 4 | 5 |
| 3 | 1 | 2 | 4 | 5 |
| 4 | 1 | 2 | 3 | 5 |
| 5 | 1 | 2 | 3 | 4 |

Figure 2.5: An illustration of k-fold cross validation approach.

has to train $n$ times. This can be very time consuming when the analysis model comprises two processes: mining and learning since both processes should be performed solely on the training dataset. The potential for the training process to become intractable is increased dramatically if every single model is slow to fit.

### 2.8.3 k-fold Cross Validation Approach

The $k-$fold CV approach randomly splits a dataset into $k$ non-overlapping partitions or folds of roughly equal size, where $k-1$ partitions are used to fit the model, and the remaining set is used to validate its performance as shown in Figure 2.5. The $k-$ fold CV method is used usually to overcome the drawbacks of holdout and LOOCV approaches by having fewer data to waste than holdout approach and $k$ times less expensive than $n$ times of LOOCV procedure. Furthermore, $k-$ fold validation procedure guarantees that there is no overlap between the samples of both sets which is a key factor for estimating the generalisation error rate of the prediction models accurately. In stratified CV, the folds are stratified so that each fold contains approximately the same proportions of response groups as in the original data, and there is evidence that this can enhance the estimation process [320]. The bias error is minimised with LOOCV method when $k = n$, however, our concern is not coming only from the risk of high bias. For small sample size datasets, the risk of high variance is increased, thus the $k-$fold is more appropriate than LOOCV estimator, which shows high variance. Commonly used values of $k$ are 5 and 10 because these values have been proven empirically to estimate generalisation error rate that can provide a good compromise for bias-variance trade-off [148]. Repeated $k-$ fold cross validation procedure should be employed to account for variance in performance estimation, and average results should be reported [41].

## 2.9 Metrics for Estimating Predictive Performance

When developing a machine learning model to perform classification, the main objective is to develop a classifier function from labelled training samples and measure its predictive performance on testing samples (that were not seen during the training process). Binary classification is the problem, where the observations belong to two response groups: Positive vs Negative. In medicine and biology, the positive group refers to cases affected by a medical condition, where the negative group refers to control cases. True Positives (TP) are the cases when the actual group is positive, and the predicted group is positive (e.g. when the model assigned the case to a cancer group, and the patient is actually suffering from cancer). True Negatives (TN) are the cases when the actual and predicted groups are negative (e.g. when the model assigned the case to control group and the case is not having cancer). False Positives (FP) are the cases when the actual group is negative (disease free), and the predicted group is positive (when the model assigned the patient to the diseased group and the case is not having cancer). False Negatives (FN) are the cases where the actual group that the patient belongs to is positive (i.e. suffering from cancer), and the model predicted the patient as belonging to the group without the disease. A variety of evaluation metrics are derived from these measurements.

Accuracy is the most common and simple evaluation metric, and it can be defined as the number of correct predictions $(TP + TN)$ classified by the model overall predictions $(TP + TN + FP + FN)$. However, the confidence of the class prediction (e.g. 0.57 or 0.97) is discarded in the accuracy assessment metric. Therefore, the estimation of the accuracy metric may not be reliable for datasets whose response groups have considerable disparate sizes, due to its bias towards the majority. True Positive Rate (TPR), *Sensitivity*, or *Recall* metric measures the proportion of positives ($TP/(TP+FN)$),(e.g. the proportion of actual cancer cases that were assigned by the model to the cancer group). True Negative Rate (TNR), or *Specificity* metric measures the proportion of negatives ($TN/(TN+FP)$) (e.g. the proportion of control cases, that were diagnosed by the model as non-cancerous). *Precision* or *Positive Predictive Value (PPV)* metric measures the ability of not diagnosing positive case as negative ($TP/(TP+FP)$) (e.g. the proportion of cases that were diagnosed as having cancer, and they actually had). A model with a high level of performance should have a high percentage of precision and recall.

The Area Under the Curve (AUC) of the (Receiver Operating Characteristics) (ROC) can be utilised for measuring the predictive performance of learning models. The focus of AUC is essentially on two measurements, which are TPR and FPR. By considering other evaluation metrics, the recall is identical to TPR, thus essentially the difference is between precision and FPR. The precision metric relies on false positives while FPR measures the true negatives. In imbalanced class data, when the majority are the negative cases, high percentage of $TN$ are more likely to be existent in the FPR due to ($FP/FP+TN$), resulting in smaller FPR. On the other hand, the majority of negative cases would not impact the precision due to the fact that this metric quantifies the number of $TP$ out of ($TP+FP$). Therefore, precision tends to the positive group than to the negative group. As a result, in imbalanced group data, when the minority are positive cases, and the interest is on identifying correct positives than correct negatives, the precision metric can be utilised for reliable estimation. In this research project, we have imbalanced group breast cancer datasets, where the majority of samples for both response groups are positives (ER+, PR+). Detecting the negative cases will be hard due to the insufficient number of cases. Therefore, AUC based on FPR and TPR can quantify the quality of the classification model when the positives are the majority.

## 2.10 Internal and External Validation

To assess how well our feature mining models work, we need to quantify their performances. To evaluate the performance of a model, the objectives of the model should be identified. The aim of proposing these mining models is to discover biomarkers that have good predictive power to distinguish the positive observations from the negatives effectively. That means the predictive performance of classification models is employed to assess the quality of the discovered biomarkers. In the literature, the assessment of feature mining methods has been restricted to the predictive performance of a classification model built on the selected subset of candidate predictors. Although most omics data analysis studies aim to select differentially expressed genes, to be used as biomarkers, there are only very few genes in common. Therefore, the reliability of the reported genes and their biological significance have raised doubts and questions [145]. The lack of overlap among the published genes is essentially caused by that various selected subsets of features produce similar predictive performances. Selecting a set of relevant predictors from small datasets could be very sensitive to which observations are included in the mining stage. Therefore, evaluating the outcomes of feature mining models based on the metric of classification's performance is not sufficient to detect true biomarkers from false positives robustly.

To detect true biomarkers, another criterion should be examined. The stability of selection is considered as another important aspect. Stability refers to the insensitivity of the feature mining model to the variations in training data. In other words, stability examines how the variations in the data can impact the feature preferences of the feature mining models. The stability can be conducted on the same dataset by having multiple random training-validation partitions, thus it can be described as an internal validation metric. If with different sets of training samples, the identified subsets of features differ radically, then the feature mining model is unstable. If the model is unstable, the confidence in the discovered biomarkers is decreased to prevent drawing unreliable biological conclusions. Jurman et al. [153] state that the predictivity and the stability of the detected biomarkers should be considered equally. The researchers in [110] argue that *"Identifying reproducible yet relevant features is a major challenge in biolog-*

*ical research"*. Therefore, data scientists need to assure not only the predictive power of the selected predictors but also their robustness to the variations in the data. In the literature, different metrics have been proposed to quantify the stability like a ranking, a weighting or a subset of features [67,81,112,155,258,306]. However, there is no universally accepted measure to estimate the true stability of a model's generalisation performance. Also, most of the stability measurements like scoring or ranking methods need a threshold to produce a stable subset. Thus, in this research, the stability of the proposed feature mining models is examined through the consistency of selection over different splits of a dataset. *"If the same features are selected in multiple independent iterations, they more likely are reliable biomarkers"* [100].

The external validation of feature mining models is of significant importance due to its role in evaluating whether the proposed models will produce generic biomarkers over multiple independent datasets, which are collected from different sources and for different perspectives. However, few studies in the literature have adopted another dataset that is collected from different studies for validation despite the abundance of publicly available omics data. Recently, a review study in metabolomics [198] has stated that more than 900 research papers have been introduced during five years for biomarker discovery in this field only, however, the number decreases dramatically when some validation metrics are adopted. According to [33] *"External validation is essential before implementing prediction models in clinical practice"*. The discovered generic biomarkers can be used to develop reliable prediction models that would be helpful in making trustable clinical decisions. If the features are discovered across multiple independent datasets, they are more likely to be true biomarkers. *"External validation using data from a completely different study provides the highest irrefutable evidence that a tool validates"* [287].

## 2.11 Discussion

This chapter discusses the major research challenges of extracting relevant knowledge from HDSSS omics data, using the state-of-the-art approaches presented in the literature. Classical statistical techniques utilised in biology and medicine for biomarker discovery were covered in Section 2.4 with an emphasis on its technical advantages and limitations. The greatest challenges for ML algorithms to handle is the high dimensionality and relatively small sample size of omics data, overviewed in Section 2.5 and the directions for addressing these issues were discussed. Therefore, the feature mining was presented in Section 2.6 as one of the main possibilities that can be employed to address the curse of dimensionality issues and detect a small group of candidate predictors that could not be detected using conventional statistical techniques and learning methods alone. The feature mining approach was discussed in terms of how to assess and how to search thousands of genes, as illustrated in Section 2.6.1 and Section 2.6.2. The critical discussion of both Sections 2.6.1 and 2.6.2 highlighted the importance of solving the biomarker identification problem from HDSSS omics data as an optimisation problem leveraging the advantages of EC method and the hybrid evaluation measurement. Furthermore, the requirement for an ensemble feature mining model is asserted to enhance the robustness of the identified subsets of candidate predictors.

The latest innovations in DL was explored in Section 2.7 to investigate its potential to formulate nonlinear models that are able to effectively discover salient invariant biomarkers from omics data. The incorporation of deep neural network models in different problem domains was reviewed in Section 2.7, emphasising the importance of having substantial data to train these models effectively. Therefore, the technological gaps and needs for the development of new DL-inspired models were stated for the goal of inferring useful models from HDSSS omics data. The groundwork for the new deep feature learning model was laid in Section 2.7.1 by considering the unsupervised pre-training approach to be an essential characteristic of a DL model developed to exploit the unknown structure of HDSSS omics data.

The main obstacle of such DL models is the lack of transparency, which means the inability to understand why the models behave as they do. This may not be an issue for some domains because it can easily validate the obtained results. However, providing explainability to diagnosis and prognosis systems is a crucial factor to develop a reliable prediction model that can be understood by clinicians, and thus it can be employed in clinical practice. Therefore, Section 2.7.2 explored the few attempts in the literature that have been conducted to understand the machinery of the deep networks and interpret its outcomes, along with its main limitations and drawbacks. Therefore, in this thesis, we develop a novel weight interpretation method to add explanatory power to our deep feature learning model for determining the candidate input features that force the different biomarker classification behaviours.

The assessment of the outcomes of the proposed feature mining models for biomarker identification from omics data involves several underlying issues that need to be properly handled in order to report reliable findings. Therefore, diverse quantitative quality metrics for validating and estimating the prediction performance were discussed in Sections 2.8 and 2.9, while the external validation using multiple independently generated datasets was explained in Section 2.10.

# Chapter 3

# Datasets and Experimental Methodology

## 3.1 Introduction

This chapter focuses on the datasets and the experimental methodologies used for model fitting and selection and therefore to validate the feature mining models proposed. Typical high-throughput biological data are more likely to contain a large number of noisy variables and molecules with unreliable measurements that can be considered indistinguishable from noise. This chapter discusses the methods of filtering out genes from genomic datasets that are not reliably expressed. This thesis utilises 18 publically available HDSSS biomedical datasets to examine the potential of the presented feature mining models to discover robust and generic biomarkers.

Four breast cancer datasets are illustrated in this chapter, along with the explanation of the pre-processing step required to produce the required information. The adopted breast cancer datasets are linked into two response groups, which are Estrogen Receptor (ER) status and Progesterone Receptor (PR) status forming eight breast cancer datasets. Moreover, the breast invasive carcinoma datasets are collected from different studies, but using the same microarray technology, are integrated in a number of different approaches to have more substantial and balanced group data. The integration provides an additional nine breast cancer datasets: three breast cancer datasets with ER groups, and six datasets with PR groups. On the other hand, the ovarian cancer dataset was employed in the preliminary experiments performed in this research to develop and validate the proposed models. The empirical assessment of the performance of the introduced feature mining models was conducted in terms of the objectives of these models, which are predictivity, stability, and generalisability.

## 3.2 Filtering Methods

Gene expression datasets typically contain thousands of genes, not all of these information are relevant. Bichsel et al. [32] state that *"very few genes are in fact significantly changed in expression in a way that is distinguishable from biological and measurement variation and noise"*. Therefore, the genes that seem to generate uninformative signals can be considered as noise. In the microarray literature, several studies have revealed the potential of filtering out genomic datasets from genes with unreliable measurements to enhance the detection of differentially expressed genes [106, 208, 281, 292]. Diverse filtering methods, based on different criteria, have been proposed for excluding genes that are not reliably expressed or represent experimental noise.

Gene expression datasets typically contain genes that exhibit little variation in their profile. A gene with small profile variance across the samples would not differ significantly among response groups. In this thesis, a filtering method based on variation criterion is utilised to remove gene expression profiles with a variance less than the 10th percentile from further analysis. Furthermore, gene expression datasets could have genes whose range of values may not well distributed (spiking behaviour). A filtering method based on low entropy criterion is utilised in this research to measure the amount of information about a variable and remove genes with low entropy expression values (i.e less than the 10th percentile). A more detailed discussion of these rudimentary filtering methods can be found in [166].

## 3.3 Breast Invasive Carcinoma Datasets

The development of the breakthroughs for extracting useful knowledge from omics data is at the core of personalised and precision medicine. However, we are able to benefit from these biological data only if they are publicly available. Recently, there is increasing pressure from funding providers and the patient community to gain the maximum benefit from produced data by sharing it with the research community regardless of whether biomedical studies are funded publicly or privately [169]. Analysing omics data over several research studies can help to control the risk of false positives, offer possibilities to innovative discoveries, and to report significant and reliable findings [57]. The availability of biomedical data repositories such as The Cancer Genome Atlas (TCGA)[1] and the International Cancer Genome Consortium (ICGC)[2] bring tremendous opportunities to health care research to benefit from this abundance of cancer genomic data.

TCGA is a collaboration between the National Cancer Institute (NCI)[3] and the National Human Genome Research Institute (NHGRI)[4] to understand the molecular basis of cancers, through the utilisation of genome analysis technologies. TCGA is one of the largest genomic data repositories, including more than eleven thousand cases, representing 33 cancers. TCGA has produced different types of genomic datasets like somatic mutation, copy number, gene expression, miRNA expression, DNA methylation, reverse protein phase array and clinical information for different types of cancers. These biological datasets are publicly available for every clinician, bioinformatician, statistician, and computer scientist to employ them for developing a wide range of analysis models. Since the completion of the TCGA project, TCGA data analysis is becoming a priority in order to provide a better understanding of the complicated mechanism of cancer so that the inferred knowledge can be transferred to personalised and precision medicine [164].

With the availability of a vast amount of genomic data, several web portals have been created to help researchers and graduate students to access and use

[1] http://cancergenome.nih.gov

[2] https://icgc.org

[3] https://www.cancer.gov/

[4] https://www.genome.gov/

these cancer datasets, using different types of exploration and analysis tools. cBioPortal [48,102] is an open-access repository for multidimensional cancer genomics datasets, and was originally developed at Memorial Sloan Kettering Cancer Center (MSKCC)[1]. cBioPortal aims to minimise the complexity of accessing various genomics projects and allow for diverse analysis and visualisation tools to be utilised. In this research, three breast cancer datasets were downloaded from the cBioPortal website, which are Breast Invasive Carcinoma (TCGA, Nature 2012), Breast Invasive Carcinoma (TCGA, Cell 2015), and Breast Invasive Carcinoma (TCGA, Provisional).

### 3.3.1 Breast Invasive Carcinoma (TCGA, Nature 2012)

This dataset originated as part of the TCGA (i.e the cancer study identifier is brca_tcga_pub), and it was used by a collaborative study between NCI and NHGRI published in 2012 [219]. The study was conducted on 825 patients with breast cancer invasive tumours and found that four main subgroups of breast cancer are caused by various types of genetic and epigenetic variations. According to the website of the National Institutes of Health (NIH), *"TCGAs comprehensive characterisation of their high-quality samples allow researchers an unprecedented look at these breast cancer subgroups"*. Various genomic and clinical datasets are included in (Nature 2012) data. The focus of the research in this thesis is on mRNA expression data. The mRNA expression data was carried out using Agilent microarray and contains 17268 genes and 526 observations. Two response groups were defined in this research, which are ER Status and PR Status. The samples with missing values/others (e.g. Performed but Not Available, Not Performed, Indeterminate) in both response groups were removed from the analysis as illustrated in Figure 3.1.

The integration of the mRNA expression data and ER clinical data, which has 780 observations, resulted in a dataset of 519 observations. According to the group distribution, 401(77.26%) are samples with ER+ tumours, and 118(22.74%) are ER- samples. The unification of the mRNA expression data and PR clinical data, which has 777 observations, resulted in a dataset of 518 observations,

[1] https://www.mskcc.org/

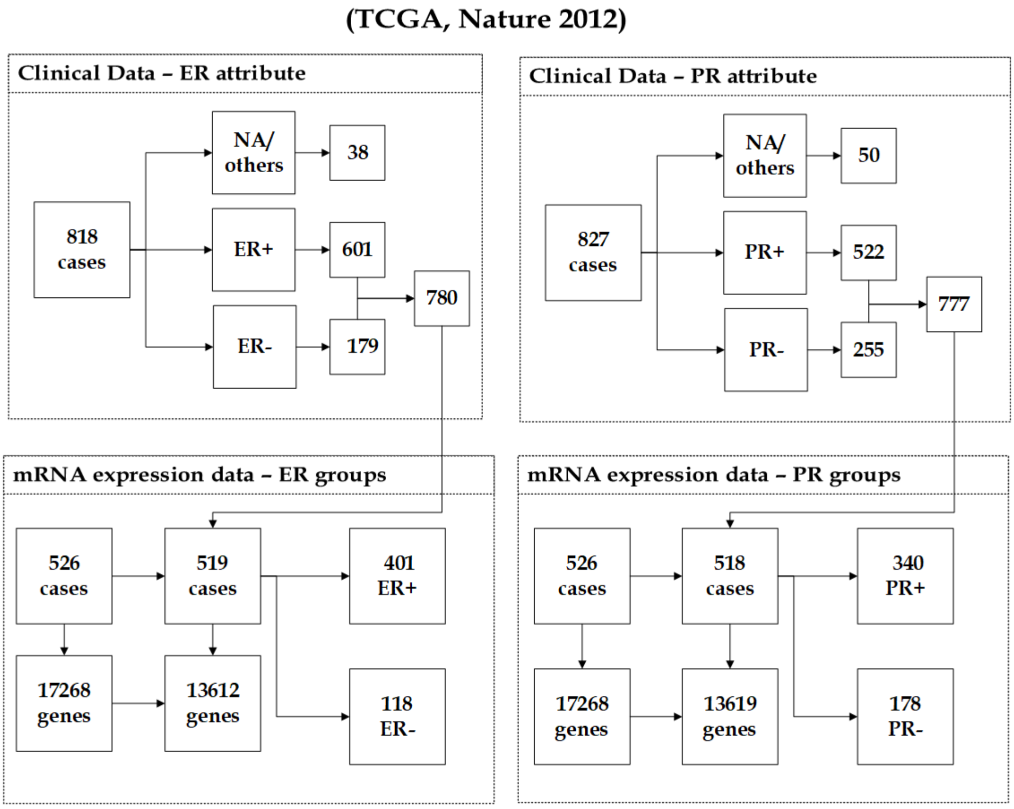


Figure 3.1: The description of (Nature 2012) dataset showing the unification of clinical data and mRNA expression data.

340(65.64%) being patients with PR+ tumours, and 178(34.36%) being PR- samples. Each mRNA sample contains 17268 genes. The filtering methods (which are described in Section 3.2) were utilised before the feature mining models take place to filter out the less reliably expressed genes from mRNA expression data. The number of remaining genes in (Nature 2012) dataset with ER groups is 13612 and 13619 with PR groups. Figure 3.1 illustrates the number of mRNA samples before and after the unification with the ER and PR clinical data, ER and PR group distribution across samples, and the number of mRNAs before and after performing the filtering methods.

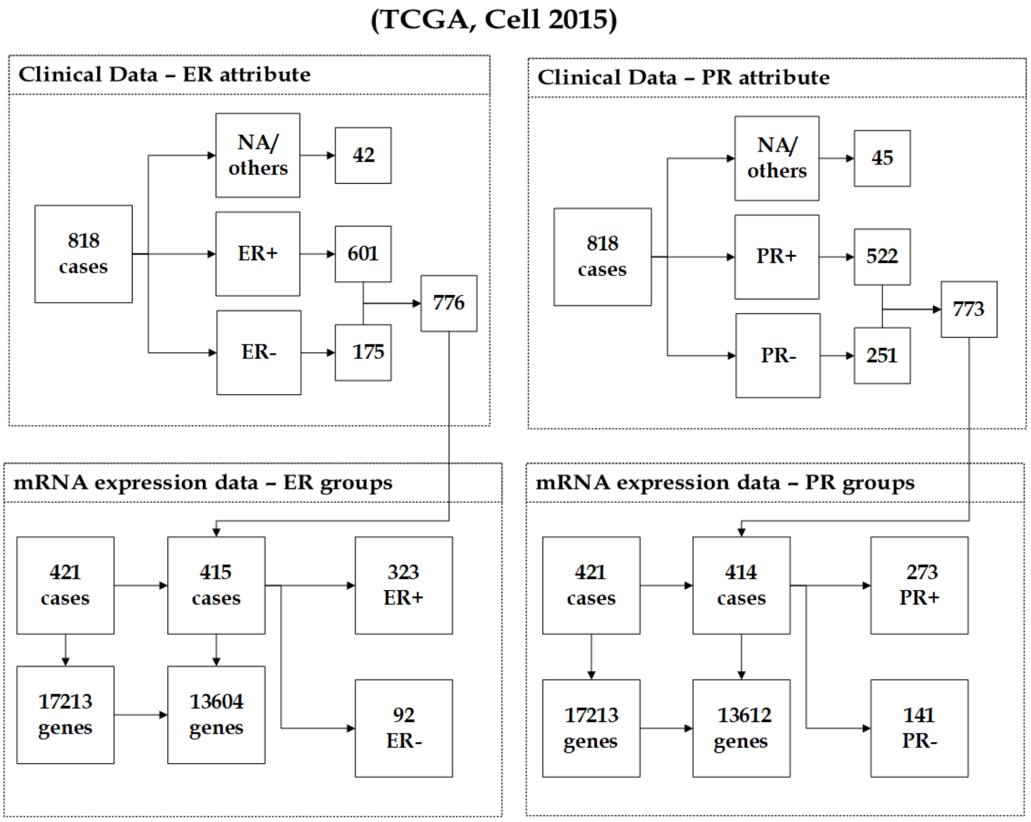


Figure 3.2: The description of (Cell 2015) dataset showing the unification of clinical data and mRNA expression data.

### 3.3.2 Breast Invasive Carcinoma (TCGA, Cell 2015)

The availability of TCGA data with high standard samples has motivated us to adopt other breast cancer datasets to examine the generalisability of the proposed feature mining models to a wider population. This dataset originated as part of the TCGA (i.e. the cancer study identifier is brca_tcga_pub2015), and it was used by the analysis study [55], which found that mixed tumours can be assigned into their subgroups using genetic features. Different genomic and clinical datasets are involved in (Cell 2015) data. The focus of the research in this thesis is on mRNA expression data. The mRNA expression dataset was carried out using Agilent microarray and contains 17213 genes and 421 observations. The samples with missing values/others (e.g. Not Available, Indeterminate) in both response groups were removed from the analysis, as explained in Figure 3.2.

The integration of the mRNA expression data and ER clinical data, which contains 776 observations, resulted in a dataset of 415 observations. According to the group distribution, 323(77.83%) are patients with ER-positive tumours, and 92(22.17%) are ER- samples. The unification of the mRNA expression data and PR clinical data, which has 773 observations produced a dataset of 414 observations, in which 273(65.94%) are PR+ patients, and 141(34.06%) are PR- samples. The number of the remaining genes of the mRNA expression dataset after applying the filtering methods is 13604 genes with ER groups and 13612 genes with PR groups as shown in Figure 3.2.

### 3.3.3 Breast Invasive Carcinoma (TCGA, Provisional)

This dataset originated as part of the TCGA (i.e. the cancer study identifier is brca_tcga), and it was used by the analysis study [234]. The study was conducted on 1098 breast cancer invasive tumours to estimate the prognosis of invasive breast cancer. The outcome is that ten genetic variations were detected to be statistically associated with histologic grade, which is one of the most important microscopic features. Diverse biomedical datasets are involved in (Provisional) data, including copy number alterations, gene mutation, mRNA and protein expression, clinical and pathological data. The focus of the research in this thesis is on mRNA expression data. The mRNA expression dataset was carried out using Agilent microarray and contains 17814 genes and 529 observations. The samples with missing values/others (e.g. Not Available, Indeterminate) in both response groups were removed from the analysis, as shown in Figure 3.3.

The integration of the mRNA expression data and ER clinical data, which has 1046 observations resulted in a dataset of 519 observations. According to the group distribution, 402(77.46%) tumours were derived from ER+ samples, and 117(22.54%) tumours were derived from ER- samples. The unification of the mRNA expression data and PR clinical data, which contains 1043 observations produced a dataset of 518 observations. The number of cases that were derived from patients with PR+ tumours is 341 out of 518, so the percentage of positives is 65.83%, while 177(34.17%) tumours were derived from PR- samples. After

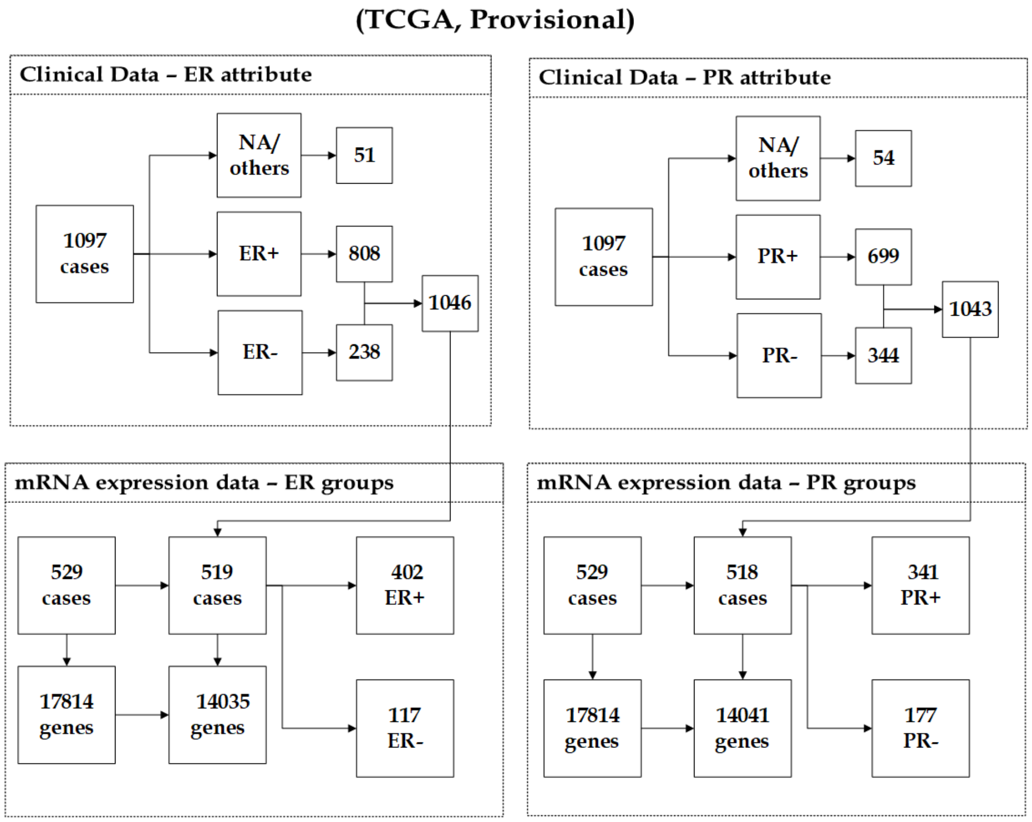


Figure 3.3: The description of (Provisional) dataset showing the unification of clinical data and mRNA expression data.

performing the filtering methods, the number of remaining genes of the mRNA expression dataset is 14035 genes with ER groups and 14041 genes with PR groups. Figure 3.3 illustrates the number of genes and mRNA samples before and after the pre-processing step, along with the distribution of both ER and PR groups across samples.

## 3.4 The Integrated Breast Invasive Carcinoma Datasets

In this section, the breast invasive carcinoma datasets that are collected from different studies, but were carried out using Agilent microarray, which are (Nature 2012), (Cell 2015) and (Provisional) are integrated in a number of different approaches to enhance the imbalanced class distribution of these datasets and also to have more substantial data. Moreover, the integration of the breast invasive carcinoma datasets helps to investigate the consistency of selection of the proposed feature mining models to a wide range of variations in breast cancer samples. To ascertain achieving the best possible balanced group datasets, three integrated datasets are created with ER groups, and six integrated datasets are created with PR groups, as discussed in the following sections.

## 3.4.1 The Integrated Datasets with ER groups

This section discusses the mechanism of integrating the breast invasive carcinoma datasets with ER groups. To attain a good-level of balanced group distribution, (**N**ature 2012), (**C**ell 2015) and (**P**rovisional) datasets are fused jointly in three different approaches:

### 3.4.1.1 NCP1 Dataset

This dataset was created, based on the integration of the negative samples of (Cell 2015) and (Provisional) datasets with (Nature 2012) data, which has 519 samples, 401 tumours were derived from patients with ER+, and 118 tumours came from ER- samples. The number of the negative samples of (Cell 2015) is 92, and of (Provisional) is 117. These ER-negatives were integrated with (Nature 2012) dataset to generate an integrated dataset called NCP1, which has 728 observations, as shown in Figures 3.4. The distribution of ER groups is that 401(55.08%) are samples with ER+ tumours and 327(44.92%) are ER- samples. As illustrated in Figure 3.4, the unification of the genes across the breast invasive carcinoma datasets resulted in a dataset that contains 13212 genes.

### 3.4.1.2 NCP2 Dataset

Alternatively, the creation of this dataset depends on that the negative samples of (Nature 2012) and (Provisional) datasets were added to (Cell 2015) data, which has 415 observations, 323 being patients with ER+ tumours and 92 being ER- samples. The ER- samples of (Nature 2012) (i.e. 118), and the ER- samples

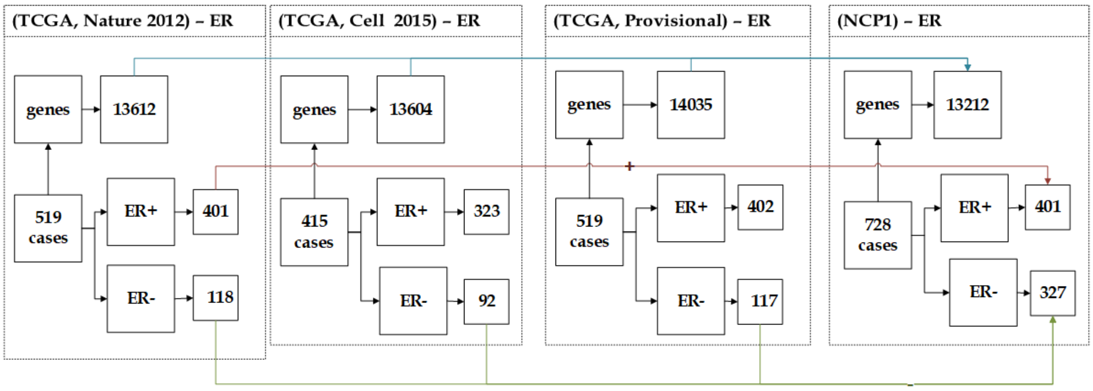


Figure 3.4: The description of NCP1 dataset showing the unification of (Nature 2012), (Cell 2015), and (Provisional) datasets.

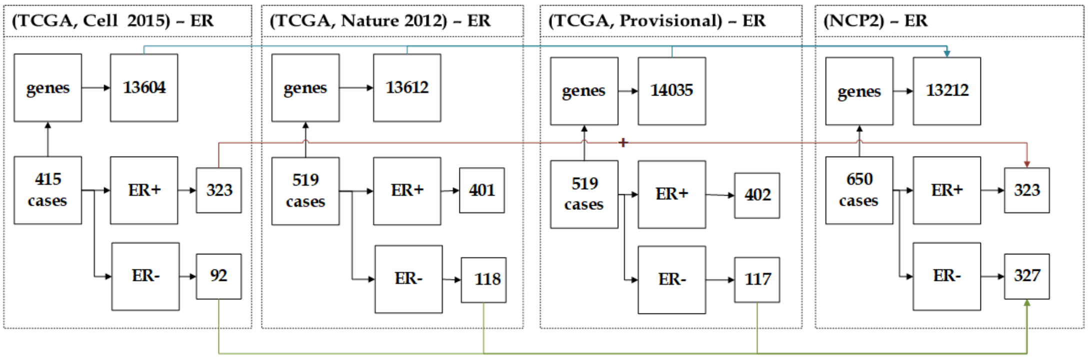


Figure 3.5: The description of NCP2 dataset showing the unification of (Nature 2012), (Cell 2015), and (Provisional) datasets.

of (Provisional) (i.e. 117) were integrated with (Cell 2015) data to produce an integrated dataset called NCP2, which has 650 observations. The class distribution of NCP2 dataset is that 327(50.31%) are ER-negatives, and 323(49.69%) are ER+ samples, as illustrated in Figure 3.4. The unification of the genes over the breast invasive carcinoma datasets resulted in a dataset that comprises of 13212 genes.

#### 3.4.1.3 NCP3 Dataset

This dataset was created based on integrating the negative samples of (Nature 2012) and (Cell 2015) datasets with the (Provisional) data, which has 519 obser-

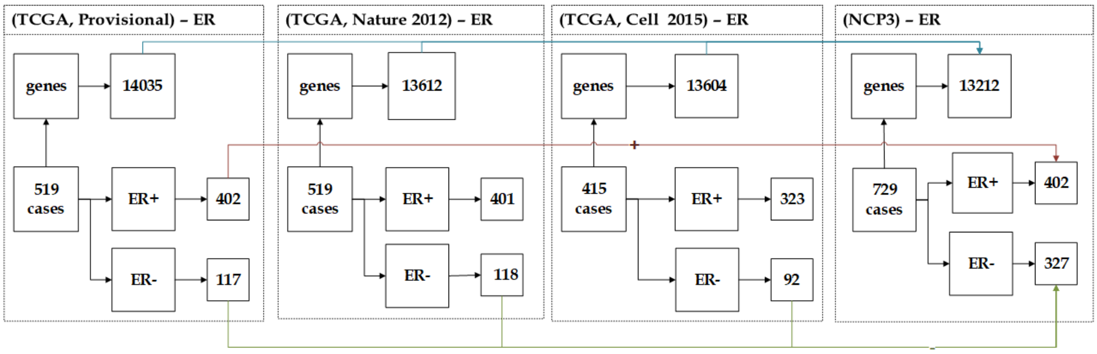


Figure 3.6: The description of NCP3 dataset showing the unification of (Nature 2012), (Cell 2015), and (Provisional) datasets.

vations, 402 being patients with ER+ tumours and 117 being ER- samples, as shown in Figure 3.6. The ER-negatives of (Nature 2012) (i.e. 118) and the ER-negatives of (Cell 2015) (i.e. 92) were integrated with (Provisional) data to create an integrated dataset called NCP3 of 729 observations. The group distribution of the NCP3 dataset is that 327(44.86%) are ER-negative samples and 402(55.14%) are patients with ER+ tumours, as clarified in Figure 3.6. Consequently, the integration of the genes over the breast invasive carcinoma datasets resulted in a dataset that contains 13212 genes.

### 3.4.2 The Integrated Datasets with PR groups

This section discusses the procedure of integrating the breast invasive carcinoma datasets that were carried out using the same microarray technology with PR groups. To achieve a good-level of balanced class distribution, two of the three datasets are fused jointly in six different approaches:

#### 3.4.2.1 NC Dataset

This dataset was created based on fusing the PR- samples of (**C**ell 2015), which are 141 with (**N**ature 2012) data, which has 518 observations, 340 being patients with PR+ tumours, and 178 being PR-negative samples, as shown in Figure 3.7. The integrated dataset NC has 659 observations, 340(51.59%) being PR+ patients and 319(48.41%) being PR- samples. As explained in Figure 3.7, integrating the

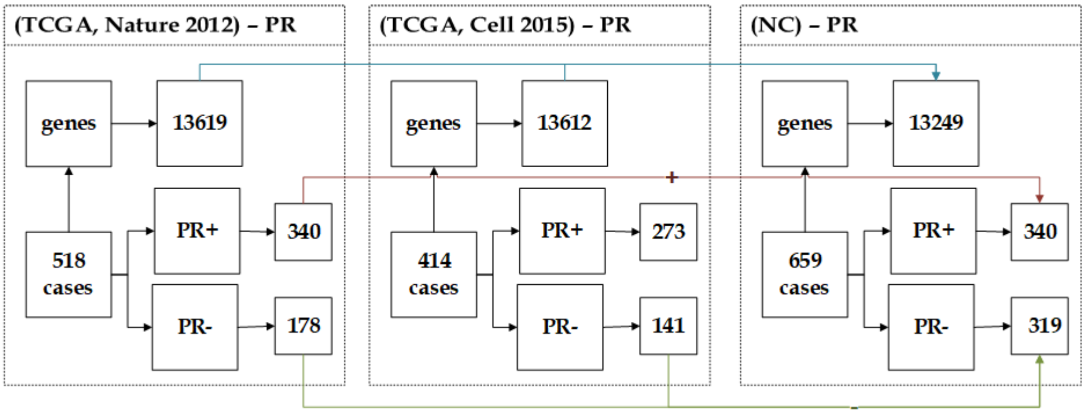


Figure 3.7: The description of NC dataset showing the unification of (Nature 2012) and (Cell 2015) datasets.

genes across the datasets resulted in a dataset that contains 13249 genes.

#### 3.4.2.2 CN Dataset

Alternatively, creating this dataset was based on that the PR-negative samples of (**N**ature 2012), which are 178 were added to (**C**ell 2015) dataset, which has 414 observations, 273 being patients with PR+ tumours, and 141 being PR-negative samples. The integrated dataset CN contains 592 observations, 273(46.11%) being PR-positives, and 319(53.89%) being PR-negatives, as illustrated in Figure 3.8. The unification of the genes over the datasets resulted in a dataset that contains 13249 genes.

#### 3.4.2.3 NP Dataset

This dataset is created based on the combination of the negative samples of the (**P**rovisional) dataset (i.e. 177) with (**N**ature 2012) data, which includes 518 observations, 340 being samples with PR+ tumours and 178 are PR- samples. The integrated dataset NP comprises 695 observations, 340(48.92%) being PR-positive patients and 355(51.08%) being PR- samples, as clarified in Figure 3.9. The integration of the genes over the datasets produced a dataset that involves 13528 genes.

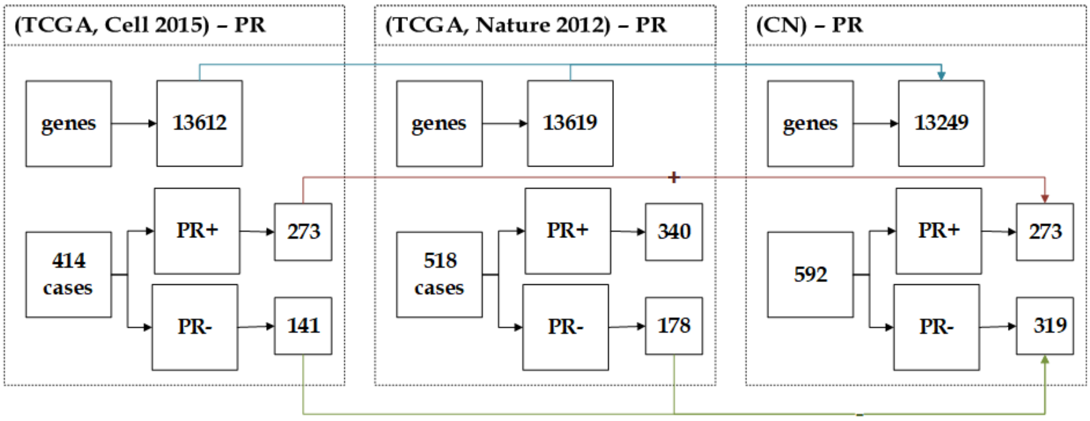


Figure 3.8: The description of CN dataset showing the unification of (Cell 2015) and (Nature 2012) datasets.

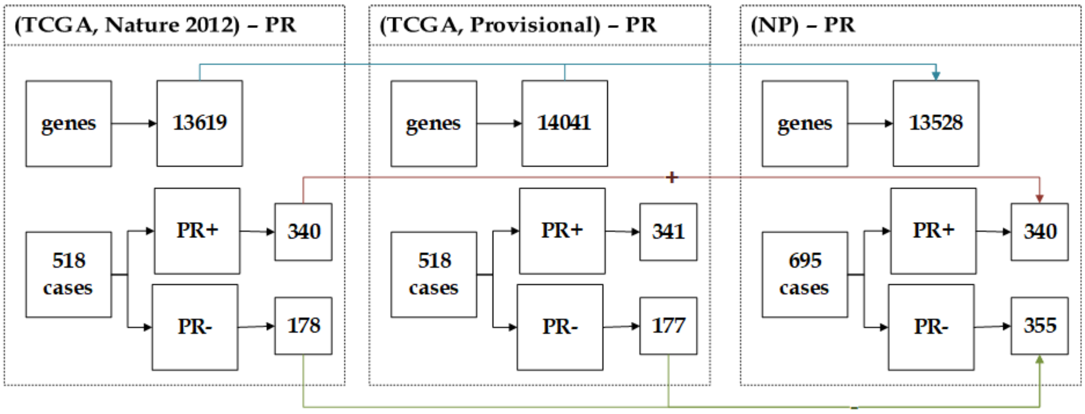


Figure 3.9: The description of NP dataset showing the unification of (Nature 2012) and (Provisional) datasets.

#### 3.4.2.4 PN Dataset

Alternatively, the creation of this dataset is based on that the negative samples of (**N**ature 2012) dataset were added to (**P**rovisional) data, which has 518 observations, 341 being samples with PR+ tumours and 177 were derived from PR- samples. The integrated dataset PN involves 696 observations, 341(48.99%) are PR-positives and 355(51.01%) are PR-negatives, as shown in Figure 3.10. unifying the genes across the datasets generated a dataset that comprises of 13528 genes.

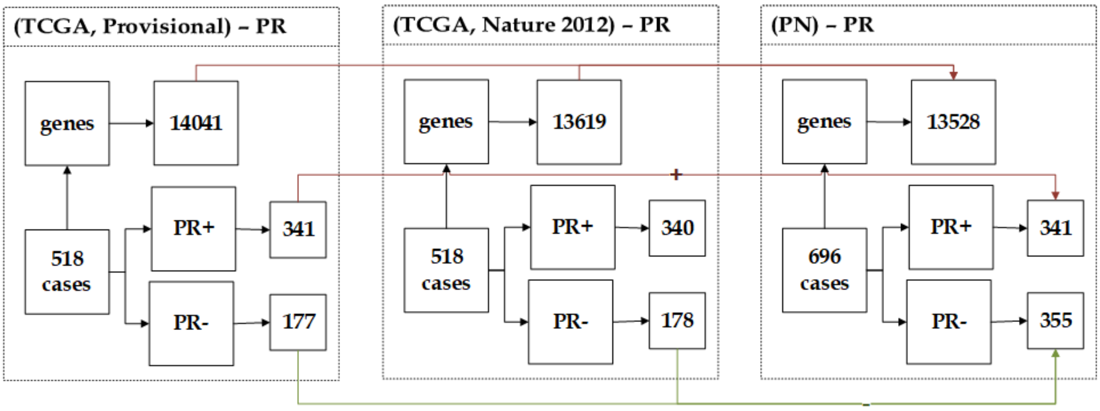


Figure 3.10: The description of PN dataset showing the unification of (Provisional) and (Nature 2012) datasets.

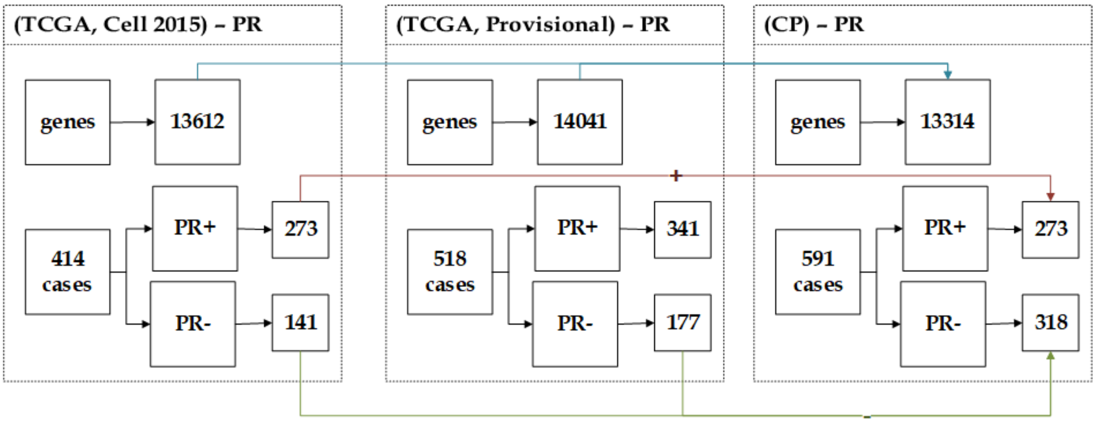


Figure 3.11: The description of CP dataset showing the unification of (Cell 2015) and (Provisional) datasets.

#### 3.4.2.5 CP Dataset

In this dataset, the PR- samples of (**P**rovisional) were added to (**C**ell 2015) data, which has 414 observations, 273 being samples with PR+ tumours and 141 being PR- samples. The integrated dataset CP contains 591 observations, as illustrated in Figure 3.11, 273(46.19%) being PR-positives and 318(53.81%) being PR-negatives. The integration of the genes over the datasets leads to a dataset with 13314 genes.

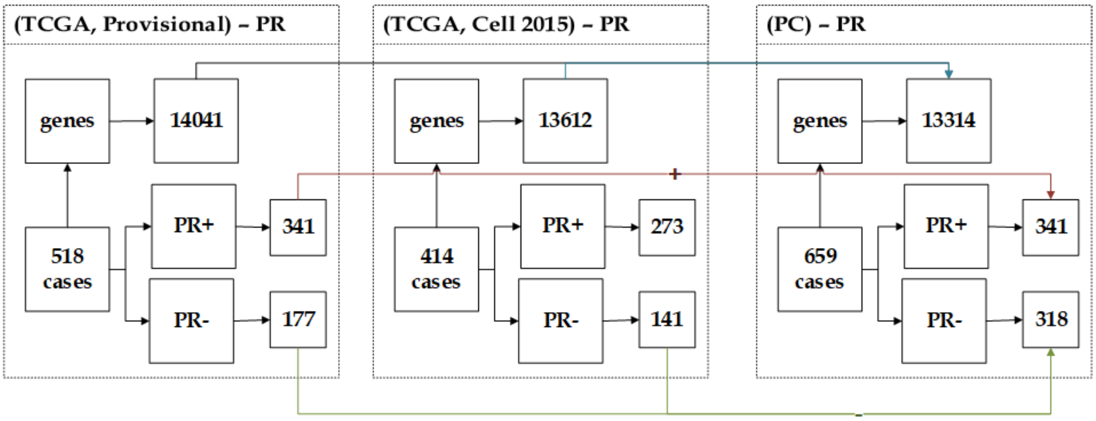


Figure 3.12: The description of PC dataset showing the unification of (Provisional) and (Cell 2015) datasets.

#### 3.4.2.6 PC Dataset

Alternatively, this datasets was created based on the combination of the PR-samples of the (**C**ell 2015) with (**P**rovisional) data, which has 518 observations, 341 being PR-positives and 177 being PR-negatives. The integrated dataset PC comprises of 659 observations, 341(51.75%) being PR-positives and 318(48.25%) being PR- samples. Unifying the genes over the datasets resulted in a dataset that has 13314 genes.

## 3.5 METABRIC Breast Cancer Dataset

This dataset was generated from METABRIC [62], [68] and downloaded from cBioPortal, where the cancer study identifier is brca_metabric. The dataset contains diverse biomedical datasets, including clinical data and two genomic datasets: gene expression, and copy number alterations. An integrative analysis study [230] performed on copy number alterations and gene expression profiles in 2000 primary breast cancer tumours emphasised the significance of genome-based stratification of breast cancer.

The mRNA expression dataset was carried out using Illumina Human v3 microarray and contains 24368 genes and 1904 observations. The integration of mRNA expression dataset and ER clinical data, which has 1980 cases, generated

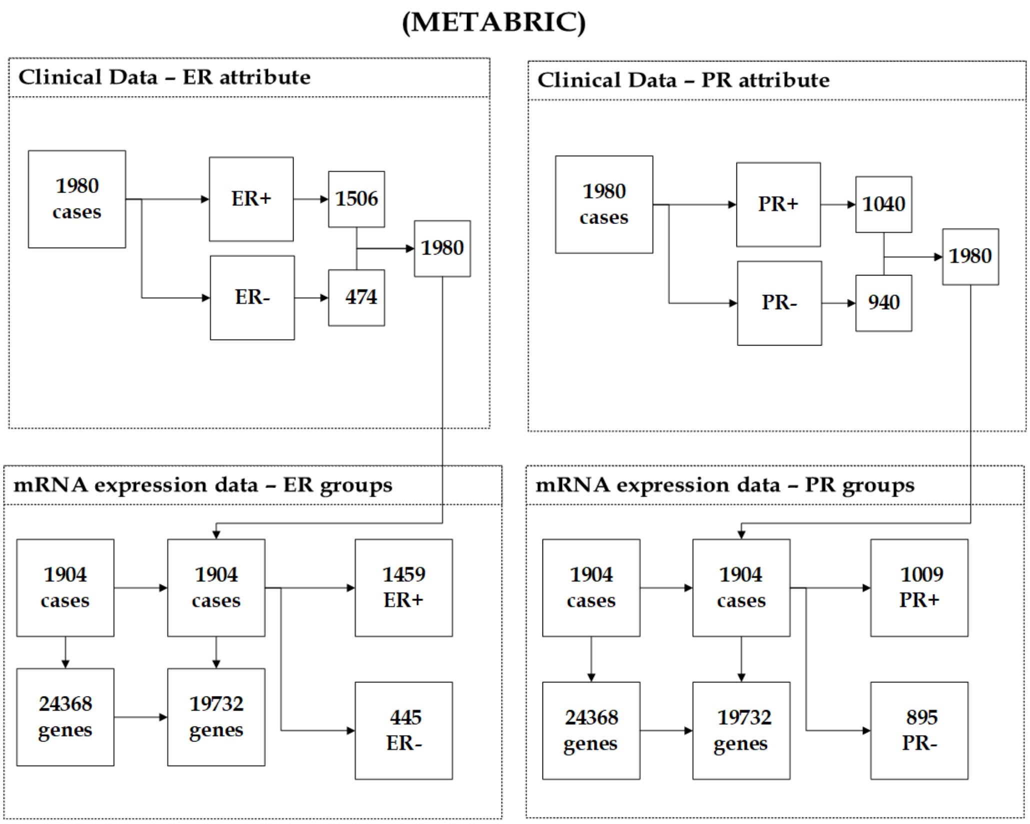


Figure 3.13: The description of (METABRIC) dataset showing the unification of clinical data and mRNA expression data.

a dataset of 1904 observations, 1459(76.63%) being samples with ER+ tumours, and 445(23.37%) were derived from ER- samples. The unification of mRNA expression dataset and PR clinical data that contains 1980 observations, resulted in a dataset of 1904 observations, 895(47.01%) being PR-negatives, comparing to 1009(52.99%) instances coming from patients with PR+ tumours. After eliminating the least promising genes from the analysis, the number of remaining genes of mRNA expression dataset with ER and PR groups is 19732. Figure 3.13 provides a summary of the number of samples and genes of mRNA expression dataset before and after the pre-processing step.

## 3.6 Ovarian Cancer Dataset

Ovarian cancer dataset is publicly available on the FDA-NCI Clinical Proteomics Program Databank website[1]. The dataset was utilised in the preliminary conducted experiments to develop and validate the introduced models. This high-resolution ovarian cancer dataset was generated using the WCX2 protein array to identify serum (blood-derived) proteomic patterns that differentiate the serum of patients with ovarian cancer from that of women without ovarian cancer. It contains records collected from 216 observations with 15000 features. Each sample has one of two possible response groups: Normal or Cancer. According to the group distribution, 121 (56%) instances were derived from patients with cancer, and 95 (44%) instances were derived from women without cancer.

## 3.7 Experimental Methodology

This section discusses the validation approaches and evaluation metrics applied to understand the performance of the proposed feature mining models.

### 3.7.1 Area Under the ROC Curve

The AUC metric is utilised as the main performance estimation metric to assess the quality of the classification models. As discussed in Chapter 2, AUC is more reliable than accuracy, more discriminative than other estimation metrics and can be measured over the range of TPR and FPR [188, 189]. Receiver Operating Characteristic (ROC) curve in Figure 3.14 shows TPR and FPR for the designed classification model using the marker on the figure. The FPR of (0.00) indicates that 0% of the validation samples are assigned incorrectly into the positive group. The TPR of (1.00) corresponds to 100% of the validation samples that are correctly classified to the positive group by the learned model. A perfect result is a right angle to the top left of the figure.

The ROC curve can be summarised into a single value by measuring the Area Under the ROC curve (AUC), which is a measure of the overall quality of the

[1] https://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

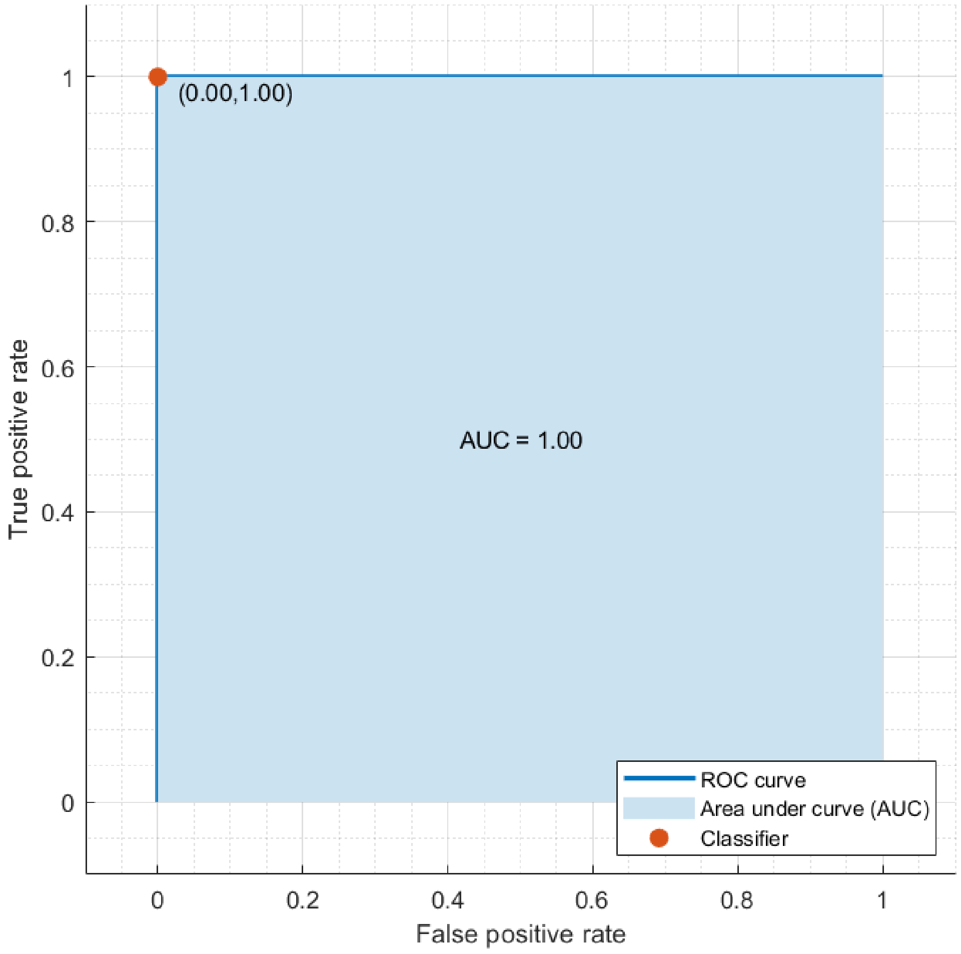


Figure 3.14: An example of Receiver Operating Characteristics (ROC) curve.

classification model. AUC resides in the range of [0, 1], and if the AUC value is equal to 1, it means the predictive performance is perfect (i.e. the classification model correctly assigned all the unseen new cases that it was given during the validation stage). If AUC = 0.5, this refers to classification by chance (random guessing), and if AUC = 0, this refers to an inverted perfect classification. Largest AUC of 1.00 indicates the optimal performance of the trained model, as shown in Figure 3.14. Thus, AUC is utilised to evaluate the predictive performance of the classification models. In this project, the AUC metric is computed with a confidence level of 0.99 to obtain a considerable level of validity and certainty.

### 3.7.2 Stability and Generalisability

Let $S = \{s_1, s_2, s_3, , s_n\}$, be a set of data samples, which can be partitioned into k non-overlapping data subsets of equal size $P = \{p_1, p_2, , ..., p_k\}$. Each subset contains approximately the same proportions of response groups as in the original data. This stratified CV procedure is repeated $k$ iterations. At each iteration $i \in \{1, 2, 3, ..k\}$, the feature mining model is applied on $P \setminus p_i$. Over iterations, a set of subsets of features $FS = \{fs_1, fs_2, ..., fs_k\}$ is produced. When $FS$ is obtained, the consistency of feature selection can be examined to define the most frequently selected features over $k$ iterations. The consistency of selection is more likely to be correlated with the predictive power of features so that the most consistently selected features should be the most relevant, whereas the least consistently selected features should be less relevant. Estimation of the predictive performance of learning models is an essential step, since it guides the process of model selection, and evaluates the quality of the chosen model. As mentioned earlier, for small datasets, there is an increased risk of obtaining high variability in the estimation error. Therefore, the choice of a suitable CV approach for small datasets has been considered carefully in order to report reliable estimations. The $5-$fold CV method is empirically established due to achieving a good compromise when attempting to address the Bias-Variance trade-off.

In this research, the generalisability metric is utilised in the mining stage in order to investigate the capacity of the proposed feature mining models to generalise to wider populations by detecting generic molecular markers for breast cancer from multiple independent genomic datasets that are collected from completely different studies. The generic biomarkers are discovered by examining the selected subsets of stable predictors, which are identified from each breast cancer dataset by the ensemble feature mining model over CV iterations. If the stable predictors are also detected across a wide range of independently generated breast cancer data samples, the more likely they are true biomarkers. As a result, predictivity, stability, and generalisability are considered equally in this research project for the goal of assessing the relevancy, robustness, and reproducibility of the discovered biomarkers across multiple independent datasets so that reliable biological findings can be reported.

## 3.8 Discussion

*"The data generated by the TCGA program comprised a vast resource that investigators will be analysing for years to come. The resource of information about breast cancer genomes will undoubtedly fuel a myriad of discoveries by the cancer research community"* - The director of NHGRI, Eric D. Green. As cancer genomic data has become more accessible, the research presented in this thesis adopts multiple independent cancer datasets, to increase the potential of discovering true biomarkers and decrease the risk of false positive. These genomic and proteomic datasets are the inputs to the proposed feature mining models for the aim of knowledge discovery. Validating and evaluating the identified biomarkers from these HDSSS omics data involves several underlying issues that need to be properly handled using the suitable experimental methodologies, effective evaluation metrics and independent validations.

The internal validation that is based on *stability criterion* and repeated 5-fold CV procedure, is employed to generate variant training sets to examine the consistency of selection of the proposed feature mining models, and variant validation sets to estimate the testing error rate reliably. The external validation that is based *generalisability criterion*, is utilised to examine the potential generalisation of the proposed mining models across multiple independent datasets. Furthermore, the sensitivity of the feature mining models to the variations in the breast cancer samples is investigated further using the integrated datasets. The *predictivity criterion* is applied using two classifiers, which are Support Vector Machine, and Bagging Decision Trees. The response groups of the breast cancer datasets before the integration approaches have considerable disparate sizes, where the majority are the positives (i.e. ER+ and PR+). Therefore, the AUC estimation metric is adopted to assess the performance of these prediction models with a confidence level of 0.99 in order to obtain a considerable level of validity and certainty.

# Chapter 4

# Evolutionary Mining Model

## 4.1 Introduction

The key challenge of the problem of knowledge discovery from omics data is searching through its high dimensional search space. The search space, $S$, which is the total number of possible candidate subsets of genes or proteins to be assessed is equal to $2^d$, and $d$ is the number of variables in the genomic or proteomic data, which is typically thousands or tens of thousands. That means that an extremely huge number of evaluations is required to find the optimal subset of candidate features, which is infeasible or computationally expensive. As discussed in details in Chapter 2, the best possible subset of key genes can be identified using Evolutionary Computational (EC) methods, a group of optimisation algorithms, which retrace the natures path to find a solution to a high dimensional complex problem in as little search time as possible. The EC methods navigate through the search space of possible candidate solutions to identify a feasible solution, which is a subset of predictors in our research problem, with respect to an objective function. As mentioned earlier in Chapter 2, Genetic Algorithm (GA) can be considered as one of the most powerful EC methods applied to feature selection problems [261], thus it is adopted in this research as the search strategy for the feature mining model. The objective function is utilised to estimate the goodness of combinations of genes so that progress toward the best possible subset can be evaluated. Assessing the feasibility of a subset of genes to solve the problem at the hand is a key factor to guide the search process of GA toward a good solution.

Therefore, this chapter introduces the GA, the methodology employed for fitness evaluation based on a hybrid selection approach. The experimental setup of the proposed evolutionary mining model is discussed, along with a detailed explanation of each step of that design. Furthermore, the experimental findings generated from the application of the ensemble evolutionary mining model to the adopted cancer datasets are presented and discussed.

## 4.2 Genetic Algorithm

The Genetic Algorithm (GA) adopts the phenomenon of adaptation as a computational process for solving general-purposes complex problems [134]. The GA is an EC method, starts with a set of initial candidate solutions (individuals), called population. The GA process creates the next population of individuals iteratively by replacing the current (parent) population with the offspring using a kind of natural selection with operators inspired by genetic variations namely *selection, crossover, mutation and elite-preservation*. The *selection operator* selects those individuals in the current population to be parents based on their *fitness values*. Individuals in the parent population that have the highest fitness values are chosen as elite individuals to be passed directly to the next population. During the reproduction process, the GA introduces some variations in the offspring. The *crossover operator* exchanges subparts of two selected individuals in the current population, while mutation operator randomly makes changes to the allele values of some locations in a single individual. The parent population is replaced with the offspring to form the next individuals. Over successive generations, the population evolves toward the best possible solution and the algorithm stops when a stopping criteria is met (e.g. the optimum is found, or a pre-defined number of generations is reached). The choice of a method for each step of the search process can significantly affect the behavior of GA. However, in the literature, there is a large body of research that has shown theoretically and experimentally that there are no universally optimal methods [69].

## 4.3 Experimental Design of the Evolutionary Mining Model

This section explains the design of the evolutionary mining model for the problem of biomarker identification from omics data. Given the genomic or proteomic dataset $D$, which is a $n \times d$ matrix of the training set, where $d$ represents the number of variables, and $n$ is the number of samples. As shown in Figure 4.1, the design starts with passing $D$ to the univariate approach to reduce the dimensionality of the data for the next optimisation phase, by eliminating the least promising genes. The selected genes by the univariate approach are sampled uniformly at random to create the initial population of the GA forming the initial candidate subsets of genes. The quality of each subset is assessed using the multivariate approach. The settings of how new search points of the next population are generated from the members of the parent population are explained in the following subsections. Appropriate choices for each step of the setup is empirically established, considering the fact that the interaction between GA components is conducted in highly nonlinear approaches.

### 4.3.1 Univariate Approach

A univariate approach is utilised firstly as a pre-processing step to reduce the exponential search space of genomic and proteomic data for the next optimisation stage. The evolutionary process of GA together with the multivariate evaluation approach assess the optimality of the candidate subsets of features, which are selected using the univariate approach. Therefore, this preliminary step contributes to decreasing the number of features that will be passed to the fitness function, based on Linear Discriminant Analysis for identifying the best combination of genes. The univariate approach is based on a statistical test that is applied to each gene individually to examine if there is any statistically significant differences between negative (normal) and positive (cancer) patients on the basis of that gene value. Therefore, the two-sample $t-$test [249] assigns a P-value to each gene as a measure of its effectiveness in distinguishing observations of different groups and thus the least discriminative genes are discarded. The gene will be

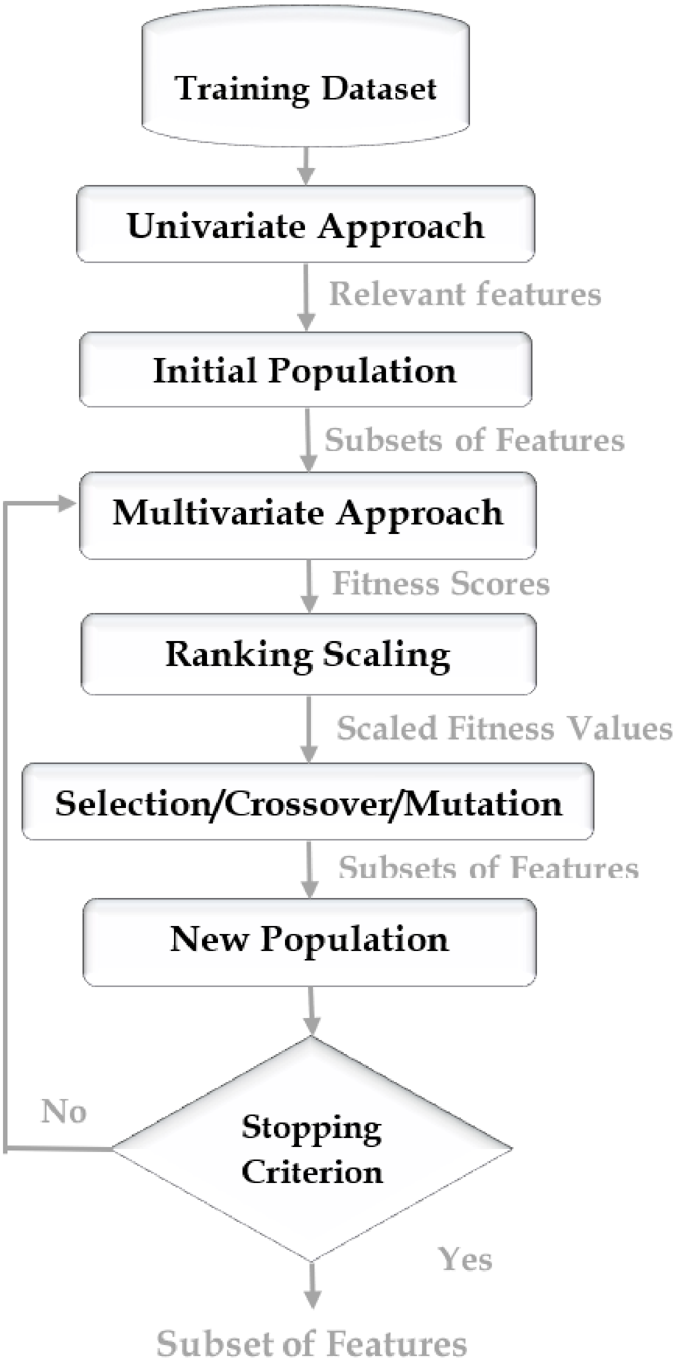


Figure 4.1: Steps of the experimental design of the evolutionary mining model.

considered relevant if the P-value is less than the significant level of the test (i.e. 0.05). Therefore, $t-$statistics check whether these two groups of samples are significantly different or not, as follows:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4.1}$$

where $\bar{x}$ and $\bar{y}$ are the sample means, $s_1$ and $s_2$ are the sample standard deviations, and $n_1$ and $n_2$ are the numbers of samples in the positive and negative groups. Two-sample $t-$test method tests the null hypothesis that the two data vectors are from populations with equal means, without the assumption that the populations also have equal variances. This is also called the Behrens-Fisher problem, which uses Satterthwaites approximation for the effective degrees of freedom. The degree of freedom $v$ for the unequal variance $t-$test is given by [215]:

$$v = \frac{(\frac{1}{n_1} + \frac{u}{n_2})^2}{\frac{1}{n_1^2(n_1-1)} + \frac{u^2}{n_2^2(n_2-1)}} \tag{4.2}$$

where

$$u = \frac{s_2^2}{s_1^2} \tag{4.3}$$

This test is sometimes called Welchs $t-$test. The selected features will form uniformly at random the initial candidate solutions of GA to be assessed using the multivariate approach.

### 4.3.2 Initial Population

The question that needs to be answered is how to represent the individuals of GA's population. A population is an array of individuals, where individuals represent potential solutions to the problem at hand, which are combinations of genes or proteins for the biomarker discovery problem. Thus, for our research problem, an individual is a fixed-length vector of $nFeat$ genes, to which the fitness function can be applied. So, for example, individuals of gene expression dataset is described by $nFeat$ genes: ⟨RPS11, PNMA1, MMP2, ZHX3, ERCC5,..., CTSC⟩. The actual space is the finite set of real numbers representable using floating-point representation. GA creates the initial population, which is a matrix $p \times nFeat$, where $p$ represents the number of individuals that are generated randomly from the selected features of the previous phase. Where $nFeat$ is the number of genes in each candidate solution, which is equivalent to the desired number of features to be detected.

### 4.3.3 Multivariate Approach

A multivariate approach based on Linear Discriminant Analysis (LDA) is adopted to measure the optimality of each subset of genes. In statistics, Discriminant Analysis is a well-known method for capturing the characteristics of the data that can best distinguish the samples in one group from those in another. When the actual groups are known, the interest is to form a rule based on the features that best characterise the differentiation between the disparate groups. According to Fishers rule [96] *"vectors in one class behave differently from vectors in the other classes, and the variance within the classes differs maximally from that between the classes"*. LDA finds linear combinations of variables in a way that the variability within-class is small and between-class is large, in order to discover structure in the dataset that guarantees maximal separability.

Suppose $A$ is a $n \times d$ matrix of the training set, where $d$ represents the number of variables, and $n$ is the number of samples. Each sample is represented by $\mathbf{x} = (x_1, ..., x_d)$. For $K$ response group, the label $Y$ ranges from 1 to $K$. The sample space of training dataset is divided into $K$ disjoint groups $(G_1, ..., G_k)$. The within-class variability can be obtained by calculating the separability (i.e. the distance) between the means of different response groups, which results in $W$ matrix of $d \times d$. While the between-class variability depends on calculating the distance between the mean and the samples of each response group, which produces the $B$ matrix of $d \times d$. LDA finds the linear combination $A\mathbf{a}$ of the variables, so that the proportion of between-class to within-class is given by $a'Ba/a'Wa$.

LDA assumes that the data within a group $k$ follows a multivariate normal distribution with mean $\mu_k$ and covariance $\Sigma_k$. When the class densities have the same covariance matrix, $\Sigma_k = \Sigma$ for all $k$, the discriminant rule is based on the

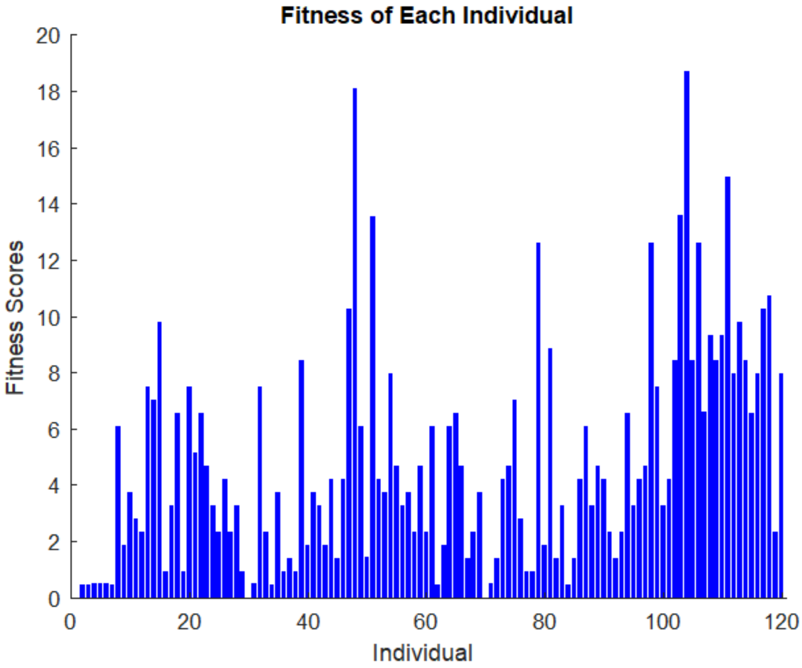


Figure 4.2: An illustration of the fitness score of each individual.

square of the Mahalanobis distance and is linear in $\mathbf{x}$, and given by the following general form [79]:

$$f(\mathbf{x}) = arg \min_{k} (\mathbf{x} - \mu_k)\Sigma^{-1}(\mathbf{x} - \mu_k)' \tag{4.4}$$

The population mean vectors and covariance matrices are estimated from a training set by the sample mean vectors and covariance matrices $\hat{\mu}_k = \bar{x}_k$ and $\hat{\Sigma}_k = S_k$. For the constant covariance matrix case, the pooled estimate of the common covariance matrix is utilised as follows [79]:

$$\hat{\Sigma} = \sum_{k} (n_k - 1)S_k/(n - K). \tag{4.5}$$

The discriminative individual is the combination of genes that can maximise the separation of positive observations from negative ones, such that the classification error rate of the multivariate model is minimised. GA assigns a fitness score for each individual in the current population according to its discriminative power as shown in Figure 4.2. Then, a sorted list of fitness values is created.

### 4.3.4 Ranking Scaling

A scaling function is utilised to convert the fitness scores of the multivariate evaluation approach to scaled values that are more appropriate for the next phase. Firstly, the ranking scaling function ranks each subset of features according to its location in the sorted list of fitness scores, for example, the rank of the best subset is 1, and the next best subset is 2. Then, the ranking scaling function scales the fitness scores of each subset on the basis of its rank, for example, the scaled score of a subset with rank $n$ is proportional to $1/\sqrt{n}$. As a result, the scaled value of the best subset is proportional to 1, and the scaled value of the next best subset is proportional to $1/\sqrt{2}$. The GA algorithm aims to minimise the misclassification rate of LDA, subsets with low scores have high scaled values. The utilisation of the ranking scaling function results in removing the impact of the spread of the fitness scores. Moreover, poorly ranked subsets become more closely equal in value using the square root compared to rank scoring. It is important here to emphasise the impact of the scaling function on the performance of GA. If the scaled fitness values are different, the highest scaled subsets reproduce rapidly, and that could lead to the insufficient exploration of the search space. On the other hand, if there is a little variation in the scaled fitness values, all subsets may be reproduced equally leading to slow convergence. The subsets of the next generation are selected according to their scaled fitness values. Subsets with high scaled values have a higher chance of selection.

### 4.3.5 Selection

The first genetic operator in the GA reproduction process is the *selection*, which specifies how to choose subsets in the parent population to generate offsprings for the next generation. Therefore, the selection operator could act like driving the search process of GA towards interesting parts of the search space by mimicking the concept of the survival of the fittest. The GA lays out a line in a way where each subset is assigned to a segment of the line that is proportional to the scaled fitness score of that subset. The GA moves along the line in steps of equal size and

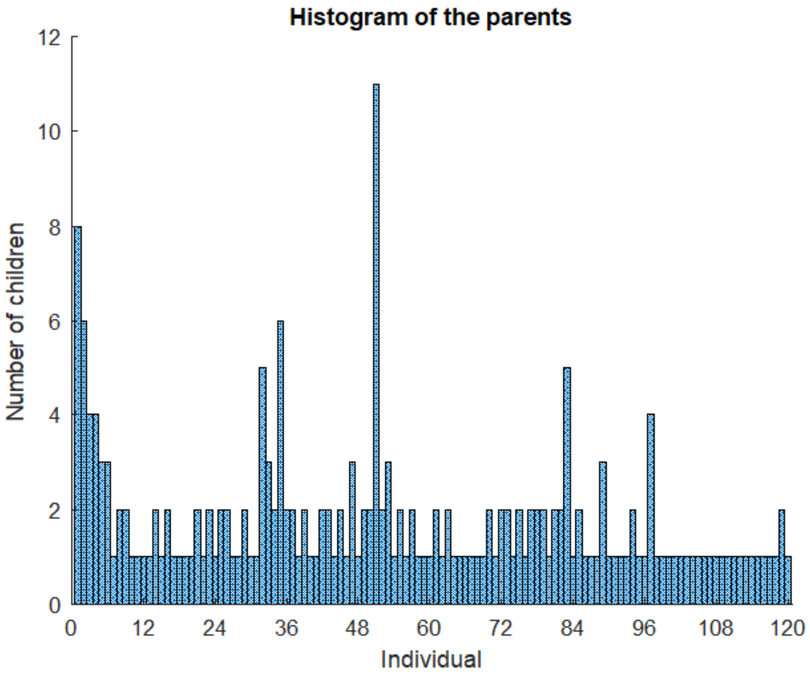


Figure 4.3: Histogram of parent individuals.

selects a parent subset from the segment it lands on. The first step is a uniform random number less than the step size. A subset of features can be chosen more than once to be a parent, and that means the features of the subset contribute to form more than one offspring as shown in Figure 4.3. Some of the subsets in the current population that have the best-scaled fitness values are chosen as elite. These elite subsets are passed directly to the next population. The fraction of subsets in the parent population are guaranteed to survive to the next generation is equal to (i.e. 0.05 multiplied by the size of the population).

### 4.3.6 Reproduction

Beside elite children, GA combines pairs of subsets in the parent population to produce *crossover children* for the next generation. The crossover operator generates a random binary vector and selects the genes from the first parent subset where the binary vector is equal to 1, and the genes from the second parent subset where the binary vector is equal to 0 and combines the genes to form the

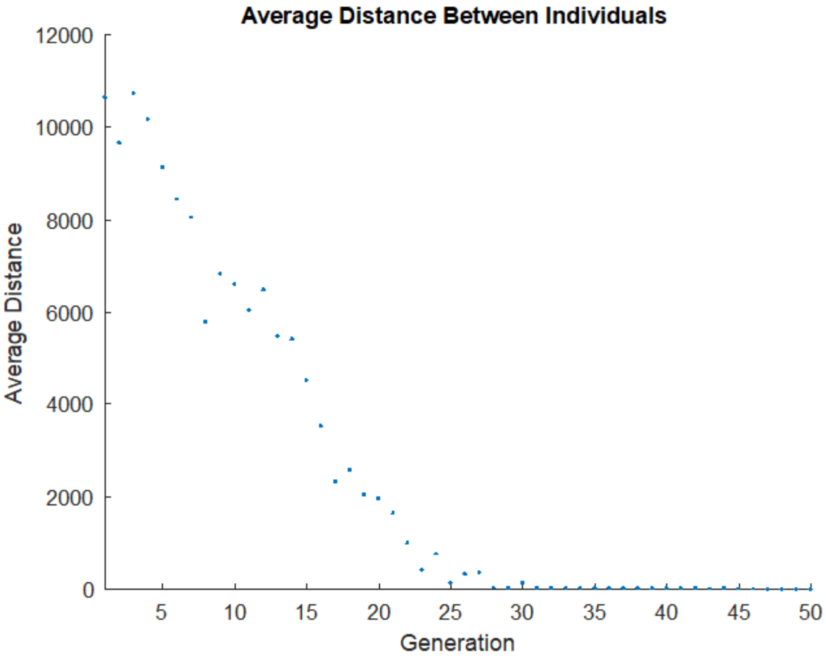


Figure 4.4: An illustration of the average distance between individuals at each generation.

new offspring. For the crossover operator, the amount of variation introduced when generating a new subset may rely on the number of the crossover points, so when the number is increased, adequate amount of variations can be produced. However, adopting a fixed number of crossover points causes that for a subset, genes that are close together are more likely to be inherited as a combination than if these genes are separated. Therefore, selecting the number of crossover points based on stochastic schema leads to produce crossover points anywhere from zero to $nFeat - 1$. The fraction of the next generation that is reproduced by crossover is 0.8.

At *the mutation stage*, GA makes random changes in the subsets in the population to create mutated children. Since the values of genes of GA's individuals are real numbers, a small perturbation of an inherited gene value is the natural way to implement mutation. Therefore, the mutation operator adds a random number to each entry of the parent subset chosen from a Gaussian distribution $G(0, \sigma)$ with a mean of zero and a standard deviation of $\sigma$. For mutation operator,

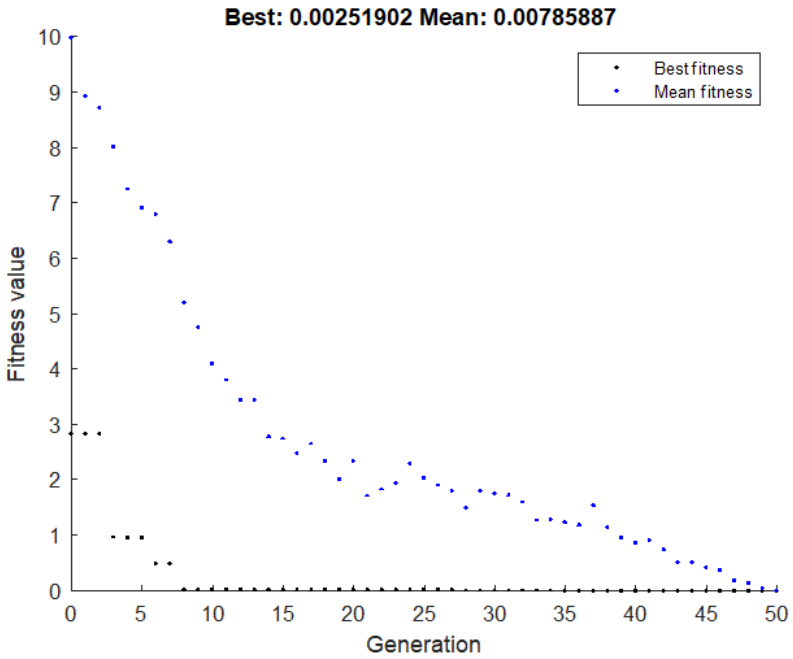


Figure 4.5: An illustration of the algorithm's evolution over generations.

the amount of introduced variation when creating new offspring may rely on how many genes are to be mutated and the amount of change in a genes value. The case when the number of mutated genes is low (e.g. one gene) and the amount of change is increased; this can be effective when the genes contribute independently. However, this may not be effective when there is an interaction between genes because improvements in GA performance require mutating multiple genes simultaneously. Therefore, mutating all genes contributes to a significant improvement in GA's performance.

To avoid any disruption that might result from the perturbation of multiple genes, GA controls the average amount of mutation through generations by decreasing the standard deviation linearly so that the amount of mutation decreases to 0 at the final step. As shown in Figure 4.4, the average distance between individuals at each generation is large, in order to make progress and the diversity declines in the last generations due to the drop in the mutation . Therefore, the mutation operator contributes to the diversity of the population and increases the likelihood that the algorithm will search a broader space and generate individuals with better fitness values. The parent population is replaced with the produced children to constitute the next generation. The algorithm iterates until the average relative change in the best fitness function value over generations is less than or equal to (i.e. 1e-6). Over successive generations, the population evolves toward the best solution as shown in Figure 4.5.

## 4.4 Results and Discussion

This section presents the analysis executed to evaluate the performance of the proposed evolutionary mining model to infer useful knowledge from genomic and proteomic data that can be employed to construct reliable prediction systems using the SVM and BDT learning models. Firstly, stratified 5−fold CV procedure is employed to randomly partition each dataset into training-validation sets as illustrated in Appendix A. At each iteration, the SVM and BDT learning models that are discussed in Chapter 2, were trained using the training set that contains only the discovered biomarkers, and then validated using the corresponding validation set as shown in Appendix B. Over CV iterations, the average predictive performance of the classification models is estimated using the quantitative quality metric, AUC. The experimentally obtained results are presented together with the discussion first, for the ovarian cancer dataset in the following subsection, followed by METABRIC, breast invasive carcinoma datasets with ER and PR groups and the integrated datasets with ER and PR groups.

### 4.4.1 Results and Discussion of Ovarian Cancer Dataset

This dataset (which is discussed in Section 3.6) was utilised in the preliminary experiments conducted in this research to develop the proposed evolutionary mining model and assess its outputs. Initially, the 5−fold CV procedure divided ovarian cancer dataset randomly into training-validation sets, as shown in Appendix A - Table 1. To examine the consistency of selection of the evolutionary mining model over the variations in the data, the CV procedure was iterated 50 times. In other words, the 5−fold CV procedure was re-partitioned randomly 10 times to generate 50 different training sets, which were used by the evolutionary mining

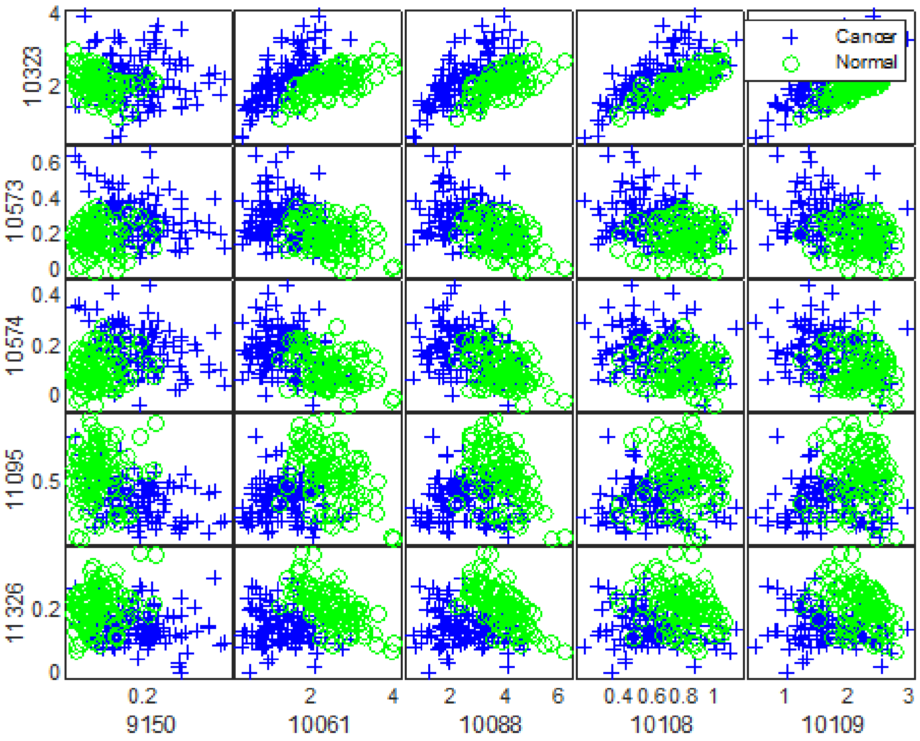


Figure 4.6: Scatter plots matrix of the stable predictors (index) of ovarian cancer dataset.

model to produce 50 subsets of features. Then, the obtained groups of candidate features were compared to define a subset of consistently selected predictors. The outcomes of our experiment identified a subset of 10 stable predictors from the ovarian cancer dataset, as shown in Figure 4.6. The discovered predictors were plotted in the X-axis and Y-axis, ascendingly according to their index, as illustrated in Figure 4.6.

At this point, it is relevant to observe that the intensity values of the stable proteins for cancer patients differ significantly from those in the normal group. Typical biomarkers identification models adopt the idea that the genes or proteins that exhibit the greatest variations across the biological conditions can be considered as potential biomarkers. Therefore, the detected proteins could act as potential biomarkers for ovarian cancer. To examine the predictivity of these proteins, the SVM and BDT classification models were trained using the training set restricted to the discovered subset of those proteins and then validated using the corresponding validation set, as shown in Appendix B - Figure 7, represented by the confusion matrices and the ROC curve plots of the SVM and BDT models for the final iteration. The average AUCs of the SVM and BDT classifiers over 50 iterations are **0.9788** and **0.9528** respectively. The experimental outcomes of ovarian cancer dataset verify that the introduced evolutionary mining model was able to capture meaningful structure in serum (blood)-derived proteomic data that can differentiate the serum of patients with ovarian cancer from that of women without ovarian cancer with a high-level of predictivity and robustness.

### 4.4.2 Results and Discussion of METABRIC Dataset

The ensemble evolutionary mining model was applied to METABRIC breast cancer dataset to extract relevant mRNA markers to the ER and PR status. Initially, the 5−fold CV procedure divided the METABRIC dataset randomly into training-validation sets, as shown in Appendix A - Table 1. Repeated CV procedure was employed to examine the sensitivity of the evolutionary mining model to 50 different training datasets. Therefore, 50 subsets of candidate features were discovered and the comparison of those subsets led to identifying a list of 10 stable predictors from the METABRIC dataset with ER groups, as shown in Figure 4.7, as well as 10 stable predictors from the METABRIC dataset with PR groups, as shown in Figure 4.8. The mRNA markers were plotted in the X-axis and Y-axis alphabetically using their names, as illustrated in Figure 4.7 and Figure 4.8. Both figures illustrate how the expression levels of the discovered genes for ER+/PR+ patients differ significantly from samples with ER/PR-negative, which verifies the potential of these mRNAs to be indicators for breast cancer and ER/PR positivity.

Table 4.1: The performance of the SVM and BDT models built on the stable predictors of METABRIC datasets with ER and PR groups.

| METABRIC | SVM | BDT |
|---|---|---|
| ER | 0.9854 | 0.9897 |
| PR | 0.9854 | 0.9832 |

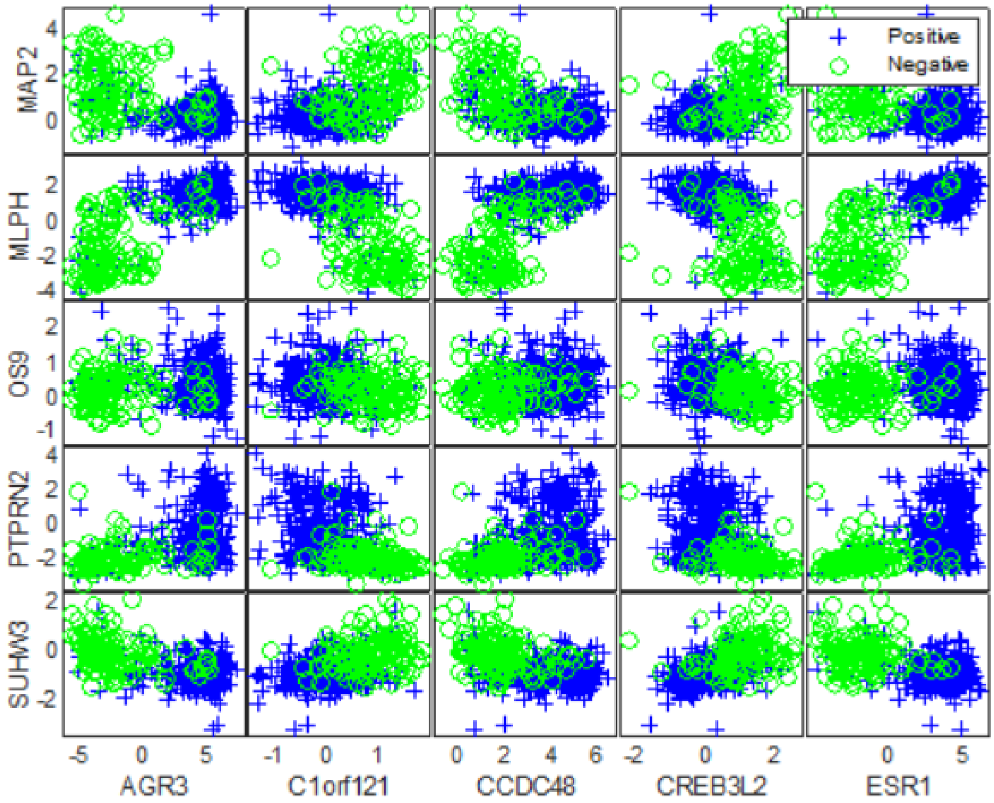


Figure 4.7: Scatter plots matrix of the stable predictors of METABRIC dataset with ER groups.

The predictive power of the identified subsets of robust molecular markers for both response groups was assessed using the SVM and BDT classifiers, as shown in Appendix B - Figure 8 and Figure 9, represented by the confusion matrices and the ROC curve plots of these learning models for the final iteration. The average AUCs are reported in Table 4.1. The obtained results reveal that the discovered genes contributed to constructing highly accurate and robust prediction models for both ER and PR groups as shown in Table 4.1. As a result, the outcomes of our experiments provide strong evidence that supports the capacity of the proposed evolutionary mining model to capture interesting complexity from this HDSSS cancer genomic dataset.

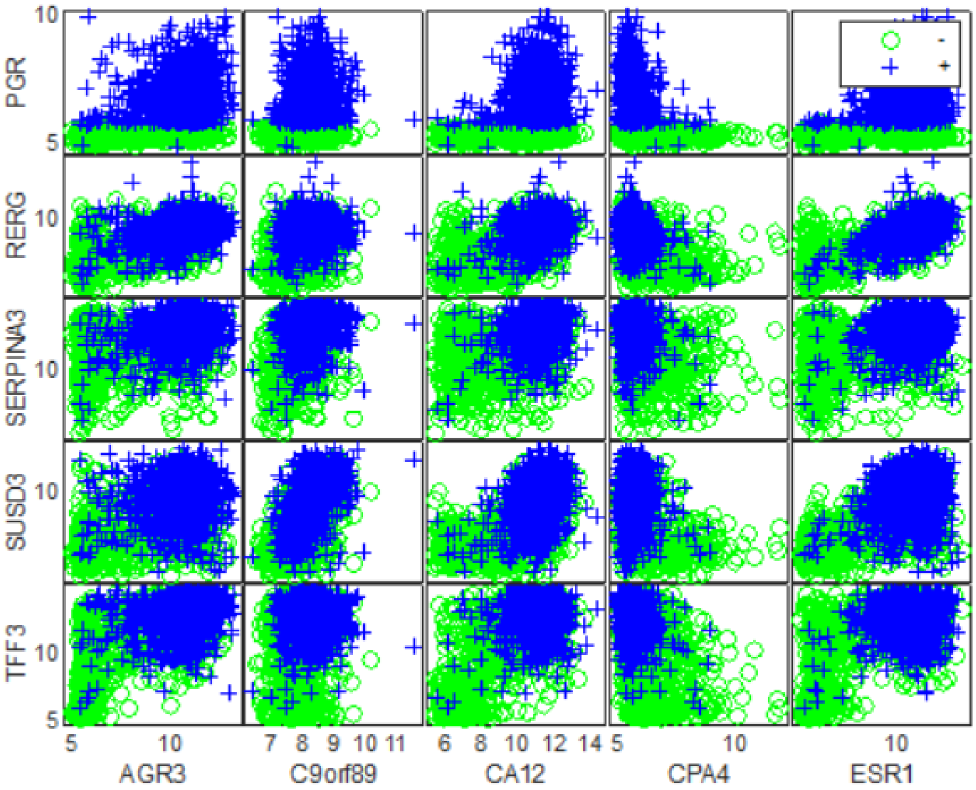


Figure 4.8: Scatter plots matrix of the stable predictors of METABRIC dataset with PR groups.

### 4.4.3 Results and Discussion of Breast Invasive Carcinoma Datasets

In this section, the evolutionary mining model was applied to the breast invasive carcinoma datasets: (Nature 2012), (Cell 2015), and (Provisional), which are explained respectively in Sections 3.3.1, 3.3.2, 3.3.3. The aim is to discover key genes that underlie the biological process of ER and PR. The 5−fold CV procedure was utilised firstly to divide each dataset into 10 random partitions, as shown in Appendix A - Table 3. The evolutionary mining model based on the ensemble approach produced 50 subsets of candidate genes, in order to investigate the gene preferences of the proposed model across different training sets and detect a subset of stable predictors for each dataset. As discussed in Section 2.10, the external validation is necessitated to test the generalisability of the feature mining model to wider populations. Therefore, the discovered groups of consistently selected predictors were compared to find generic mRNA markers across (Nature 2012), (Cell 2015), and (Provisional) datasets. The generic biomarkers will be discussed in the next subsections according to their relevancy to the ER groups and PR groups.

#### 4.4.3.1 ER Groups

Two biomarkers were found to be generic across the breast invasive carcinoma datasets, which are **{'ESR1', 'AGR3'}** as shown in Figure 4.9 - subfigures with ER, which can illustrate the capability of these biomarkers to distinguish the samples with ER+ tumours from those with ER- effectively. The discovered mRNAs were plotted in the X-axis and Y-axis alphabetically, using their names, as illustrated in Figure 4.9. Furthermore, the performance of the SVM and BDT models formed from the training sets that contain only these biomarkers was assessed using the corresponding validation sets, as shown in Appendix B - Figure 10 and Figure 11, represented by the confusion matrices and the ROC curve plots of these learning models for the final iteration. The average AUCs of both classifiers over 50 iterations are presented in Table 4.2. The experimental outcomes show that the SVM and BDT classification models achieved a high level of predictive performance and robustness over all the datasets, which reflects the robustness of the identified genes. This is evidence showing that the presented evolutionary mining model was able to recognise individual markers that are more insensitive to the variations in the data, while simultaneously maintaining as much of the knowledge about the input data as possible.

Table 4.2: The performance of the SVM and BDT models built on the generic biomarkers of the breast invasive carcinoma datasets with ER and PR groups.

| Dataset | SVM-ER | BDT-ER | SVM-PR | BDT-PR |
|---|---|---|---|---|
| (Nature 2012) | 0.9239 | 0.9205 | 0.8555 | 0.8477 |
| (Cell 2015) | 0.9379 | 0.8958 | 0.8599 | 0.8581 |
| (Provisional) | 0.9370 | 0.8954 | 0.8679 | 0.8603 |

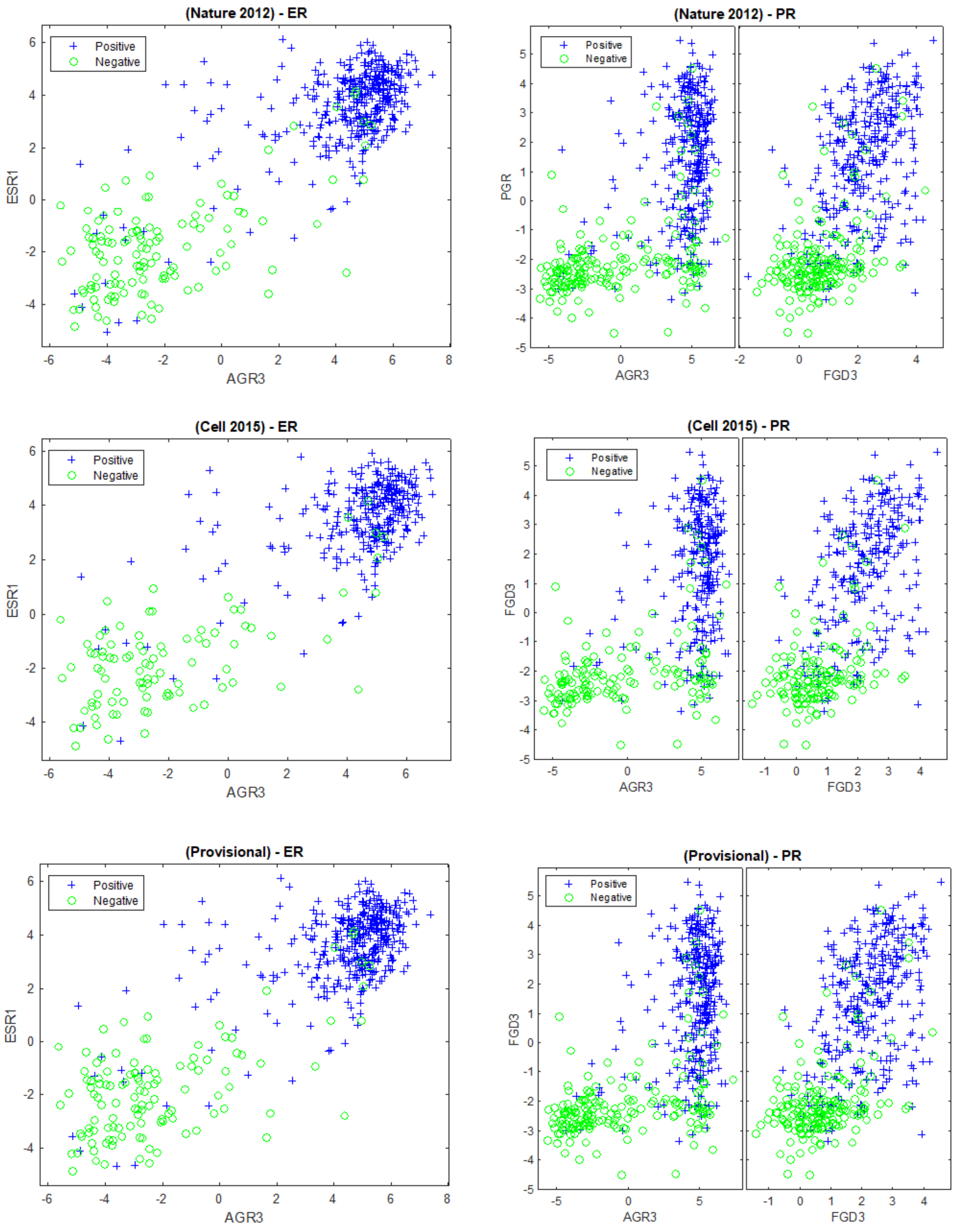


Figure 4.9: Scatter plots matrices of the generic biomarkers of the breast invasive carcinoma datasets with ER and PR groups.

#### 4.4.3.2 PR Groups

Three generic biomarkers were recognised across the breast invasive carcinoma datasets with PR groups, which are **{'PGR', 'AGR3', 'FGD3'}**, as shown in Figure 4.9 - subfigures with PR. The discovered mRNA markers were plotted in the X-axis and Y-axis alphabetically using their names, as illustrated in Figure 4.9. Herein, it is important to observe the capability of the detected biomarkers to differentiate the patients with PR+ tumours from PR-negatives effectively. That means that the expression levels of these mRNAs differ significantly for patients in PR+ group from the samples in PR- group. Therefore, the discovered mRNA markers have the potential to be true biomarkers for breast cancer and PR positivity. The predictive power of the discovered biomarkers was assessed using the SVM and BDT models, as shown in Appendix B - Figure 12 and Figure 13, represented by the confusion matrices and the ROC curve plots of these learning models for the final iteration. The average AUCs over 50 iterations are reported in Table 4.2. The experimental results reveal that the classification models achieved a good level of predictive performance over all the datasets, which reflects the predictivity and robustness of the discovered biomarkers. This is again another pieced evidence that demonstrates the capability of the proposed evolutionary mining model to navigate through the large search space of genomic data and identify an ensemble subset of robust and reproducible biomarkers.

### 4.4.4 Results and Discussion of Integrated Breast Invasive Carcinoma Datasets

The proposed evolutionary mining model was then applied to the integrated breast invasive carcinoma datasets with ER groups: NCP1, NCP2, NCP3, which are illustrated respectively in Sections 3.4.1.1, 3.4.1.2, 3.4.1.3. Furthermore, the evolutionary mining model was applied to the integrated datasets with PR groups: NC, CN, NP, PN, CP, PC, which are explained respectively in Sections 3.4.2.1, 3.4.2.2, 3.4.2.3, 3.4.2.4, 3.4.2.5, 3.4.2.5. Initially, the 5−fold CV procedure was employed to partition each dataset into 10 random splits, as shown in Appendix A - Table 7 with ER groups and Table 8 with PR groups. The internal validation of the evolutionary mining model was verified over 50 iterations to discover the most consistently selected predictors. Then, the external validation was adopted to detect generic biomarkers from the subsets of stable predictors across the integrated datasets for both response groups. The generic mRNAs will be discussed in the following subsections based on their relevancy to the ER and PR groups.

#### 4.4.4.1 ER Groups

Two biomarkers were detected to be generic, across the integrated invasive carcinoma datasets with ER groups, which are **{'ESR1', 'CA12'}** as shown in Figure 4.10. The discovered biomarkers were plotted in the X-axis and Y-axis alphabetically using their names, as illustrated in Figure 4.10. It can be observed from this figure the potential of the discovered biomarkers to separate the ER-positive patients from ER-negatives effectively, which refers to the variability in their expression levels between ER+ and ER- groups. The predictivity of the detected mRNA markers was assessed using the SVM and BDT classifiers, as shown in Appendix B - Figure 14 and Figure 15, represented by the confusion matrices and the ROC curve plots of these learning models for the final iteration. The average AUCs over iterations are introduced in Table 4.3.

The obtained results of both classification models reveal a high-level of predictivity, as well as robustness, which validates the capability of the identified biomarkers to build highly accurate and reliable prediction systems. This, in turn, validates the performance of the evolutionary mining model to extract useful knowledge from these integrated datasets. Furthermore, the outcomes of our experiments show that the BDT classifier achieved a higher level of predictive performance than the SVM model, which verifies the importance of training this model using more substantial data, whose response groups are well-balanced.

Table 4.3: The performance of the SVM and BDT models built on the generic biomarkers of the integrated datasets with ER groups.

| Dataset | SVM | BDT |
|---|---|---|
| NCP1 | 0.9344 | 0.9654 |
| NCP2 | 0.9356 | 0.9635 |
| NCP3 | 0.9374 | 0.9664 |

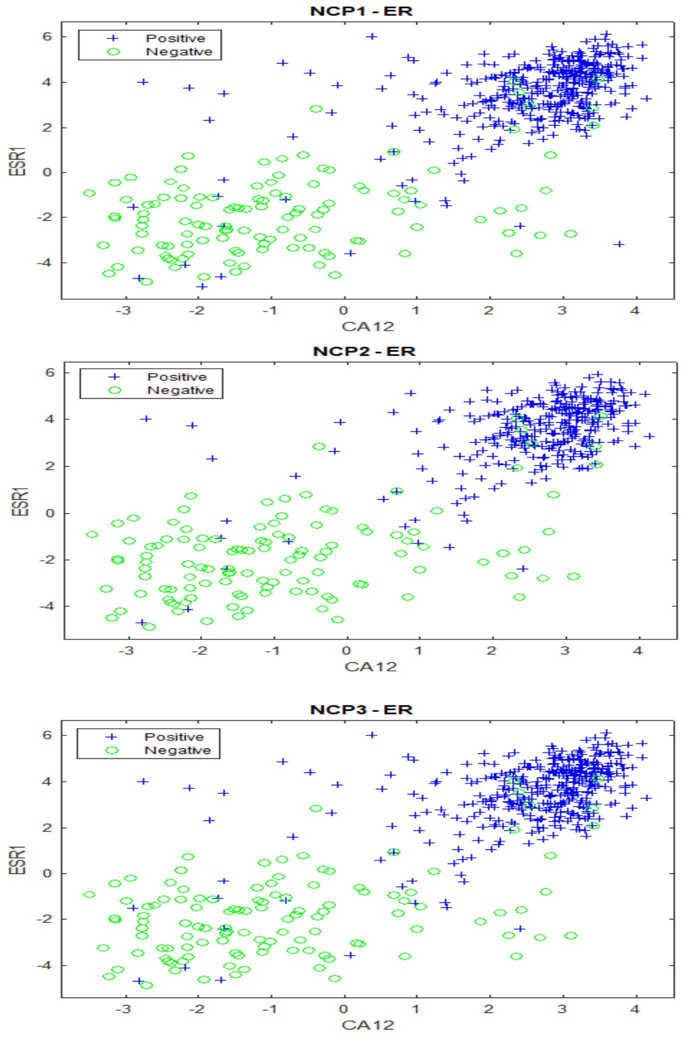


Figure 4.10: Scatter plots matrices of the generic biomarkers of the integrated datasets with ER groups.

#### 4.4.4.2 PR Groups

Four biomarkers were found to be generic between the subsets of stable predictors of the integrated invasive carcinoma datasets with PR groups, which are **{'AGR3', 'FGD3', 'PGR', 'GFRA1'}**, as shown in Figure 4.11. The discovered biomarkers were plotted in the X-axis and Y-axis alphabetically using their names, as shown in Figure 4.11. It is important to observe how the detected biomarkers exhibit a discrimination capability among PR+ and PR- groups, in

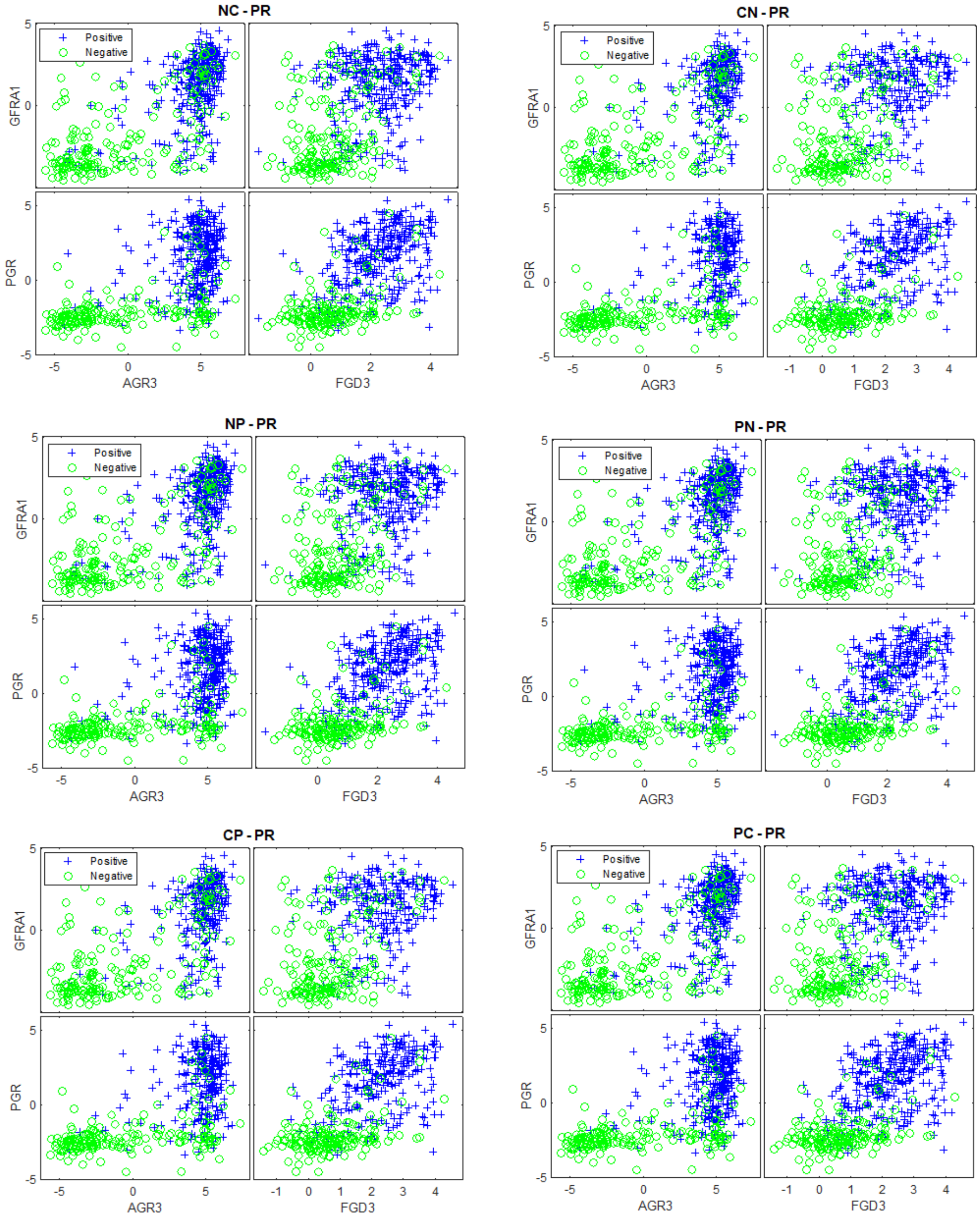


Figure 4.11: Scatter plots matrices of the generic biomarkers of the integrated datasets with PR groups.

which their expression levels show a significant difference across the samples of different groups. To examine the predictive power of these biomarkers, the SVM and BDT learning models were trained and then validated using the training-validation sets that contain only the generic mRNA markers, as shown in Appendix B - Figures 16, 17, 18, and 19, represented by the confusion matrices and the ROC curve plots of these learning models for the final iteration. The average AUCs over 50 iterations are introduced in Table 4.4. The outcomes of our experiments reveal that the prediction models achieved a high-level of generalisability and robustness over all the datasets, which demonstrates the relevance of the discovered biomarkers to the status of PR.

## 4.5 Discussion

This chapter investigated the usefulness of state-of-the-art evolutionary computation methods, with the goal of developing an effective knowledge discovery model from HDSSS omics data. To mitigate the limitations reported in the research literature, the parallel adaptive search of the GA is integrated with the hybrid evaluation measurement, based on univariate and multivariate statistical techniques. Furthermore, the ensemble mining model is employed to provide additional randomness to the selection process, based on GA and produce an ensemble subset of robust biomarkers. The experimental outcomes generated from the application of the evolutionary mining model to the ovarian, METABRIC, and breast invasive carcinoma datasets individually and collectively for both ER and PR groups are

Table 4.4: The performance of the SVM and BDT models built on the generic biomarkers of the integrated datasets with PR groups.

| Dataset | SVM | BDT |
|---|---|---|
| NC | 0.8733 | 0.9286 |
| CN | 0.8712 | 0.9269 |
| NP | 0.8807 | 0.9301 |
| PN | 0.8793 | 0.9284 |
| CP | 0.8760 | 0.9289 |
| PC | 0.8807 | 0.9319 |

presented and discussed in this chapter. The results of our experiments reveal the capability of the proposed model to detect individual indicators that are robust to irrelevant variabilities in the input, while simultaneously capturing the required information to recover the data. Furthermore, these generic molecular markers exhibit a high-level of predictivity, in which their expression levels show significant differences between the patients with ER+/PR+ tumours and samples with ER/PR-negative. A high-level of robustness these biomarkers also exhibit over multiple independent datasets strongly indicate their pervasiveness amongst a broad range of breast cancer patients.

Leveraging the merits of traditional statistical techniques with evolutionary computational methods for the purpose of knowledge discovery from HDSSS omics data contributed to developing an efficient feature mining model. The only concern about this feature mining model is that the number of the detected generic biomarkers across the datasets is small. This has driven our research to explore state-of-the-art deep neural network models, for inferring high-level abstract features from HDSSS omics data. More specifically, to ascertain whether the automatically extracted hierarchical features produced by such neural network models will offer a distinct advantage over our current evolutionary mining model for extracting salient biomarkers from omics data. Therefore, in the next chapter, we will discuss our innovative Deep Feature Mining Model.

# Chapter 5

# Deep Mining Model

## 5.1 Introduction

Learning useful knowledge from high dimensional data automatically, without the need for hand designed features that require domain expertise or ad-hoc specific methodologies and techniques, is highly desirable. This kind of automated learning has the potential to identify high-level abstract representations that aid predictions relevant to precision medicine. **The question is**: what are the required elements of a feature learning algorithm to be able to exploit large and noisy spaces of omics data effectively and discover robust biomarkers? Given the fact that omics data are more likely to be non-linear in nature [297], there is a necessity for *nonlinear learning* that avoids the linear assumptions of traditional statistical models, in order to discover enough of the meaningful intricacies underlying these high-throughput data.

In the literature, it has been shown that the shallow architectures of learning algorithms could lead to a poor generalisation ability, unless a huge number of samples and resources are provided [21], therefore, there is a significant requirement for feature learning based on *deep architectures.* Shallow architectures are more likely to capture low-level features of the input, encoding more noise, and lacking the variance in training data to constrain the weights and thus representations. With deep architectures, the dimensionality can be substantially reduced, thus the problem can be further abstracted by learning high-level features from low-level representations, allowing better generalisation performance and knowledge transfer [24, 38, 176]. This necessitates the need for deep feature learning models that consist of multiple levels of input transformation of increasing abstractions, in order to mitigate against the curse of dimensionality of omics data.

In addition to the curse of dimensionality, biomarker discovery from omics data has the additional problem of small sample sizes such that the number of variables vastly exceeds the number of observations. Research has shown that for a small training set, the unsupervised pre-training approach that is discussed in Chapter 2 produces consistently better generalisation performance and prevents the risk of overfitting [87]. However, as discussed previously, the dimensionality of omics data is high (i.e. tens of thousands of molecules), and that means that there is an exponential number of possible input configurations. Therefore, the available biological samples become even increasingly sparse making the process of discovering plausible and robust input configurations a very difficult task. Moreover, in genomic datasets, very few genes are expressed reliably at biologically significant levels and distinguishably from noise and measurement variation [32]. Consequently, a new feature learning model is introduced based on a set of non-linear sparse Auto-Encoders that are deliberately constructed in an under-complete manner to force the network to find progressively the complex featural representations necessary to capture enough of the important variations underlying the biological samples. The proposed deep feature learning model is utilised to discover and interpret important signals from omics data that aid prediction relevant to precision medicine.

The proposed deep feature learning model applies multiple levels of projections to the input features to abstract the problem and capture high-level dependencies for achieving a high-level of generalisability. This would be a powerful feature learning model for high dimensional classification problems. However, for the problem of knowledge discovery, it is hard to interpret which subsets of genes were responsible for deriving such predictions. To overcome the inherent issue of poor explanatory power associated with the deep learning paradigm, a new weight interpretation method will be presented that aids the researcher in opening up the so-called black box of the network to ascertain which genes were dominant within its internal representations. An interpretation method that can provide explainability to the black-box problem is crucial to approach AI, so that a new horizon of knowledge in a wide range of domains can be discovered. The novel interpretation technique introduced will aid bioinformatics researchers to open the black box and thereby discover important biomarkers from the latent representations form such DL models.

Some existing deep learning methods are able to handle curse of dimensionality issues and improve generalisability. However, this is typically at the expense of *long training times, a need for substantial data to train the models, and lack of transparency in that it is not able to unambiguously state which input features are responsible for its behaviour.* To alleviate these limitations, a novel deep feature mining model is introduced in this thesis with an explanatory technique that can be used for discovering robust molecular markers from HDSSS omics datasets. Unlike other models, our deep mining model can perform deep classification whilst simultaneously revealing the key factors underlying its hidden representations. The output decisions of the proposed model were further validated using appropriate evaluation metrics and independent model validations, thus providing significant confidence as to the relevance, robustness, and reproducibility of the discovered biomarkers.

## 5.2 Experimental Design of the Deep Mining Model

A new deep feature learning model called a Stacked Sparse Compressed Auto-Encoder is proposed in this thesis to infer useful knowledge from HDSSS omics data for modelling reliable prediction systems. The Sparse Compressed Auto-Encoder (SCAE) is simply a feedforward neural network trained with a variant of backpropagation to reproduce its input signal on its output layer, resulting in a hidden or latent feature layer of neurons representing the underlying transformation performed. The principle idea behind our SCAE model is to transform the original high dimensional omics data into a reduced feature space so that enough of the interesting complexity can be retained whilst not requiring additional observations to further constrain the model. This reduced description of the omics data is further realised through a regularisation technique within SCAE that maximises the likelihood of retaining important input signals describing much of the variance within the data, whilst filtering out the less important and noisy signals.

The Stacked Sparse Compressed Auto-Encoder (SSCAE) is composed of a sequence of SCAE trained in a dependent and co-operative manner, where the hidden feature layer of one model feeds as input to another. The underlying complexity of omics data is compactly represented with multiple levels of abstraction, therefore, we apply a greedy recursive approach to transforming the input signals containing tens of thousands of genes into a hidden representation of a lower dimension and higher abstraction, which is then provided as input to another SCAE, which encodes this further at a higher abstract level and so on. The resulting abstract hidden layer is then provided as input to the final layer of SSCAE (i.e. the output layer), which is a softmax classification layer trained to classify the input as belonging to either a patient with or without cancer.

In addition, we augmented a novel weight interpretation feature into SSCAE such that we were able to determine which original features on the input layer were most highly predictive, positively and negatively associated with the positive patient groups e.g. cancer, ER+/PR+. Therefore, two types of outcomes were revealed by our deep mining model, both indicating strong likelihoods of a patient having cancer. The first outcome indicated a subset of highly positively-weighted genes whereby the amplifications and gains in the gene expression levels were associated with the likelihood of a patient having cancer. Conversely, the second outcome revealed another subset of genes that were highly negatively-weighted and coincided with significant downregulation in the gene expression levels, and again indicated the strong likelihood of a patient having cancer.

### 5.2.1 Auto-Encoder

An autoencoder (AE) is a neural network model that is trained to map an input $\mathbf{x}$ into a hidden representation $\mathbf{y}$ using an encoding function $f$, where $g$ is a decoding function that transforms $\mathbf{y}$ to construct $\mathbf{z}$ as closely as possible to $\mathbf{x}$.

The encoder is a non-linear sigmoid function $s$ that transforms the input vector $\mathbf{x}$ into the hidden representation $\mathbf{y}$, which is expressed as $f_\theta(\mathbf{x}) = s(\mathbf{W}\mathbf{x} + \mathbf{b})$ with parameters $\theta = \{W, b\}$. The weight matrix $\mathbf{W}$ is $d' \times d$, where $d$ corresponds to the dimension of $\mathbf{x}$ and $d'$ corresponds to the dimension of $\mathbf{y}$, and $\mathbf{b}$ is an offset vector of dimensionality $d'$. The decoder is a non-linear sigmoid function $s$ that transforms back the hidden representation $\mathbf{y}$ to construct the vector $\mathbf{z}$ of dimensional $d$, which is expressed as $\mathbf{z} = g_\theta(\mathbf{y})$, where $g_\theta(\mathbf{y}) = s(\mathbf{W}'\mathbf{y} + \mathbf{b}')$ with the parameters $\theta' = \{W', b'\}$. The learning process relies on finding the parameters $\theta$ that significantly minimise the cost function, which measures the discrepancy between the original data $\mathbf{x}$ and its reconstruction $\mathbf{z}$.

### 5.2.2 Sparse Compressed Auto-Encoder

A Sparse Compressed Auto-Encoder (SCAE) is an AE that adds sparsity penalty to the compressed representations to react to distinctive generic features from HDSSS data. Sparsity refers to render the units of hidden layers to be at or near zero so that most factors become irrelevant and few are relevant and insensitive to irrelevant variations. Under-complete or compressed representations corresponds to that the code dimensions (i.e code refers to the hidden layer with the lowest number of dimensions that captures the most abstract features encoded) tend to be smaller than input dimensions. For the SCAE, $\mathbf{z}$ is not supposed to be an exact reconstruction of $\mathbf{x}$, but rather it is meant to be a rough approximation (within an allowable error tolerance) that is less sensitive to variations from the training data leading to avoid the risk of overfitting where very low bias and high variance might be obtained. Moreover, generating a rough approximation will force the network to learn some kind of meaningful relationships between variables. Furthermore, placing constraints on the compressed AE leads to activate hidden neurons in response to given input contributing to distilling effectively enough of the interesting complexity underlying the representative samples that can approximate the input distribution.

Let $\hat{\rho}_i = \frac{1}{n}\sum_{j=1}^{n} a_i x_j$ be the activation of hidden neuron $i$ over a collection of training examples. Neuron $i$ is considered active if the average activation value over all the training examples is close to 1, or inactive if the average value over all the training examples is close to 0. Enforcing the constraint $\hat{\rho}_i = \rho$, where $\rho$ is the sparsity parameter, which takes a small values close to zero (e.g. $\rho = 0.05$). As explained previously, a low activation value means that the hidden neuron reacts to a small number of the training examples, which means different groups of hidden neurons assigned to different statistical features. These patterns of activation can be statistically more efficient since a large number of possible sets of features can be activated in response to given input. Therefore, a regulariser is added to the cost function to enforce the values of $\hat{\rho}_i$ to be low as follows:

$$\Omega_{sparsity} = \sum_{i=1}^{d'} \rho \log(\frac{\rho}{\hat{\rho}_i}) + (1-\rho)\log(\frac{1-\rho}{1-\hat{\rho}_i}). \tag{5.1}$$

In order to reduce the magnitude of the weights and avoid the risk of overfitting so that the learned representations rely on the input features rather than the deep network structure, $L2$ regularisation term on the weights is added to the cost function as follows:

$$\Omega_{weights} = \frac{1}{2}\sum_{l}^{L}\sum_{j}^{n}\sum_{i}^{k}(W_{ji}^{l})^2, \tag{5.2}$$

where $L$ is the number of hidden layers, $n$ is the number of examples, and $k$ is the number of variables. The loss function of training the SCAE is sparse mean squared error (MSE) function, which is formulated as follows:

$$MSE = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}(x_{kn} - z_{kn})^2 + \lambda \times \quad \Omega_{weights} + \quad \beta \times \quad \Omega_{sparsity}, \tag{5.3}$$

where $\lambda$ controls the impact of the weight regulariser in the cost function, and $\beta$ controls the impact of the sparsity regulariser in the cost function. When handling a high dimensional problem, deep network models involve adjustment of thousands of weights, thus the optimisation techniques should be applicable to these large-scaled problems. Several research studies [98, 109, 128, 238] have shown the feasibility of the scale conjugate gradient descent method to deal with high dimensional problems in an effective way. Therefore, the SCAE is trained with scaled conjugate gradient backpropagation method [212].

### 5.2.3 Stacked Sparse Compressed Auto-Encoder

The Stacked Sparse Compressed Auto-encoder (SSCAE) can be developed using a series of SCAEs. The encoding procedure of the SSCAE that has $l$ layers can be expressed as follows: $\mathbf{y} = f_l(...f_i(...f_1(\mathbf{x})))$, where $f_i$ is the encoding function of the module $i$, while the decoding procedure can be defined as: $\mathbf{z} = g_l(...g_i(...g_1(\mathbf{y})))$, where $g_i$ is the decoding function of the level $i$. A series of CV experiments are conducted to assess the performance of the selected modules and identify the best performing one based on validation performance. Therefore, the SSCAE is designed with four layers of dimensions 500, 200, 100, and 50. Then, the output of the fourth layer is employed to train the softmax layer for classification by forcing the output layer of the SSCAE to sum to 1, so that it is forcing backpropagation to be aware of the whole output layer hence this activation function transforms a vector rather than a scalar (net input) like a sigmoid function. The SoftMax neural network layer was trained in a supervised fashion based on the Cross-Entropy (CE) function:

$$CE = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{k} t_{ij}\ln y_{ij} + (1 - t_{ij})\ln(1 - y_{ij}). \tag{5.4}$$

where $n$ is the total number of the training examples, and $k$ is the number of the response groups, $t_{ij}$ is the $ij$entry of the group matrix, which is $k \times n$ matrix, and $y_{ij}$ is the $i$th output from the SCAE when the input vector is $x_j$. The CE function of the SoftMax layer is optimised using the scaled conjugate gradient method [212]. The response group was represented in the output layer coded as 0 for Normal and 1 for Cancer for ovarian cancer dataset. For the METABRIC dataset with Estrogen Receptor, the response groups were encoded in the output layer as 0 for Negative Estrogen Receptor (ER-) and 1 for Positive Estrogen Receptor (ER+). For the METABRIC dataset with Progesterone Receptor, the response groups were encoded in the output layer as 0 for Negative Progesterone Receptor (PR-) and 1 for Positive Progesterone Receptor (PR+). The SSCAE is trained in a supervised fashion based on the CE function of Equation 5.4 and

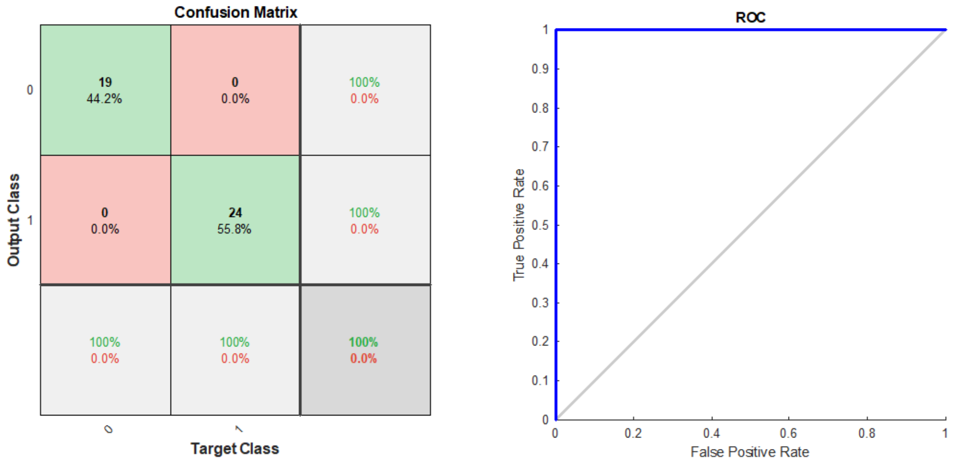


Figure 5.1: An illustration of the validation performance of the SSCAE.

the SCG optimisation method [212], using the full training set and then it is validated using the full corresponding validation set, as shown with performance, represented by the confusion matrix and the ROC curve plots of Figure 5.1 - (ovarian cancer dataset at fold 1). To account for variance in the performance estimation, the SSCAE is trained using variant sets of training samples and the average predictive performance is reported. Furthermore, the performance of each trained SCAE module is validated using the MSE, between the validation set and its reconstruction, which is predicted by the SCAE that was trained on the corresponding training set as shown in Appendix A. The capability to form deep feature hierarchies by stacking the unsupervised modules with the SoftMax classifier results in forming highly non-linear representations that preserve the key determinants within the original data. The high-level representations capture high-level dependencies between features, and this leads to discovering the underlying abstractions needed for solving this complex detection problem.

However, due to the multiple levels of transformations that the SSCAE performs to the input features, it is hard to recognise which subsets of genes or proteins constituted the latent representations of the SSCAE and were responsible for playing a significant role in deriving such predictions. This may not be an issue for some domains because it can easily validate the obtained results. However, providing explainability to disease diagnosis and prognosis systems is a crucial factor to develop a reliable prediction system that can be employed in clinical practice. Furthermore, stating which phenotypes are responsible for such predictions increases the certainty in the decision-making process. The difficulty of deconstructing DL methods remains a major obstacle for employing these advance techniques in omics data analysis for the goal of biomarker identification. In this research project, a new interpretation method called deep mining is introduced to decode the mechanism of the SSCAE so that a reduced set of highly predictive and reliable biomarkers can be derived effectively.

### 5.2.4 A New Weight Interpretation Method

Several hypotheses, that have been proposed in the literature to justify why the unsupervised pre-training approach works well, highlighted the importance of finding the appropriate weights in guiding the learning process towards discovering a good representation similar to the optimisation [177] and regularisation [87] hypotheses. The learning process of DL models can be described as fitting weight parameters in a way that can significantly minimise the loss function. For a shallow AE, the weight of each variable reflects its contribution on the node's activity so that the signal with a larger weight has a greater impact. However, given the deep architecture of the SSCAE, how can we measure the contribution of each feature?

When the SSCAE model is trained using the training set, the classification errors can be back-propagated through the layers of the SSCAE to the input layer to estimate the individual contribution of each variable. That mean that the impact of each variable on the classification accuracy is forward-propagated from the input layer through the layers of the Deep network. Since the weight is the main indicator of variables importance, the relevancy of each feature can be detected through leveraging the Input Weight matrix of the SSCAE, with its Layers Weight matrices. The Input Weight matrix (**IW**) of the SSCAE is $d' \times d$, where $d$ corresponds to the dimension of $\mathbf{x}$ and $d'$ corresponds to the dimension of $\mathbf{y^{(1)}}$. The Layer Weight matrix ($\mathbf{LW_i}$) of layer $l^{(i)}$ of the SSCAE is $d' \times d$, where $d$ corresponds to the dimension of $\mathbf{y^{(i-1)}}$ and $d'$ corresponds to the dimension of $\mathbf{y^{(i)}}$, and for $L$ layers of the SSCAE. Therefore, leveraging the (**IW**) of the SSCAE

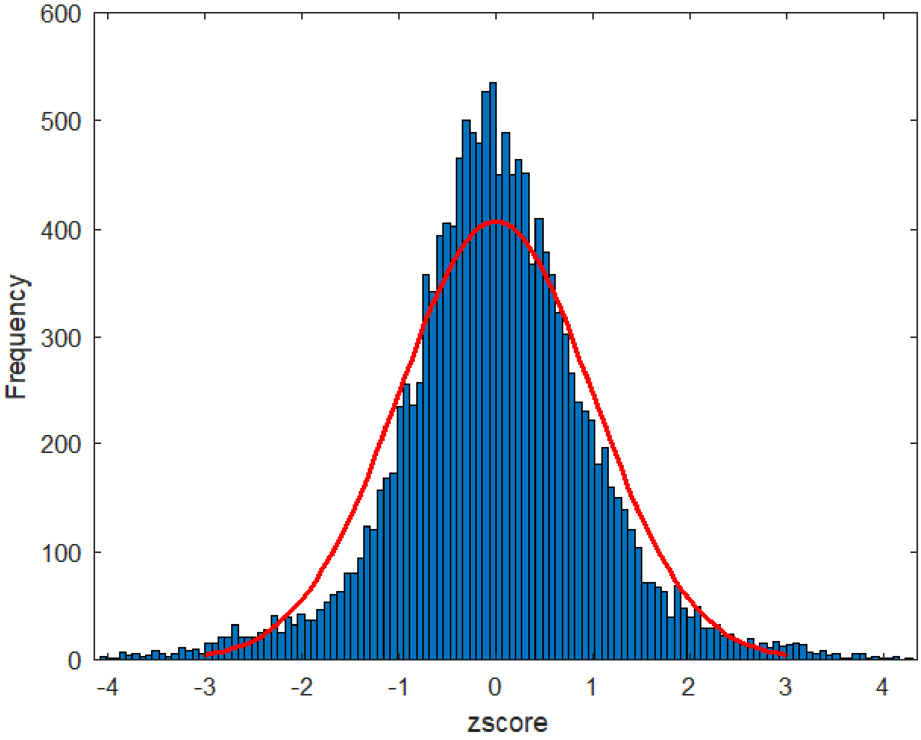


Figure 5.2: Histogram of z-scores of the weight vector.

with its Layers Weight (**LWs**) matrices results in defining the importance of each variable and as follows:

$$\mathbf{DM} = \mathbf{IW}^{\top} \prod_{i=1}^{L} \mathbf{LW}_i^{\top}. \tag{5.5}$$

which results in a $d \times 1$ weight vector called **DM**, where $d$ corresponds to the number of features in the original datasets. Therefore, each gene has a weight score that indicates its integrated impact over the depth of the SSCAE and reflects its contribution. The weights of the features in **DM** are distributed symmetrically and roughly center at 0, and the weight vector **DM** resembles a normal distribution, as shown in Figure 5.2 - (ovarian cancer dataset at fold 1). A small percentage of features in the **DM** exhibit High Positive (HP) or High Negative (HN) weight, as shown in Figure 5.2. Two lists of genes with a length of the code dimension (i.e. 50): 1) with HP weight and 2) with HN weight are detected.

To examine the consistency of feature extraction of the SSCAE together with the deep mining model across the variations in the training data, $k$ weight vectors **DMs** are obtained over CV iterations, thus $k$ lists of genes with HP weight and $k$ lists of genes with HN weight are generated. The positive lists are compared to find the most frequently selected predictors and the negative lists are examined to declare the most consistently detected predictors. As mentioned previously in Section 2.10, the generalisability criterion can be considered the highest validation tool for verifying the outcomes of the feature mining models proposed for biomarker identification from omics data. Therefore, the discovered subsets of consistently selected predictors with HP weight are examined to detect generic molecular markers across multiple independent datasets. By the same way, the identified subsets of stable predictors with HN weight are investigated to detect a generic subset of biomarkers over a wide range of independently generated data samples.

*This weight interpretation method expands our deep learning model to include a feature selection method in addition to the feature extraction capacity already inherent within this paradigm.* As a result, two smaller subsets of robust molecular markers are produced, one corresponding to those genes that are highly expressed for most of the patients from the positive group compared to the negatives; and the other subset refers to those genes that are highly expressed for most of the samples in the negative group compared to the positives. Our novel deep mining model provides yet another arrow within the quiver of bioinformaticians for discovering and evaluating new biomarkers that may help further the endeavour of producing more effective and personalised medicine.

## 5.3 Results and Discussion

This section presents the analysis performed to evaluate the empirical performance of the proposed SSCAE and deep mining model. Firstly, the stratified $5-$fold CV procedure was used to randomly partition each dataset into 10 training-validation random splits, as illustrated in Appendix A. At each iteration, the SSCAE was applied to the training dataset, in order to learn compact and meaningful representations from its high dimensional space for developing a robust DL prediction model. The performance of the SSCAE was validated using the corresponding validation set, as illustrated in Appendix B, and the average AUC is reported over iterations. Each trained SCAE module was also assessed using the MSE, between the validation set and its reconstruction, which was predicted by the SCAE that was trained using the corresponding training set as shown in Appendix A.

Simultaneously, the proposed deep mining model was applied at each iteration to define two lists of candidate features with HP and HN weight. Over CV iterations, the five identified groups of features with HP weight were compared to provide a subset of stable predictors for each dataset and by the same way, a subset of stable predictors with HN weight was produced. The subsets of stable predictors with HP weight were examined and the subsets of stable predictors with HN weight were compared to define generic molecular markers across a wide range of independently generated data samples. The discovered subsets of the generic biomarkers with HP and HN weight were used to build prediction models individually and collectively using the SVM and BDT classifiers in order to evaluate their relevance to the clinical outcomes.

### 5.3.1 Results and Discussion of Ovarian Cancer Dataset

This section presents and discusses the experimentally obtained results of applying the proposed SSCAE and deep mining model to the ovarian cancer dataset, (which is discussed in Section 3.6). Initially, the 5−fold CV procedure was utilised to divide the dataset randomly into 10 random partitions, as shown in Appendix A - Table 1. At each iteration, the SSCAE was trained using the training observations and validated using the corresponding validation set, as shown in Appendix B - Figure 1, represented by the confusion matrix and the ROC curve plot of the SSCAE for the final iteration. The average AUC of the SSCAE over CV iterations is **0.9843**. The high performance of the SSCAE reveals that this deep feature learning model was able to discover a relevant and robust representation from the ovarian cancer data, thus a highly accurate and reliable prediction model was formed. Furthermore, the performance of each trained SCAE module was also evaluated using the MSE between the validation set and its reconstruction

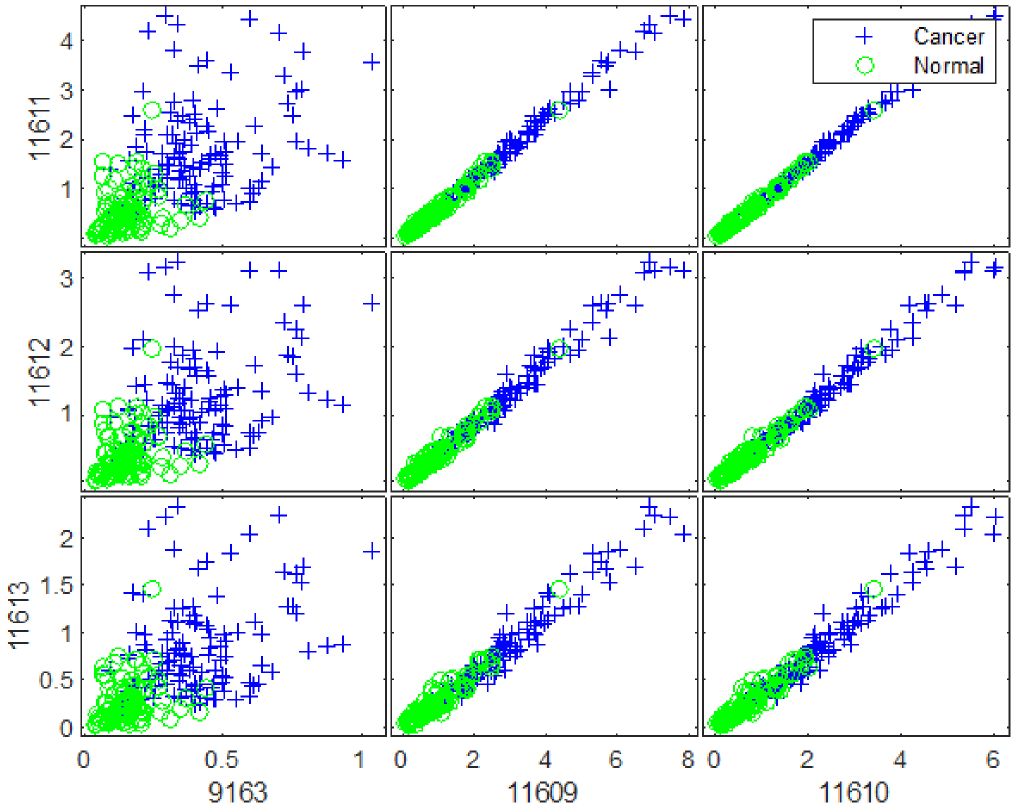


Figure 5.3: Scatter plots matrix of the stable predictors (index) with HP weight of ovarian cancer dataset.

as shown in Appendix A - Table 2.

Simultaneously, the deep mining model was executed at each iteration to decipher which combination of key features constituted the latent representations of the SSCAE. Examining the ten obtained lists of proteins resulted in finding 6 stable predictors with HP weight, as shown in Figure 5.3 and 13 stable predictors with HN weight, as shown in Figure 5.4. The biomarkers were plotted in the X-axis and Y-axis ascendingly, using their index, as illustrated in Figures 5.3, with HP weight and 5.4 with HN weight. At this point, it is relevant to observe that the intensity values of the stable proteins with HP weight for the patients who suffer from cancer are more likely to be higher than their intensity values for most of the normal samples, contrary to the intensity distributions of the

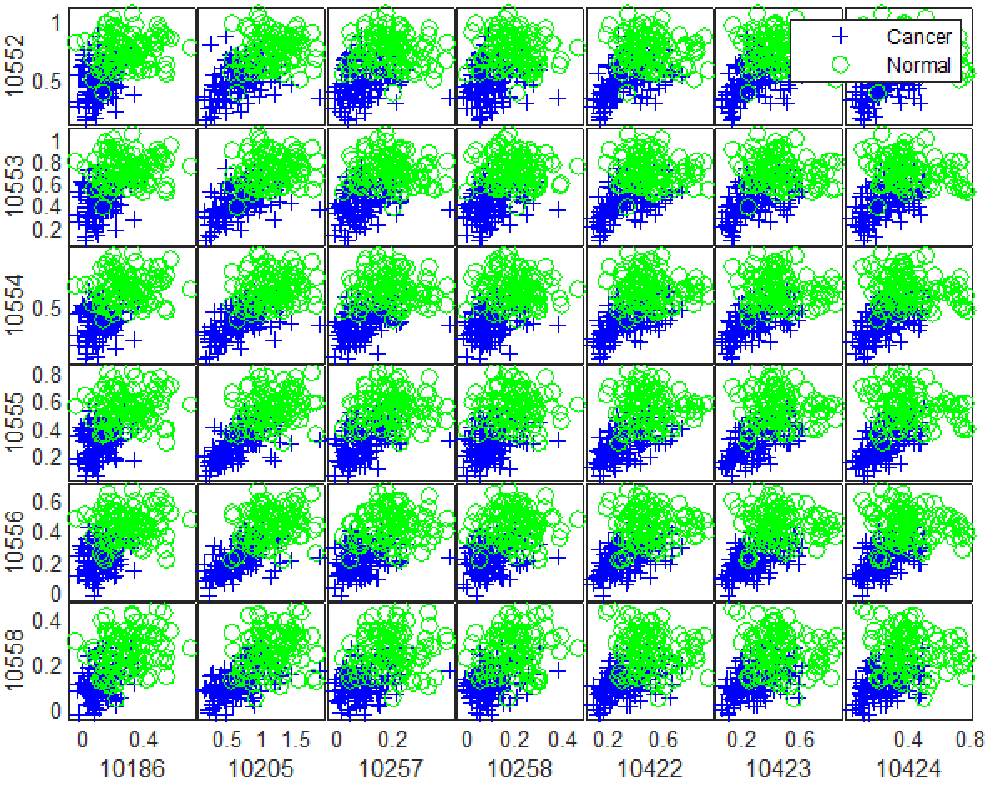


Figure 5.4: Scatter plots matrix of the stable predictors (index) with HN weight of ovarian cancer dataset.

stable proteins with HN weight, where their values for the normal observations are more likely to be higher than their intensity values for most of the ovarian cancers. Firstly, this demonstrates the efficiency of the proposed SSCAE to capture intrinsic structure in serum (blood)-derived proteomic data. Secondly, it is a strong indicator that the proposed deep mining model was able to deconstruct the SSCAE and interpret its weight matrices effectively, so that the proteomic patterns that can differentiate between the patients with ovarian cancer from the women without ovarian cancer were detected in two forms.

The subsets of stable predictors, that are shown in Figures 5.3 and 5.4, were used separately and collectively to construct the SVM and BDT classifiers and the average performance of these prediction models is presented in Table 5.1.

The obtained results show that both classification models achieved a high-level of accuracy, using the ensemble subset of stable predictors with HP and HN weight (i.e. All). The experimentally obtained outcomes of METABRIC dataset is presented in the following section along with a detailed discussion.

### 5.3.2 Results and Discussion of METABRIC Dataset

This section presents and discusses the experimental outcomes of applying the SSCAE and deep mining model to the METABRIC dataset with ER and PR groups, (which is illustrated in Section 3.5). Initially, the $5-$fold CV procedure was utilised to divide the dataset randomly into non-overlapping training-validation sets, as shown in Appendix A - Table 1. At each iteration, the SSCAE was trained using the training dataset and validated using the corresponding validation observations as shown in Appendix B - Figure 1 represented by the confusion matrices and the ROC curve plots of the SSCAE for the final iteration. The average AUCs of the SSCAE over CV iterations are **0.9884** and **0.9380** for ER and PR groups respectively. The outcomes of our experiments reveal that the newly learned features by the SSCAE contributed to developing a highly accurate and robust prediction model. Furthermore, the performance of each trained SCAE module was validated using the MSE between the validation set and its reconstruction as shown in Appendix A - Table 2.

For ER groups, the application of the deep mining model, based on the internal validation over variant groups of samples, generated a subset of 25 consistently selected predictors with HP weight, to be associated with ER groups, as shown in Figure 5.5. The mRNA markers were plotted in the X-axis and Y-axis alphabetically using their names, as illustrated in Figure 5.5. This figure illustrates how the expression levels of the detected genes with HP weight differ significantly be-

Table 5.1: The performance of the SVM and BDT models built on the stable predictors of ovarian cancer dataset.

| SVM-HP | SVM-HN | SVM-All | BDT-HP | BDT-HN | BDT-All |
|---|---|---|---|---|---|
| 0.8886 | 0.8975 | 0.9227 | 0.8726 | 0.8828 | 0.8964 |

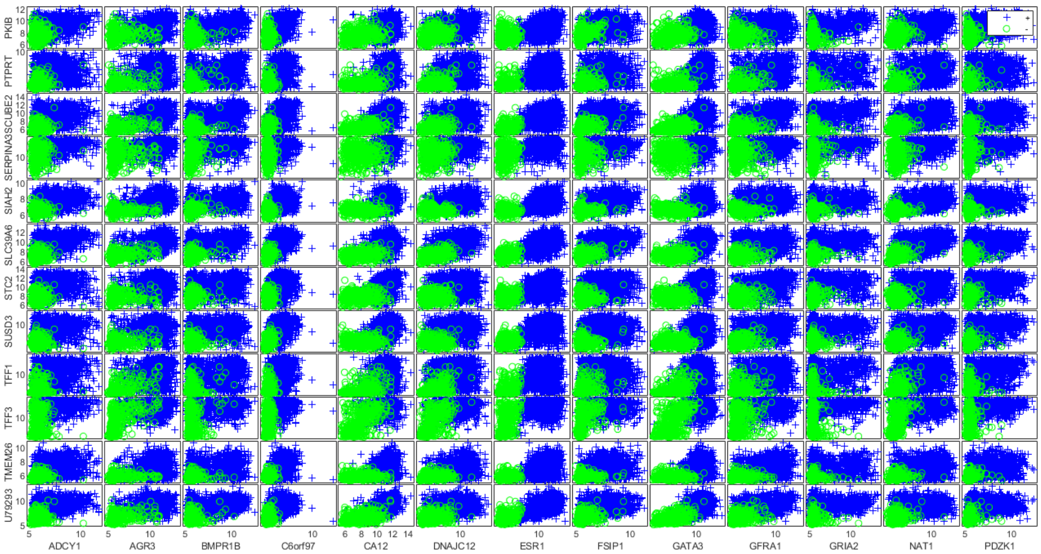


Figure 5.5: Scatter plots matrix of the stable predictors with HP weight of METABRIC dataset with ER groups.

tween ER-positives and ER-negatives. Furthermore, the examination of the five discovered subsets of candidate genes with HN weight produced 7 stable predictors, as presented in Figure 5.6. As mentioned previously, the mRNA markers were plotted in the X-axis and Y-axis alphabetically using their names, as illustrated in Figure 5.6. It can be observed in this figure that the expression levels of the recognised set of stable genes with HN weight for the patients with ER+ tumours are more likely to be lower than their expression levels for most of the samples from ER- group.

For PR groups, investigating the consistency of selection of the proposed SSCAE together with the deep mining model resulted in finding 6 consistently selected predictors with HP weight, as shown in Figure 5.7. The discovered mRNAs were plotted in the X-axis and Y-axis alphabetically using their names, as illustrated in Figure 5.7. This figure demonstrates the discrimination power of these robust genes to differentiate the patients with PR-positive tumours from PR- samples effectively. Moreover, the comparison of the selected lists of candidate mRNAs with HN weight led to identifying 5 stable predictors as illustrated in Figure 5.8. The discovered mRNAs were plotted in the X-axis and Y-axis

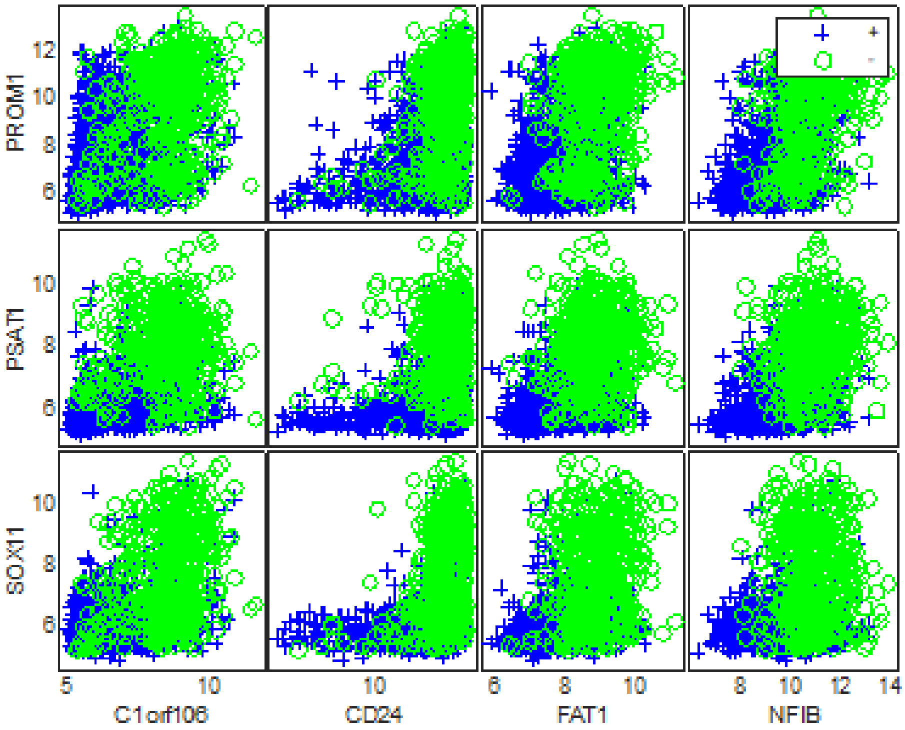


Figure 5.6: Scatter plots matrix of the stable predictors with HN weight of METABRIC dataset with ER groups.

alphabetically using their names, as illustrated in Figure 5.8. It can be observed in this figure that the expression levels of the discovered set of stable genes with HN weight are more likely to be higher for the PR-negatives than most of the patients from PR+ group, as shown in Figure 5.8.

Herein, it is very important to observe that the expression levels of HP weighted mRNA markers are more likely to be higher for the patients with ER+/PR+ tumours than most of the ER/PR-negative samples, as shown in Figures 5.5 and 5.7. In contrast, the biomarkers with HN weight exhibit higher expression levels for the observations from ER-/PR- groups, compared to the ER/PR-positive patients, as shown in Figures 5.6 and 5.8. This mechanism has also been recognised with the experimental outcomes of the ovarian cancer dataset, where HP weighted proteins exhibit high intensity values for the cancer

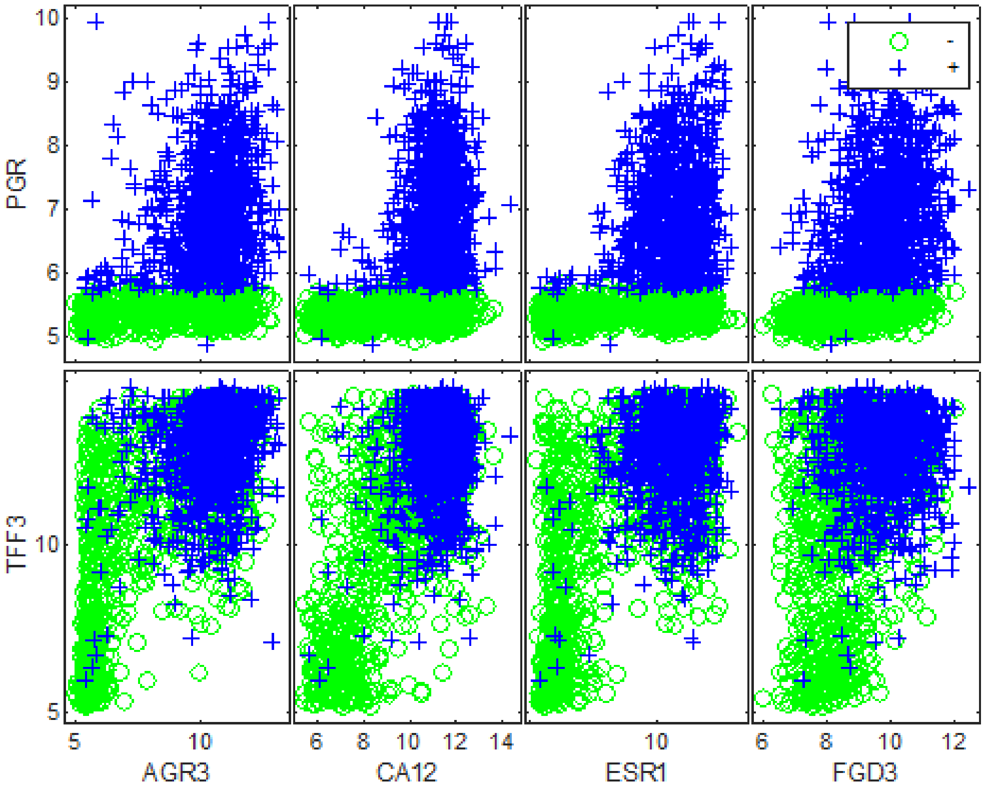


Figure 5.7: Scatter plots matrix of the stable predictors with HP weight of METABRIC dataset with PR groups.

patients, compared to the normal samples in contrast to the discovered proteins with HN weight, which show low intensity values for the cancers, in comparison to the normals. Thus, this is significant evidence that verifies firstly the effectiveness of the SSCAE to discover robustly differentially expressed genes from HDSSS genomic data and assign reliably HP and HN weight to these potential biomarkers. Secondly, this demonstrates the capability of deep mining model to interpret the weight matrices of the SSCAE and identify effectively the key genes underlying its latent representation that are positively and negatively associated with breast cancer and ER/PR positivity.

The SVM and BDT classifiers were trained using the selected subsets of stable mRNA markers with HP and HN weight separately and collectively. The predic-

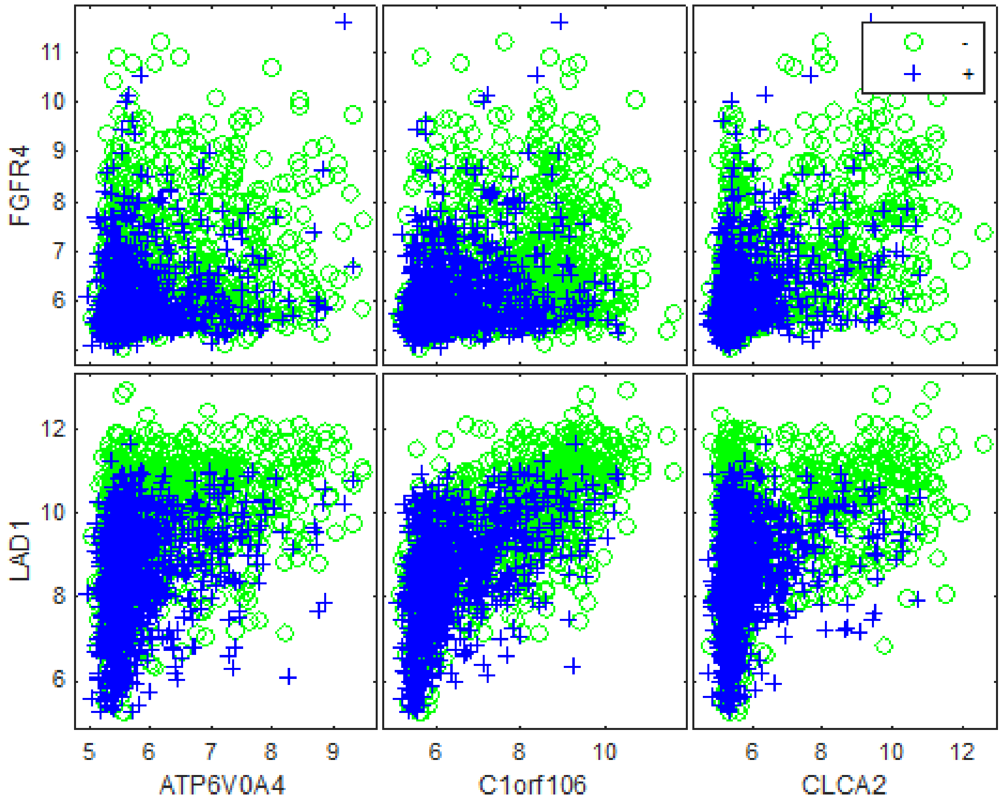


Figure 5.8: Scatter plots matrix of the stable predictors with HN weight of METABRIC dataset with PR groups.

tive performance of the learning models was validated using the corresponding

Table 5.2: The performance of the SVM and BDT models built on the stable predictors of METABRIC dataset with ER and PR groups.

| The subset of | SVM | BDT |
|---|---|---|
| HP genes with ER | 0.9820 | 0.9853 |
| HN genes with ER | 0.9023 | 0.8838 |
| All genes with ER | 0.9855 | 0.9850 |
| HP genes with PR | 0.9825 | 0.9815 |
| HN genes with PR | 0.7290 | 0.7143 |
| All genes with PR | 0.9815 | 0.9824 |

validation set and the average AUCs over CV iterations are presented in Table 5.2. For ER groups, the experimental outcomes show that the HP weighted biomarkers contributed to constructing more highly accurate and robust prediction models than HN weighted biomarkers, and integrating the subsets has improved the performance of SVM model only very slightly. Similar findings were also obtained for PR groups, such that the performance of the classification models built on the HP weighted biomarkers is significantly higher than its performance when trained using the mRNA markers with HN weight. Moreover, integrating both subsets has improved the performance of the BDT model only and very slightly.

### 5.3.3 Results and Discussion of Breast Invasive Carcinoma Datasets

This section presents and discusses the application of the proposed SSCAE together with deep mining interpretation method to the breast invasive carcinoma datasets: (Nature 2012), (Cell 2015), and (Provisional), which are explained respectively in Sections 3.3.1, 3.3.2, 3.3.3. Initially, the 5−fold CV procedure was employed to partition each dataset into 10 random training-validation sets, as shown in Appendix A - Table 3, with ER groups and Table 4 with PR groups. At each iteration, the SSCAE was trained using the training samples and validated using the corresponding validation samples, as shown in Appendix B - Figures 2, with ER groups and 3 with PR groups, represented by the confusion matrices and the ROC curve plots of the SSCAE for the final iteration. Then, the average predictive performance of the SSCAE is introduced. Furthermore, the performance of each trained SCAE module was validated using the MSE between the validation set and its reconstruction, as shown in Appendix A - Table 5, with ER groups and Table 6 with PR groups. The outcomes of our experiments for ER and PR groups will be discussed in the following subsections.

#### 5.3.3.1 ER Groups

The aim of applying the proposed SSCAE to the breast invasive carcinoma datasets is to extract relevant knowledge for estimating the status of ER. Therefore, the predictive performance of the SSCAE was assessed using AUC as shown in Table 5.3. The obtained results reveal that the SSCAE achieved high-levels of generalisation and robustness for all of the datasets, which provides strong evidence that very useful representations were discovered.

The deep mining model based on the internal validation approach was applied to interpret the weight matrices of the SSCAE. As a result, a subset of consistently selected predictors with HP weight and a subset of stable predictors with HN weight, were generated over iterations for each breast cancer dataset. Then, the external validation approach was applied to examine the potential of the proposed SSCAE and deep mining model to generalise to wider populations. Thus, the outcome was that 16 mRNA markers with HP weight were found to be generic across the datasets. The generic biomarkers are: {'AGR3', 'ESR1', 'GFRA1', 'SIAH2', 'SLC39A6', 'SCUBE2', 'C6orf97', 'ANXA9', 'CA12', 'NAT1', 'GATA3', 'PCP2', 'FSIP1', 'EVL', 'LRRC56', 'IGFALS'}, as shown in Figure 5.9. The discovered mRNAs were plotted in the X-axis and Y-axis alphabetically using their names, as illustrated in Figure 5.9. Furthermore, applying the generalisability criterion resulted in finding 16 mRNA markers with HN weight to be generic across (Nature 2012), (Cell 2015) and (Provisional) datasets. The generic biomarkers are: {'PSAT1', 'PPP1R14C', 'TMEM40', 'VGLL1', 'C1orf106', 'BBOX1', 'SOX11', 'PROM1', 'DKK1', 'PARRES1', 'S100A8', 'S100A9', 'TRPV6', 'B3GNT5', 'KRT16', 'KRT81'}, as shown in Figure 5.10. As mentioned previously, the discovered mRNAs were plotted in the X-axis and Y-axis alphabetically using their names, as illustrated in Figure 5.10.

Both Figures 5.9 and 5.10 illustrate the capability of the identified biomarkers with HP and HN weight to separate the patients with ER+ tumours from the ER-

Table 5.3: The performance of the SSCAE of the breast invasive carcinoma datasets with ER groups.

| Dataset | AUC |
|---|---|
| (Nature 2012) | 0.9404 |
| (Cell 2015) | 0.9406 |
| (Provisional) | 0.9385 |

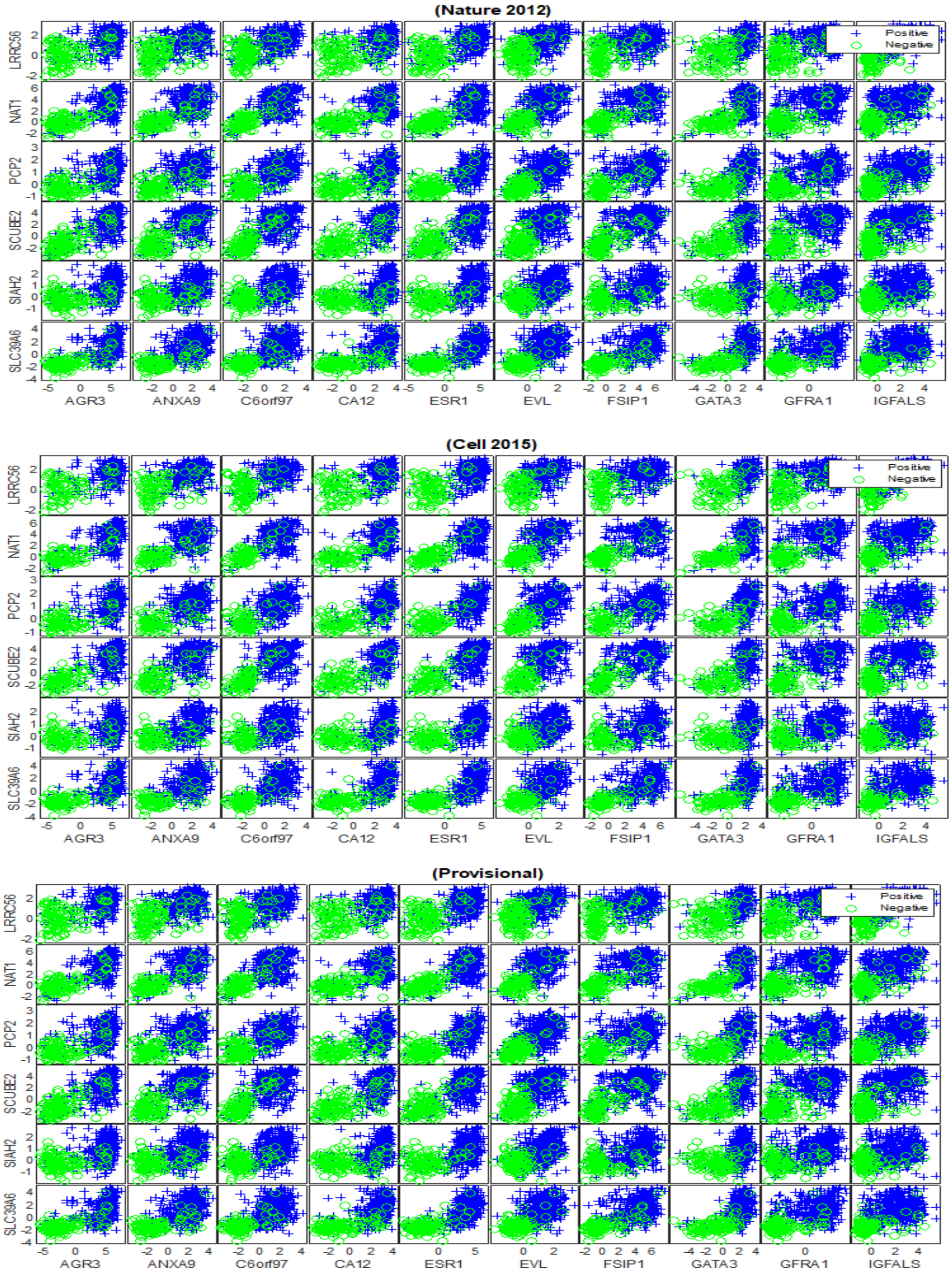


Figure 5.9: Scatter plots matrices of the generic biomarkers with HP weight of the breast invasive carcinoma datasets with ER groups.

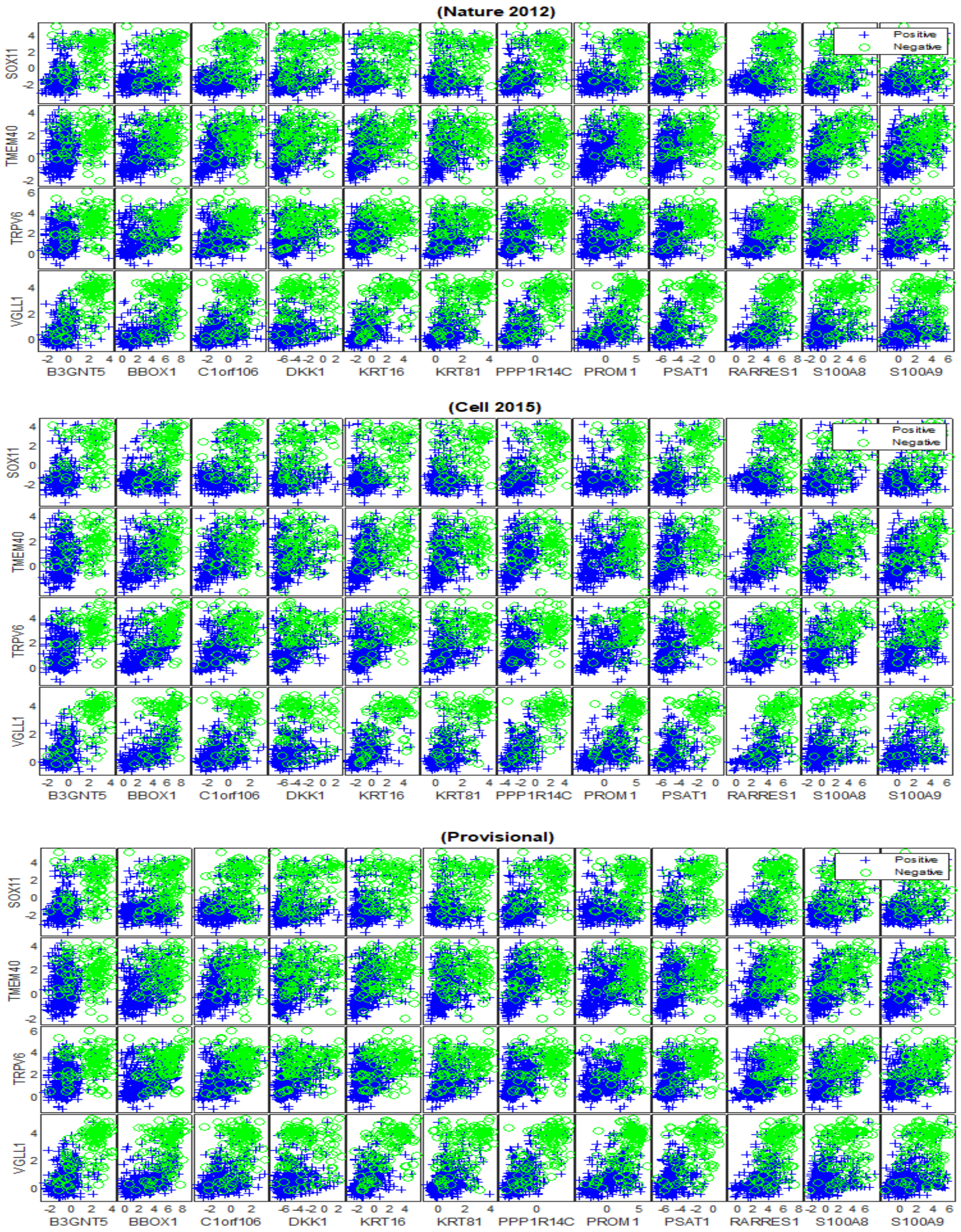


Figure 5.10: Scatter plots matrices of the generic biomarkers with HN weight of the breast invasive carcinoma datasets with ER groups.

samples over all the datasets. Herein, it is important to observe that the generic biomarkers with HP weight are highly expressed for the observations from ER+ group compared to the ER- samples, as shown in Figure 5.9. In contrast, the HN weighted mRNA markers are highly expressed for the ER-negatives compared to the ER-positives, as shown in Figures 5.10. This has also been recognised with the ovarian cancer dataset in Section 5.3.1 and the METABRIC dataset in Section 5.3.2. Therefore, there is a high potential that HP weight had been assigned by the SSCAE to the differentially expressed genes or proteins whose values for the positive cases are more likely to be higher than their values for the negatives. While the SSCAE had assigned HN weight to the deferentially expressed genes or proteins whose levels for the negative samples are more likely to be higher than their levels for the positive patients. Firstly, this mechanism demonstrates the potential of the SSCAE to exploit the unknown structure of genomic and proteomic data and capture high-level abstract and generic features. Secondly, this is strong evidence that supports the validity of the new weight interpretation method to overcome the issue of poor explanatory power associated with the deep learning and aid the researcher in opening up the so-called black box of the network to ascertain which genes were dominant within its internal representations.

The relevancy of the discovered subsets of the generic mRNA markers with HP and HN weight to the status of ER was evaluated individually and collectively using the SVM and BDT classifiers. The average predictive performance of both prediction models is shown in Table 5.4. The obtained results reveal that SVM and BDT models built on the generic biomarkers with HP weight achieved higher levels of performance than when they were trained using the generic biomarkers

Table 5.4: The performance of the SVM and BDT models built on the generic biomarkers of the breast invasive carcinoma datasets with ER groups.

| Dataset | SVM-HP | SVM-HN | SVM-All | BDT-HP | BDT-HN | BDT-All |
|---|---|---|---|---|---|---|
| (Nature 2012) | 0.9331 | 0.8695 | 0.9304 | 0.9052 | 0.8585 | 0.9034 |
| (Cell 2015) | 0.9340 | 0.8673 | 0.9340 | 0.9177 | 0.8726 | 0.9300 |
| (Provisional) | 0.9388 | 0.8714 | 0.9233 | 0.8847 | 0.8650 | 0.9244 |

with HN weight. Furthermore, the ensemble subset of the generic biomarkers (i.e. All) has improved only the predictive performance of the BDT model for (Cell 2015) and (Provisional) datasets. The application of the SSCAE together with the deep mining model based on the defined assessment criterion: predictivity, stability, and generalisability to the breast invasive carcinoma datasets produced two subsets of relevant, robust, and reproducible biomarkers. In this research, it has been shown how the discovered mRNA markers with HP weight exhibit a positive association with ER positivity where an inverse association was recognised between the identified biomarkers with HN weight and high ER levels. A detailed discussion about the type of relationship between the discovered biomarkers and breast cancer and ER/PR positivity will be overviewed in Chapter 6.

#### 5.3.3.2 PR Groups

The proposed SSCAE was also applied to the breast invasive carcinoma datasets to capture relevant knowledge for estimating the status of PR. Therefore, at each iteration, the SSCAE was trained using the training set and validated using the corresponding validation observations and its average predictive performance is shown in Table 5.5. The empirical outcomes show the capability of the SSCAE to estimate the status of PR over all the datasets with a good-level of predictivity, which reflects the usefulness of the newly discovered features.

Afterwards, the deep mining model based on the internal validation approach was applied to identify a subset of consistently selected predictors with HP weight and a subset of stable predictors with HN weight for each dataset. The investigation of the generalisation capability of the SSCAE together with the deep mining model led to detect 10 generic mRNA markers with HP weight across

Table 5.5: The performance of the SSCAE of the breast invasive carcinoma datasets with PR groups.

| Dataset | AUC |
|---|---|
| (Nature 2012) | 0.8892 |
| (Cell 2015) | 0.8975 |
| (Provisional) | 0.8846 |

(Nature 2012), (Cell 2015) and (Provisional) datasets. The biomarkers are: {‘FGD3’, ‘GFRA1’, ‘GRPR’, ‘PGR’, ‘SUSD3’, ‘GREB1’, ‘SIAH2’, ‘SCUBE2’, ‘AGR3’, ‘PGLYRP2’}, as shown in Figure 5.11. The discovered mRNAs were plotted in the X-axis and Y-axis alphabetically using their names, as illustrated in Figure 5.11. Furthermore, 10 generic mRNA markers with HN weight were detected across the breast invasive carcinoma datasets, which are: {‘LAD1’, ‘ATP6V0A4’, ‘NXPH1’, ‘C9orf58’, ‘CLCA2’, ‘FGFR4’, ‘PPP1R1A’, ‘TRPV6’, ‘C1orf115’, ‘TSPAN8’}, as shown in Figure 5.12. The discovered mRNAs were plotted in the X-axis and Y-axis alphabetically using their names, as illustrated in Figure 5.12. Figure 5.11 and Figure 5.12 illustrate the capability of the subsets of discovered biomarkers with HP and HN weight to distinguish the patients with PR+ tumours from the PR-negative samples over all the datasets.

In these matrices of plots, we can also observe that the expression levels of HN weighted mRNAs for the PR-negatives are generally higher than their expression levels for the patients with PR+ tumours, as shown in Figure 5.12, in contrast to the HP weighted biomarkers whose expression levels for the patients from the PR+ group are more likely to be higher than most of the samples from PR- group, as shown in Figure 5.11. The consistency of obtaining this decoding pattern demonstrates the effective mechanism of the deep mining model for opening up the black box of the SSCAE in a steady way. For further verification, the performance of the SVM and BDT classifiers trained using the training set that contains only the selected subsets of the generic biomarkers (separately and collectively) was validated using the corresponding validation set. The average AUCs over CV iterations are shown in Table 5.6. The outcomes of our experiments reveal that the SVM and BDT classifiers trained using the generic biomarkers

Table 5.6: The performance of the SVM and BDT models built on the generic biomarkers of the breast invasive carcinoma datasets with PR groups.

| Dataset | SVM-HP | SVM-HN | SVM-All | BDT-HP | BDT-HN | BDT-All |
|---|---|---|---|---|---|---|
| (Nature 2012) | 0.8428 | 0.8040 | 0.8532 | 0.8564 | 0.7885 | 0.8588 |
| (Cell 2015) | 0.8432 | 0.7994 | 0.8469 | 0.8654 | 0.7990 | 0.8637 |
| (Provisional) | 0.8566 | 0.8042 | 0.8521 | 0.8683 | 0.8038 | 0.8726 |

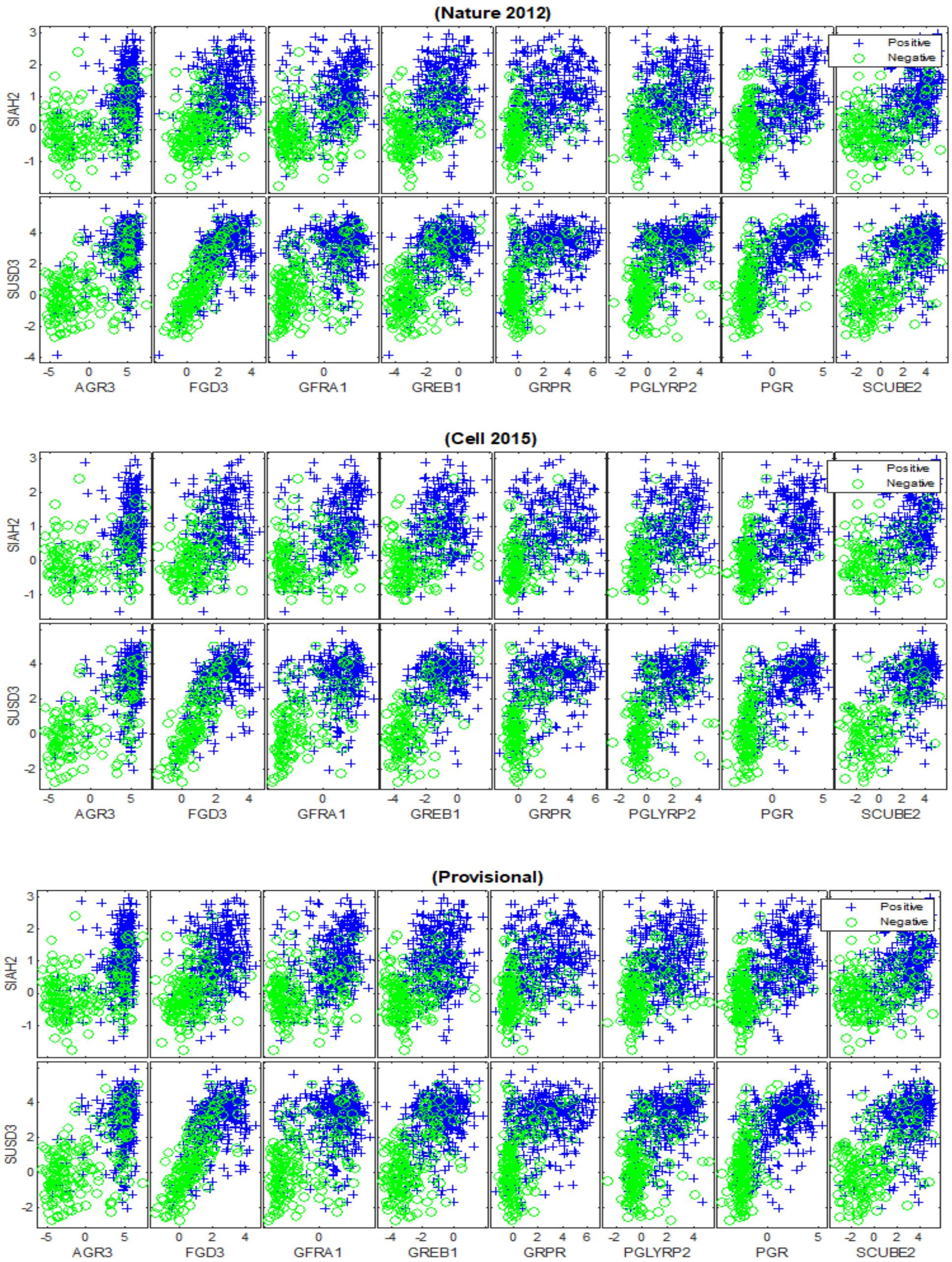


Figure 5.11: Scatter plots matrices of the generic biomarkers with HP weight of the breast invasive carcinoma datasets with PR groups.

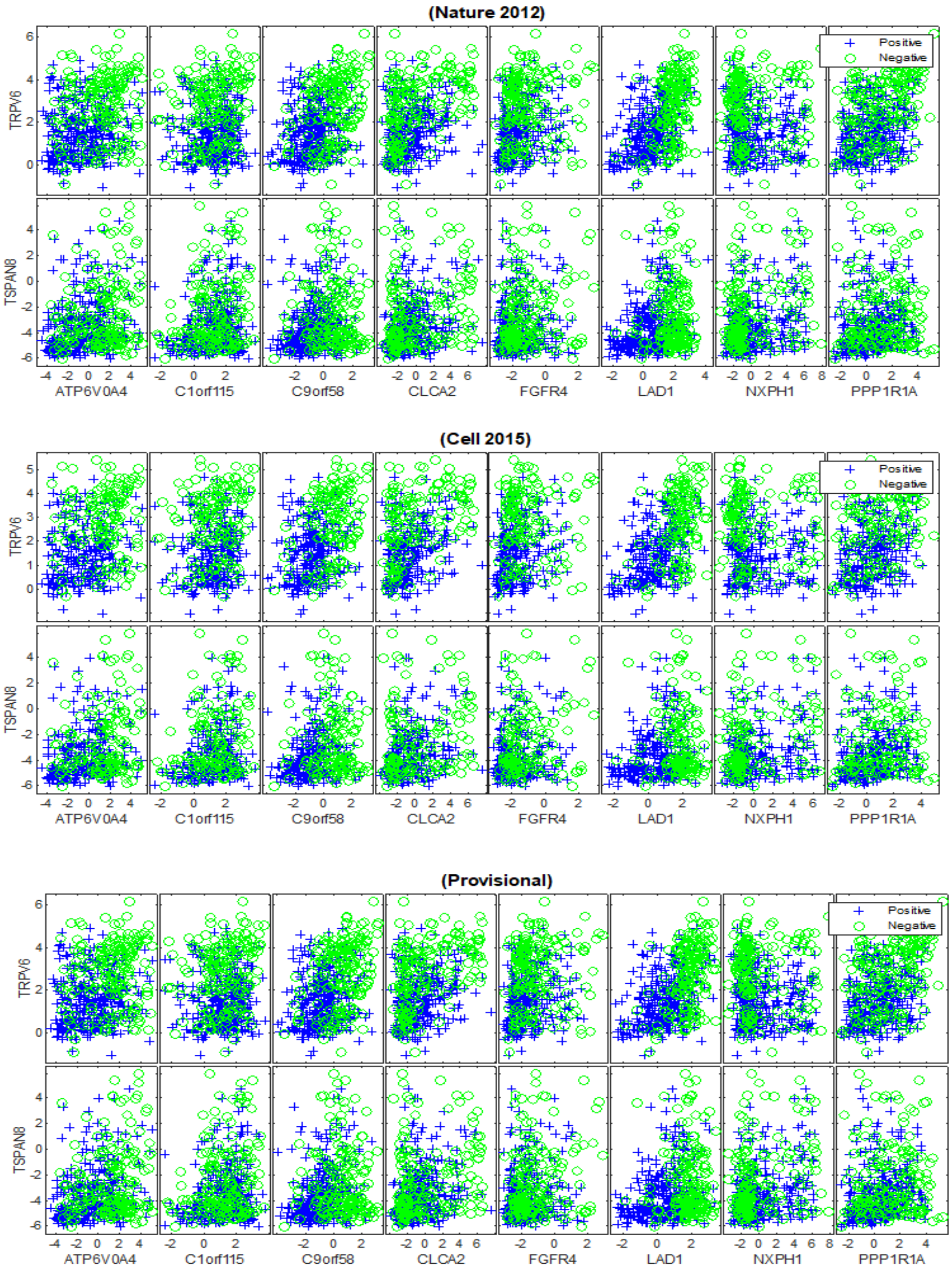


Figure 5.12: Scatter plots matrices of the generic biomarkers with HN weight of the breast invasive carcinoma datasets with PR groups.

with HP weight achieved higher levels of performance than when they trained using the generic biomarkers with HN weight. Furthermore, the integration of the biomarkers with HP and HN weight (i.e. All) has generally improved the predictive performance of both classification models very slightly. The utilisation of the SSCAE to derive cancer markers from breast invasive carcinoma datasets led to construct accurate and reliable prediction systems. Furthermore, the deep mining interpretation method contributed adding the explanatory power to that deep feature learning model and identify two subsets of salient, invariant, and generic mRNA markers that are associated positively and negatively to breast cancer and PR positivity.

### 5.3.4 Results and Discussion of Integrated Breast Invasive Carcinoma Datasets

The presented SSCAE was applied to the integrated breast invasive carcinoma datasets with ER groups: NCP1, NCP2, NCP3, which are illustrated respectively in Sections 3.4.1.1, 3.4.1.2, 3.4.1.3. Moreover, the SSCAE was utilised to learn useful representations from the integrated datasets: NC, CN, NP, PN, CP, PC, which are explained respectively in Sections 3.4.2.1, 3.4.2.2, 3.4.2.3, 3.4.2.4, 3.4.2.5, 3.4.2.5 for estimating the status of PR. Initially, the 5$-$fold- CV procedure is employed to partition each integrated dataset into 10 training-validation sets, as shown in Appendix A - Tables 7 with ER groups and 8 with PR groups. At each iteration, the SSCAE was trained using the training samples and validated using the corresponding validation samples, as shown in Appendix B - Figure 4 with ER groups and Figures 5 and 6 with PR groups, represented by the confusion matrices and the ROC curve plots of the SSCAE for the final iteration. Furthermore, the performance of each trained SCAE module was validated using the MSE between the validation set and its reconstruction, as shown in Appendix A - Table 9 with ER groups and Table 10 with PR groups. The obtained results of applying the SSCAE to the integrated datasets with ER and PR groups are presented in the following sections.

#### 5.3.4.1 ER Groups

The SSCAE was applied to the integrated datasets: NCP1, NCP2, NCP3 for predicting the status of ER based on learning high-level relevant features. The average predictive performance of the SSCAE over CV iterations is shown in Table 5.7. The outcomes of our experiments reveal that the SSCAE is performing as a highly predictive and robust classification model, which reflects its capability to learn deeply high-level abstract features that fully recovered the data. In addition, the experimental outcomes show an improvement in the predictive performance of the SSCAE when trained using the integrated datasets: NCP1, NCP2, NCP3, compared with its performance using the breast invasive carcinoma datasets with ER groups separately as shown in Table 5.3. This demonstrates the importance of having more substantial data and enough representative samples for each response group for achieving higher-levels of generalisation.

The proposed deep mining model based on the internal and external validation was also applied to the integrated datasets with ER groups to detect potential biomarkers on the basis of stability and generalisability. As a result, 12 mRNA markers with HP weight were found to be generic across the datasets: NCP1, NCP2, NCP3. The biomarkers are: {'AGR3', 'ANXA9', 'C6orf97', 'ESR1', 'GFRA1', 'NAT1', 'PCP2', 'SIAH2', 'SLC39A6', 'SCUBE2', 'CA12', GATA3'}, as shown in Figure 5.13. The biomarkers were plotted in the X-axis and Y-axis alphabetically according to their names, as illustrated in Figure 5.13. This Figure illustrates the potential of the discovered biomarkers to separate the samples with ER-positive tumours from the ER-negatives effectively. Moreover, the comparison of the identified subsets of stable predictors with HN weight produced 16 generic mRNAs which are {'B3GNT5', 'BBOX1', 'Clorf106', 'DKK1', 'HRASLS',

Table 5.7: The performance of the SSCAE of the integrated datasets with ER groups.

| Dataset | AUC |
|---|---|
| NCP1 | 0.9819 |
| NCP2 | 0.9881 |
| NCP3 | 0.9886 |

‘KRT16’, ‘PPP1R14C’, ‘PPP1R1A’, ‘PROM1’, ‘PSAT1’, ‘PARRES1’, ‘S100A8’, ‘S100A9’, ‘SOX11’, ‘TMEM40’,‘VGLL1’}, as illustrated in Figure 5.14. The mRNAs were plotted in the X-axis and Y-axis alphabetically according to their names, as illustrated in Figure 5.14. This Figure presents the ability of the detected genes to discriminate the patients with ER+ tumours from the observations in the ER-negative group effectively.

Herein, with the integrated datasets: NCP1, NCP2, NCP3, we can also observe that the SSCAE had assigned HP weight to the differentially expressed genes, which exhibit higher expression levels for the ER+ patients, in comparison to their expression levels for the ER-negatives, as shown in Figure 5.13. In contrast, the HN weights had been assigned to the key genes that are highly expressed for the ER-negative samples compared to their expression levels for the ER-positives, as shown in Figure 5.14. This is another strong evidence that demonstrates firstly the efficacy of the SSCAE to distil relevant variations from the large and noisy feature spaces of genomic data. Secondly, it promotes the validity of the proposed deep mining model as a powerful interpretation method that can deconstruct the internal state of such deep feature learning models and add explainability for the goal of identifying highly predictive and robust biomarkers that are related to the disease and the clinical outcome in two forms.

To investigate the predictivity of the discovered subsets of generic biomarkers, the performance of the SVM and BDT classifiers built on these subsets of mRNA markers separately and collectively was validated using the corresponding validation set and the average AUCs are shown in Table 5.8. Our experimental obtained results reveal that the SVM and BDT classification models, trained using the generic biomarkers with HP weight achieved higher-levels of predictive

Table 5.8: The performance of the SVM and BDT models built on the generic biomarkers of the integrated datasets with ER groups.

| Dataset | SVM-HP | SVM-HN | SVM-All | BDT-HP | BDT-HN | BDT-All |
|---|---|---|---|---|---|---|
| NCP1 | 0.9305 | 0.8843 | 0.9328 | 0.9652 | 0.9631 | 0.9664 |
| NCP2 | 0.9323 | 0.8879 | 0.9323 | 0.9675 | 0.9661 | 0.9722 |
| NCP3 | 0.9341 | 0.8846 | 0.9363 | 0.9710 | 0.9726 | 0.9735 |

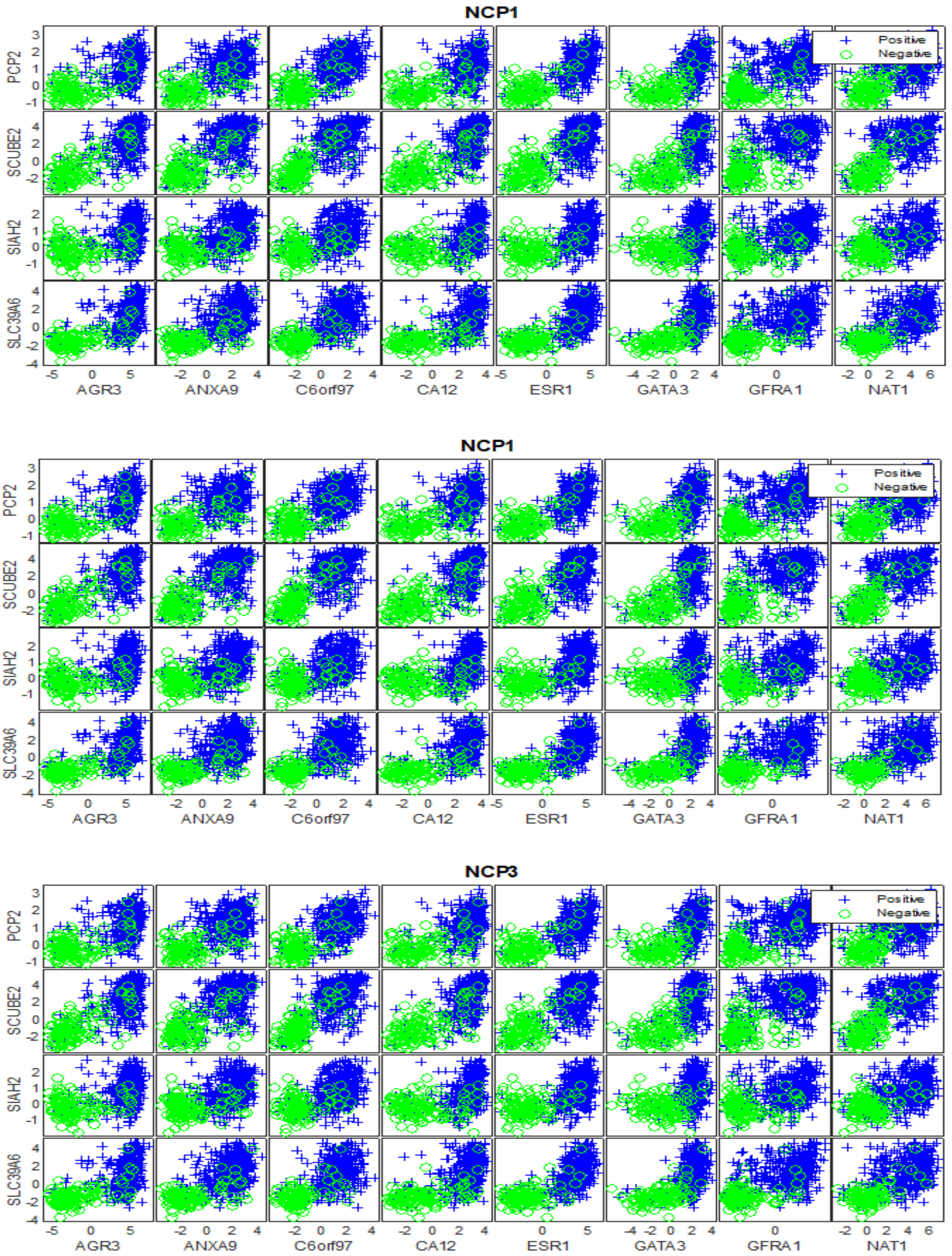


Figure 5.13: Scatter plots matrices of the generic biomarkers with HP weight of the integrated datasets with ER groups.

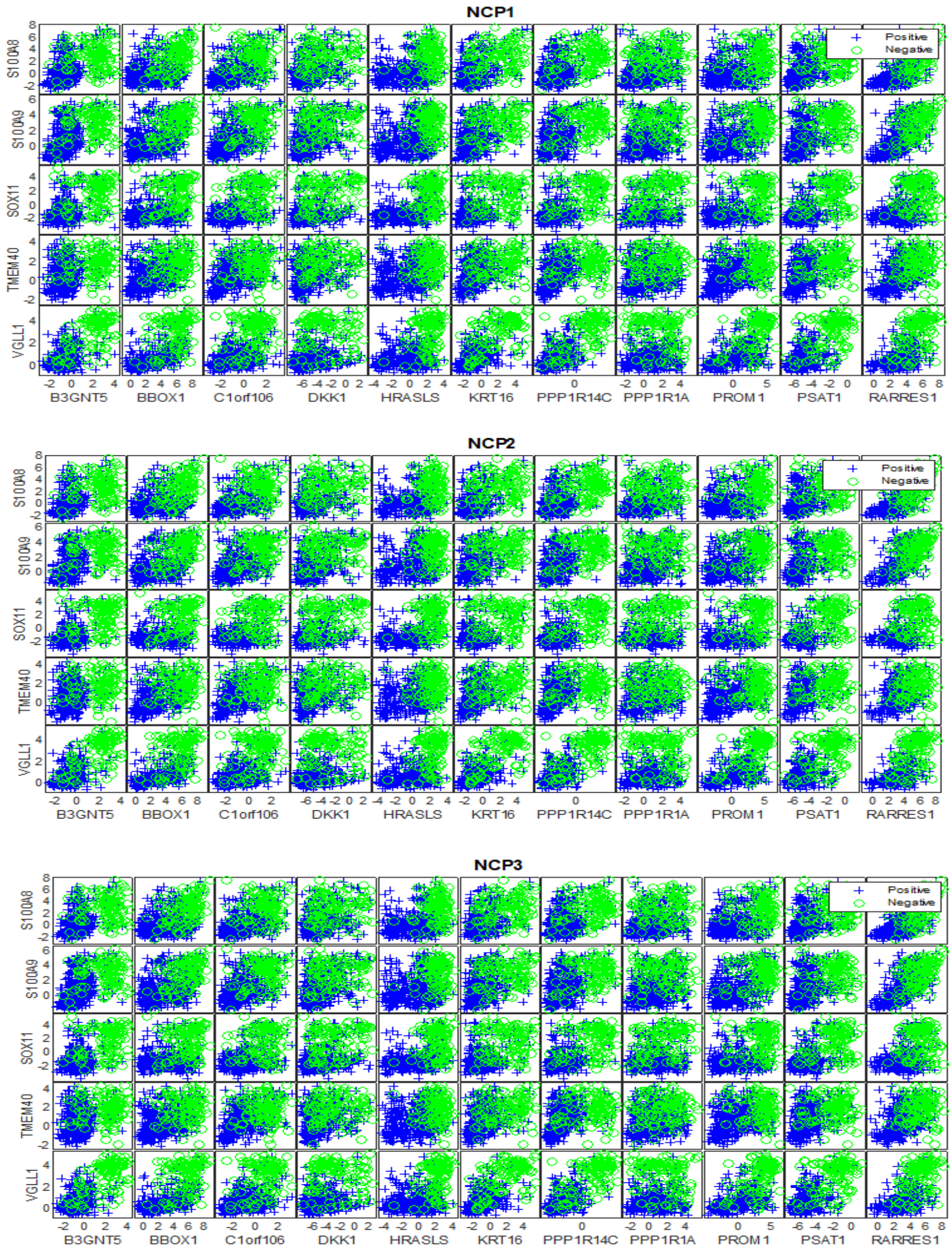


Figure 5.14: Scatter plots matrices of the generic biomarkers with HN weight of the integrated datasets with ER groups.

accuracy than when they were constructed using the generic biomarkers with HN weight. Furthermore, generally, the predictive performance of SVM and BDT classifiers is improved slightly when they trained using the integrated subset of generic biomarkers. The outcomes of our experiments also show that the BDT classification model performed better than the SVM classifier using the integrated datasets, in comparison to its performance using the breast invasive carcinoma datasets separately, which reflects the significant impact of having more substantial data, whose response groups are well-balanced on the performance of that learning model.

#### 5.3.4.2 PR Groups

The SSCAE was also applied to the integrated datasets: NC, CN, NP, PN, CP, PC, in order to detect relevant features for estimating the status of PR and the average AUC of the SSCAE is presented in Table 5.9. The outcomes of our experiments confirm discovering very useful knowledge by the SSCAE by achieving high-levels of generalisation and robustness for all of the integrated datasets, as shown in Table 5.9. Furthermore, the obtained results show an improvement in the generalisation ability of the SSCAE using the integrated dataset: NC, CN, NP, PN, CP, PC, compared with its performance using the breast invasive carcinoma datasets with PR groups separately, as shown in Table 5.5. Therefore, the construction of the SSCAE using more substantial data whose response groups are well-represented can have the great potential to improve the performance of that deep learning model.

The proposed deep mining model was applied based on the stability and generalisability criterion to detect HP and HN weighted genes. The obtained results from our experiments show that 8 mRNA markers with HP weight were found to be generic across the integrated datasets: NC, CN, NP, PN, CP, PC. The

Table 5.9: The performance of the SSAE of the integrated datasets with PR groups.

| NC | CN | NP | PN | CP | PC |
|---|---|---|---|---|---|
| 0.9584 | 0.9667 | 0.9647 | 0.9786 | 0.9661 | 0.9527 |

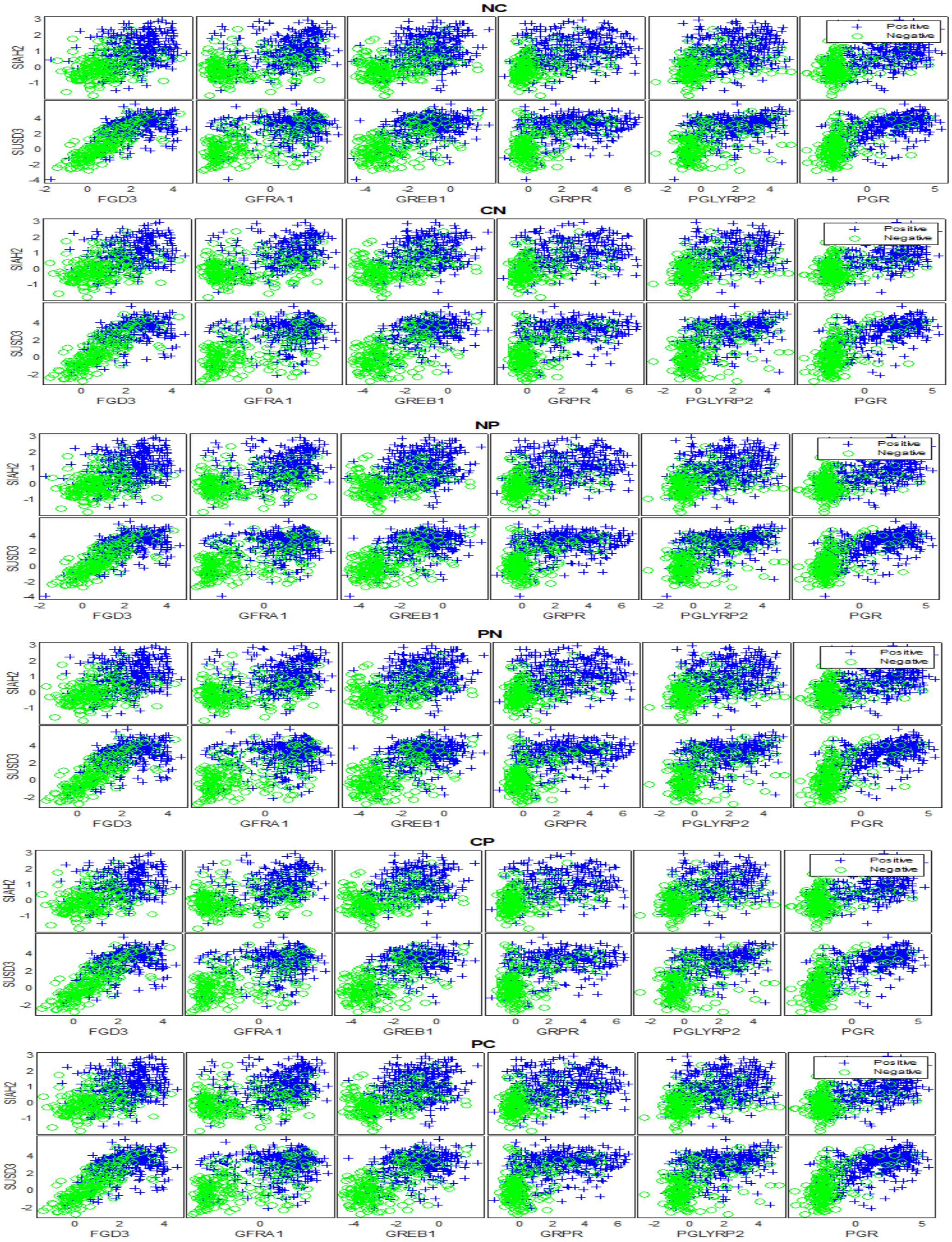


Figure 5.15: Scatter plots matrices of the generic biomarkers with HP weight of the integrated datasets with PR groups.

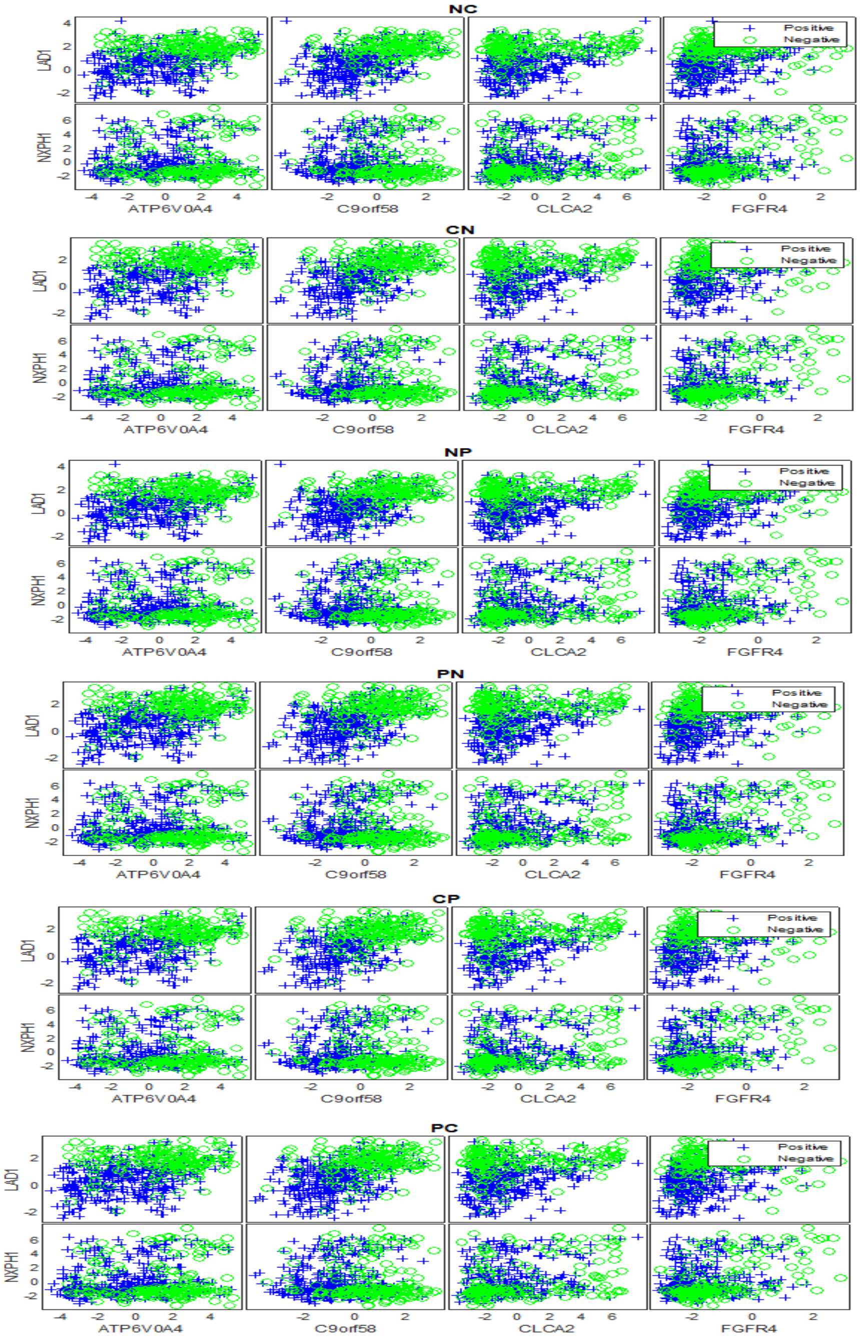


Figure 5.16: Scatter plots matrices of the generic biomarkers with HN weight of the integrated datasets with PR groups.

biomarkers are: {'GFRA1', 'GRPR', 'PGLYRP2', 'PGR', 'SIAH2', 'SUSD3', 'FGD3', 'GREB1'}, as shown in Figures 5.15. Furthermore, six mRNAs were found to be generic across the integrated datasets with PR groups. These biomarkers are: {'ATP6V0A4', 'NXPH1', 'CLCA2', 'FGFR4', 'LAD1', 'C9orf58'}, as illustrated in Figure 5.16. The discovered biomarkers were plotted in the X-axis and Y-axis alphabetically according to their names, as shown in Figure 5.15 and Figure 5.16. These Figures illustrate the capability of the recognised mRNAs to discriminate the samples with PR+ tumours from those in PR- group efficiently. Furthermore, it can also recognise the consistency of the SSCAE over multiple independent datasets in assigning HP weight to the deferentially expressed mRNAs, whose levels for the PR+ patients are high compared to the PR-negatives, whereas HN weight had been assigned by the SSCAE to the mRNA markers that are lowly expressed for the PR-positives, in comparison to the PR-negatives. These findings assert firstly the feasibility of the SSCAE as an effective feature learning model that can deeply capture intrinsic structure from HDSSS omics data. Secondly, our outcomes demonstrate the validity and capability of the proposed deep mining model for providing explainability to such deep learning models, which is a crucial element of prediction systems used by health practitioners and decision-making professionals.

The predictive performance of the SVM and BDT classification models built on training samples restricted to the identified subsets of generic biomarkers (separately and collectively) was validated using the corresponding validation samples and the average AUCs are shown in Table 5.10. The obtained results reveal that

Table 5.10: The performance of the SVM and BDT models built on the generic biomarkers of the integrated datasets with PR groups.

| Dataset | SVM-HP | SVM-HN | SVM-All | BDT-HP | BDT-HN | BDT-All |
|---|---|---|---|---|---|---|
| NC | 0.8804 | 0.8195 | 0.8929 | 0.9372 | 0.8728 | 0.9327 |
| CN | 0.8743 | 0.8132 | 0.8806 | 0.9364 | 0.8728 | 0.9345 |
| NP | 0.8744 | 0.8167 | 0.8830 | 0.9387 | 0.8889 | 0.9371 |
| PN | 0.8805 | 0.8242 | 0.8919 | 0.9531 | 0.8921 | 0.9530 |
| CP | 0.8757 | 0.8162 | 0.8888 | 0.9132 | 0.8724 | 0.9357 |
| PC | 0.8818 | 0.8302 | 0.9005 | 0.9401 | 0.8713 | 0.9371 |

the subset of generic biomarkers with HP weight contributed to constructing more highly accurate and robust prediction systems than the subset of generic biomarkers with HN weight. Furthermore, the integration of the subsets (i.e. All) has improved the predictive performance of SVM and BDT models slightly. The obtained results also reveal that the predictive performance of the BDT classifier is consistently higher than the SVM model for the integrated datasets with ER and PR groups overall the subsets of generic biomarkers.

## 5.4 Discussion

This chapter investigated the value of applying state-of-the-art DL methods to the problem of automatically determining salient biomarkers of cancers of interest from HDSSS omics data. Therefore, the outcome of our investigations was proposing the SSCAE on the basis of multiple levels of *sparse* and *compressed* representations of increasing abstractions, in order to mitigate against the key challenges that arise from handling HDSSS omics data. In addition, a novel method of interpreting the internal state of the SSCAE was developed and proved invaluable for detecting what these deep feature learning models had determined to be salient biomarkers. Considering the challenging issues of HDSSS omics data, the SSCAE was able to capture enough of the interesting complexity underlying the available biological samples and spell out a small proportion of relevant and insensitive factors, therefore the generic biomarkers were discovered robustly across a wide range of independently generated breast cancer samples. Furthermore, the empirical findings of our research emphasise the importance of using more substantial data whose response groups are well-balanced when optimising deep neural networks for high-levels of generalisability and robustness.

The introduction of the new weight interpretation method proposed in this chapter proved to be very effective in opening up the black box of SSCAE for this particular task. A detailed evaluation of the SSCAE weights revealed that the deep neural network had assigned HP weight to those genes, whose expression levels are more likely to be higher for the positive samples than for the negative samples. Likewise, the deep network had assigned HN weight for the biomarkers, whose expression levels are more likely to be higher for the negative samples than for the positives. This provides a robust discriminative basis with which to accurately classify positive and negative samples. The experimental outcomes provide strong evidence that the proposed deep mining model introduced in this chapter, is able to robustly identify salient, invariant and generic biomarkers for breast cancer. The clinical relevance of the detected biomarkers will be discussed in the following chapter.

# Chapter 6

# Biomarkers and Bioinformatics

## 6.1 Introduction

The main objective of validating and evaluating the proposed feature mining models is to assess their ability to discover robustly relevant knowledge to the cancers of interest from the HDSSS genomic and proteomic data. Therefore, the discovered biomarkers for breast cancer and the hormone receptors ER and PR were estimated in the previous chapters in terms of predictivity, stability, and reproducibility over multiple datasets that were independently generated and derived from different sets of biological samples. The outcomes of our experiments reveal that the discovered biomarkers demonstrate computational and biological relevance as well as the capability to construct highly accurate and reliable prediction models. These findings were proved using the most suitable quality assessment metrics, unlike many biomarker discovery models proposed in the literature which lacked the utilisation of robust evaluation and independent experimental validation. Since publicly available genomic and proteomic datasets are utilised in this research, thus all the required information is provided to allow the reproducibility of the results.

In addition to the assessment metrics, the verification of the clinical relevance of the detected biomarkers to breast cancer, ER and PR is another crucial step that should be adopted to indicate the scientific quality. The assessment of the clinical relevance of newly discovered biomarkers to a disease or clinically relevant conditions is usually conducted by bioinformatics studies. Therefore, the discovered biomarkers will be evaluated in this chapter with respect to their relevance to breast cancer revealed by bioinformatics research in the literature. It is important to emphasise that, at the time of writing, *each study has identified and discussed the markers for breast cancer individually, and no research has found or examined the combination of these biomarkers or some of them simultaneously.*

Furthermore, this chapter discusses the type of relationship recognised in this PhD research between each individual molecular marker and ER/PR expression levels for the goal of better understanding the biological mechanism underlying the association. Discovering robust biomarkers and identifying their relation to human breast cancers can allow more personalised medicine approaches to be developed, which could help in detecting, managing, and treating this heterogeneous disease.

## 6.2 Discovered Biomarkers with HP Weight for ER

This section discusses the relevance of the recognised biomarkers with HP weight to breast cancer and the hormone receptor ER in term of what has been conducted in the literature from bioinformatics analysis research. Furthermore, the association between each mRNA marker and the oestrogen receptor observed in this research will be examined and discussed to understand the type of existent relationship and provide conclusive evidence.

- **{'ESR1'}**. According to Cancer Genetics Website[1] *"This gene encodes an estrogen receptor. Estrogen and its receptors are essential for sexual development and reproductive function, but also play a role in other tissues such as bone. Estrogen receptors are also involved in pathological processes, including breast cancer, endometrial cancer, and osteoporosis"*. Several studies have explored the relevance of ESR1 gene to breast cancer. Holst et al. [136] discussed the findings of five studies that showed a correlation between elevated ESR1 and high ER level. Then, they summarised that *"there*

[1] http://www.cancerindex.org/geneweb/ESR1.htm

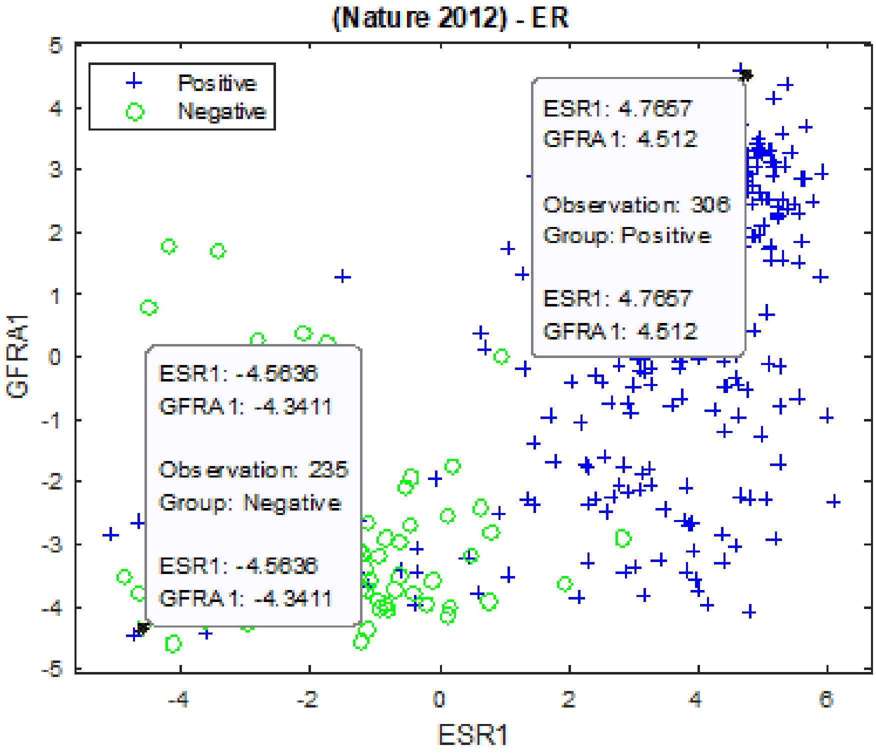


Figure 6.1: Scatter plot of ESR1 and GFRA1 of (Nature 2012) dataset with ER groups, illustrating that the observation 306 from the ER+ group has high expression levels of (ESR1, GFRA1), which are $(4.7657, 4.512)$ in comparison to the observation 235 from the ER- group, which has low expression levels of (ESR1, GFRA1), which are $(-4.5636, -4.3411)$.

*is growing evidence that ESR1 gain or amplification is a fairly frequent event in breast cancer"*. Similar findings were found by the study [174], which confirmed the existence of amplifications and gains of the ESR1 in breast cancer, where a strong positive correlation between ESR1 and ER was recognised. In the literature, several researchers have investigated the role of ESR1 as potential prognostic and predictive biomarkers for breast cancer [8, 12, 171, 178, 207, 243, 244]. In this thesis, evidence of a positive correlation was found between the expression patterns of ESR1 and ER, so that elevated ESR1 contributes to the positivity of ER, as illustrated in Figure 6.1. The increase of ESR1 mRNA expression levels in the ER+

tumours was observed over a wide range of breast cancer samples of (Nature 2012), (Cell 2015), (Provisional) and (METABRIC), as well as the integrated datasets with ER groups: NCP1, NCP2, and NCP3. Therefore, further investigations can be conducted by domain experts to examine the potential of ESR1 to be utilised in the early detection and monitoring the progression of this heterogeneous disease.

■ **{'GFRA1'}**. GFRA1 has been revealed by Immunohistochemistry (IHC) as under-expressed in normal tissue and over-expressed in subsets of breast cancers [34]. Moreover, several studies have shown that GFRA1 exhibits over-expression in the majority of breast cancers [36, 88, 89, 239, 317]. A recent study [95] that discussed the emerging role of the GDNF family in neoplasm has stated that GFRA1 mRNA expression is detected in breast tumour samples, and is associated with ER expression. Recently, GFRA1 has been identified in [34] as a breast cancer tumour associated antigens. Bhakta et al. in another recent study [31] have confirmed the abundant expression of GFRA1 in luminal *A* breast cancer tissues, whereas minimal or no expression was observed in most normal tissues. In this thesis, evidence of a positive correlation was found between GFRA1 mRNA and ER levels, as presented in Figure 6.1, and across a broad range of breast cancer samples of (Nature 2012), (Cell 2015), (Provisional) and (METABRIC), as well as the integrated datasets with ER groups. A more personalised treatment or monitoring planning for breast cancer could be developed by investigating further the mechanism underlying the association between the expression patterns of GFRA1 and ER.

■ **{'AGR3'}**. The human Anterior Gradient (AGR) family is composed of three proteins, AGR1, AGR2 and AGR3, all belonging to the protein disulfide isomerase (PDI) [5]. AGR3 was identified in [105] as a potential marker for diagnosis and prognosis of breast cancer from blood, and it was found to be significantly associated with ER. The research study [222] found AGR3 to be significantly correlated with ER. While the researchers in [295] identified AGR3 as a potential marker for triple-negative breast cancer. In this thesis, evidence of a positive association between the expression patterns

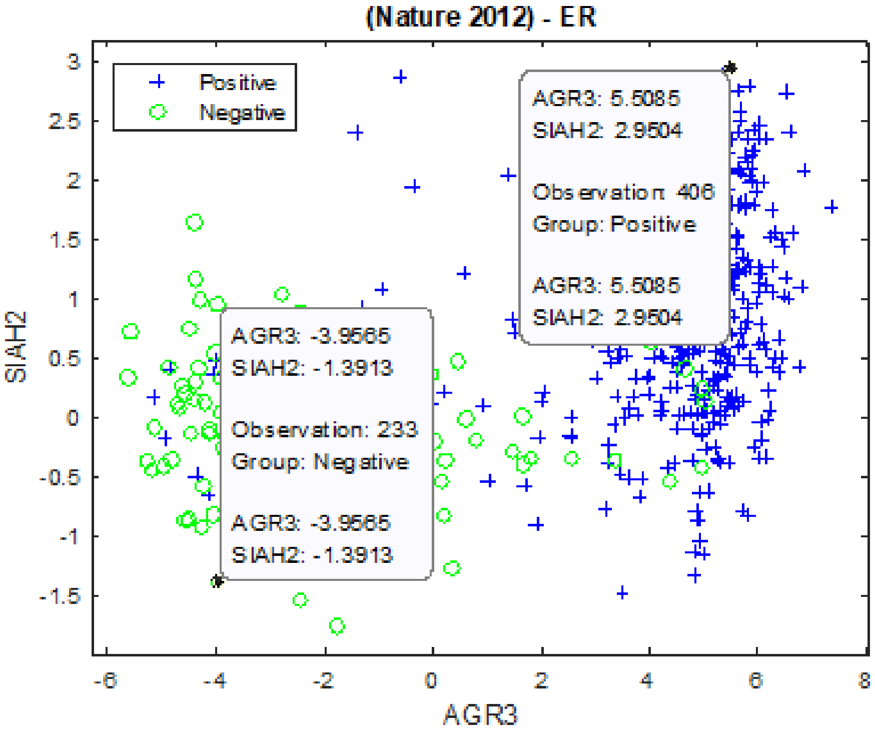


Figure 6.2: Scatter plot of AGR3 and SIAH2 of (Nature 2012) dataset with ER groups, illustrating that the observation 406 from the ER+ group has high expression levels of (AGR3, SIAH2), which are $(5.5085, 2.9504)$ in comparison to the observation 233 from the ER- group, which has low expression levels of (AGR3, SIAH2), which are $(-3.9565, -1.3913)$.

of AGR3 mRNA and ER, as shown in Figure 6.2, and across the datasets: (Nature 2012), (Cell 2015), (Provisional) and (METABRIC), as well as the integrated datasets with ER groups. Identifying a potential molecular marker in bodily fluids (e.g. serum) can contribute significantly in the early detection of breast cancer and evaluating the development of this complicated disease. Therefore, further studies are required to gain insights into the process underlying the association between AGR3 and ER.

- **{'SIAH2'}**. In the literature, it has been shown that the expression level of SIAH2 is correlated with breast cancer aggressiveness and overall patient survival [3]. According to [54], SIAH2 has been detected mainly in ER+

tumours. A positive relationship between SIAH2 mRNA and ER expression levels was detected by the researchers in [149]. Later [298], they found that in ER+ breast cancer, high levels of SIAH2 associated with unfavorable outcome in primary breast cancer and treatment outcome in metastatic breast cancer. Similar results had been described earlier by Chan et al. [49] that found a high expression level of SIAH2 is associated with an unfavorable relapse-free survival. Sun et al. [271] recognised SIAH2 to be over-expressed in invasive breast cancer comparing to normal or ductal carcinoma in situ tissues. In our biomarker discovery study, a positive relationship between the expression patterns of SIAH2 and ER was recognised, in which elevated SIAH2 is observed in the ER+ tumours, as presented in Figure 6.2. The association were detected across multiple independent datasets: (Nature 2012), (Cell 2015), (Provisional), and (METABRIC) as well as the integrated datasets with ER groups. As a result, this mRNA marker could be examined further to indicate its contribution to the heterogeneity of breast cancer.

- **{'C6orf97'}**. According to the Gene Copoeia website[1], *"the function of Chromosome 6 open reading frame 97 and its encoded protein is not known. Several genome-wide association studies have implicated the region around this gene to be involved in breast cancer and bone mineral density"*. Zheng et al. [338] found in a Chinese population, a SNP in the region between C6orf97 and ESR1 increased breast cancer risk where similar findings were detected in a European population. C6orf97 was found in [80] to be contributed to the phenotype associated with ER positivity. Yamamoto et al. [325] detected that C6orf97 shows significant worse prognostic values, especially in luminal B breast cancer. This thesis reveals a positive relationship between the expression patterns of C6orf97 and ER, as shown in Figure 6.3, and over multiple independent datasets, which are (Nature 2012), (Cell 2015), (Provisional), and (METABRIC) as well as the integrated datasets NCP1, NCP2, and NCP3. These findings can motivate the researchers and the graduate students in the bioinformatics and biology to conduct functional

[1] http://genecopoeia.com/gene/hs-c6orf97.html

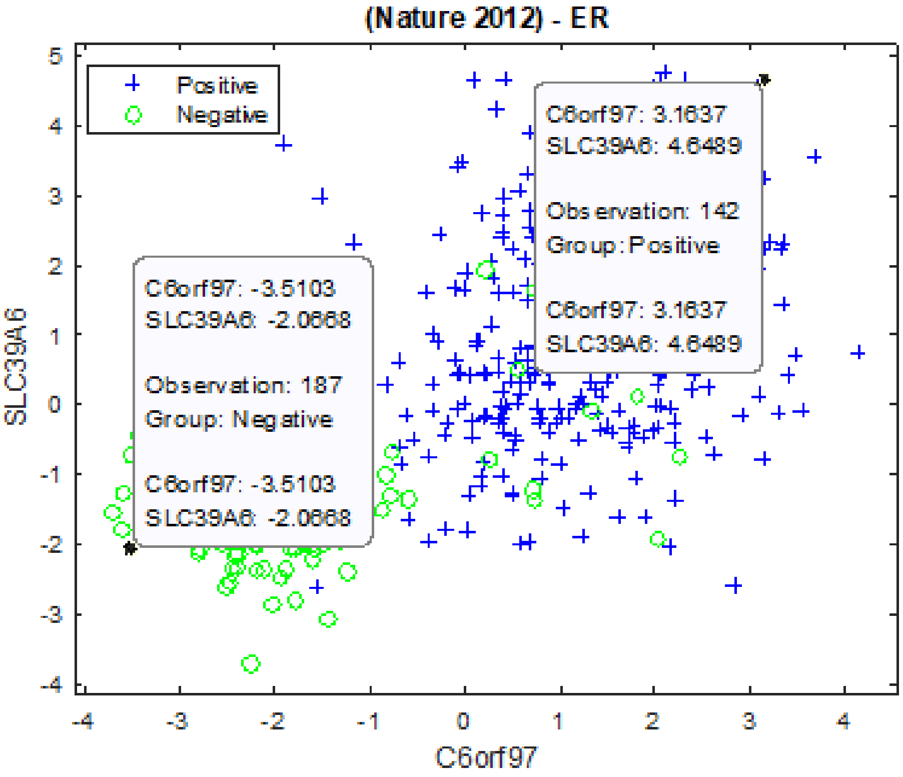


Figure 6.3: Scatter plot of C6orf97 and SLC39A6 of (Nature 2012) dataset with ER groups, illustrating that the observation 142 from the ER+ group has high expression levels of (C6orf97, SLC39A6), which are $(3.1637, 4.6489)$ in comparison to the observation 187 from the ER- group, which has low expression levels of (C6orf97, SLC39A6), which are $(-3.5103, -2.0668)$.

studies of this locus.

- **{'SLC39A6'}**: Solute carrier family 39 member 6, according to the NCBI website[1], *"zinc is an essential cofactor for hundreds of enzymes. It is involved in protein, nucleic acid, carbohydrate, and lipid metabolism, as well as in the control of gene transcription, growth, development, and differentiation. SLC39A6 belongs to a subfamily of proteins that show structural characteristics of zinc transporters"*. The research study [288] stated that SLC39A6 is estrogen regulated and existent in ER+ breast cancer as well

[1] https://www.ncbi.nlm.nih.gov/gene/25800

as in tumours that spread to the lymph nodes. Kasper et al. [158] found that SLC39A6 mRNA and protein level could act as novel biomarkers of clinical outcome in breast cancer patients. The authors in [274] emphasised targeting SLC39A6 for the treatment of metastatic breast cancer. Our genomic analysis study reveals a positive correlation between SLC39A6 and ER expression levels in breast cancer so that elevated SLC39A6 contributes to the ER positivity, as illustrated in Figure 6.3. This type of association was detected over a wide range of breast invasive carcinoma samples. The potential contribution of SLC39A6 mRNA to ER positivity can be further examined in order to enable utilising zinc transporter LIV-1 (SLC39A6) in the construction of the breast cancers prediction systems or the selection process of optimum therapy.

- **{'ANXA9'}**. ANXA9 has been found by a recent study [322] to be significantly correlated with ESR1. The strong expression of ANXA9 was also found by [264] to be correlated with the metastasis of breast cancer to the bone. The research team at Berkeley Lab [141] found ANXA9 to be highly expressed in approximately half of the patients and a significant relationship between ANXA9 and aggressive breast cancers was indicated. In this thesis, we have demonstrated the existence of high expression levels of ANXA9 mRNA in ER+ tumours, as explained in Figure 6.4, and over multiple breast invasive carcinoma datasets and the integrated datasets with ER groups. Therefore, the capability of ANXA9 to be used as a diagnostic or prognostic marker in the early detection or evaluation of breast cancers can be further studied, in which the discovered knowledge can be transferred to personalised and precision medicine.

- **{'NAT1'}**. The authors in [305] stated that there is growing evidence that demonstrates the biological role of NAT1 in the progression of breast cancer and suggested NAT1 transcripts as candidate prognostic markers in ER+ breast cancer. Adam et al. [4] confirmed NAT-1 mRNA level to be overexpressed in clinical breast cancers and a strong association of NAT-1 staining with ER+ tumours was demonstrated. The recent analysis study [45] in normal, primary breast tissues and breast cancer cell lines has suggested

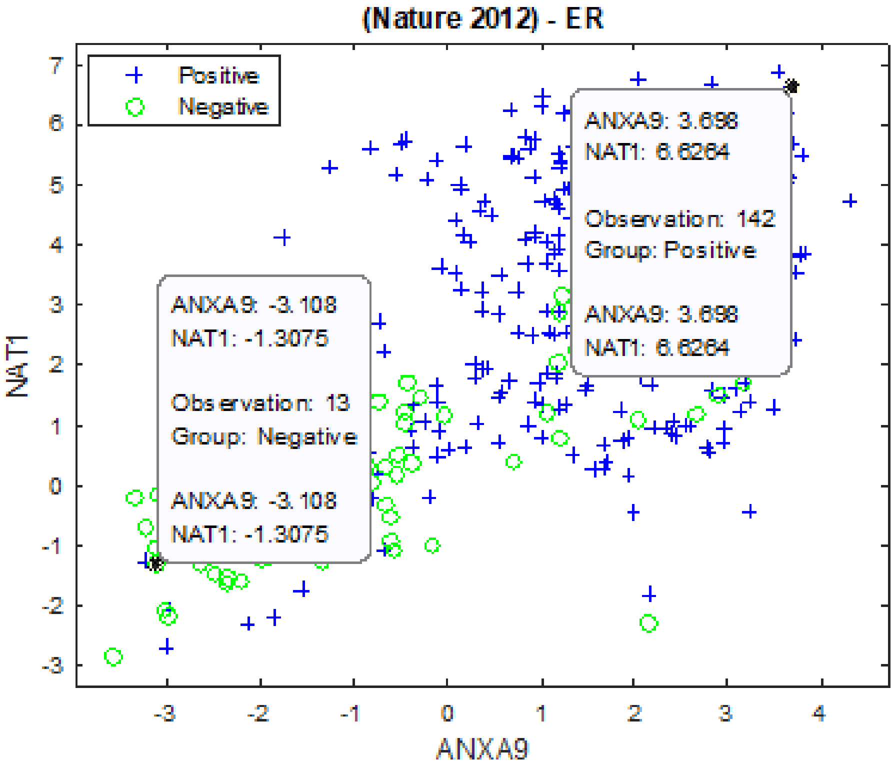


Figure 6.4: Scatter plot of ANXA9 and NAT1 of (Nature 2012) dataset with ER groups, illustrating that the observation 142 from the ER+ group has high expression levels of (ANXA9, NAT1), which are (3.698, 6.6264) in comparison to the observation 13 from the ER- group, which has low expression levels of (ANXA9, NAT1), which are (−3.108, −1.3075).

that NAT1 and ESR1 expression may have overlapping regulation. Furthermore, the NAT1 expression levels were shown in [86] to have a positive correlation with ER. In this thesis, evidence of a positive correlation was detected between NAT1 and ER, so that high mRNA levels of NAT1 can be observed in the patients with ER+ tumours compared to the ER-negative samples, as shown in Figure 6.4. The association was recognised across a wide range of breast cancer samples of (Nature 2012), (Cell 2015), (Provisional), and (METABRIC) as well as NCP1, NCP2, and NCP3. Therefore, targeting the NAT1 gene for further investigations can contribute to exploring more knowledge about this gene and its role in human breast cancers

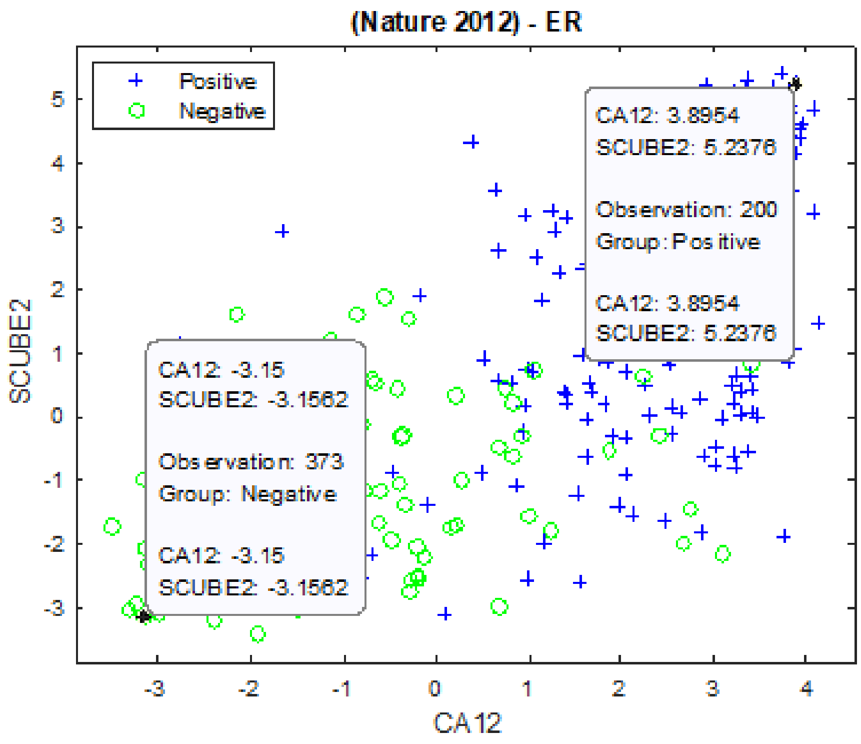


Figure 6.5: Scatter plot of CA12 and SCUBE2 of (Nature 2012) dataset with ER groups, illustrating that the observation 200 from the ER+ group has high expression levels of (CA12, SCUBE2), which are (3.8954, 5.2376) in comparison to the observation 373 from the ER- group, which has low expression levels of (CA12, SCUBE2), which are (-3.15, -3.1562).

and ER positivity.

- **{'CA12'}**. CA12 has been recognised by Barnett et al. [17] to be highly correlated with ERA in human breast tumours. The research study [185] detected that CA12 and AGR3 are up-regulated in ER+ tumours, while Watson et al. [311] found CAXII (CA12) to be frequently expressed in invasive breast carcinoma. This thesis identifies a positive association between CA12 mRNA and ER, as illustrated in Figure 6.5, and across a wide range of breast samples of (Nature 2012), (Cell 2015), (Provisional), and (METABRIC). This provides strong evidence that this gene has the potential to be a clinical biomarker for human breast cancer and therefore further studying is imperative.

- **{'SCUBE2'}**: Several studies have detected SCUBE2 expression in primary invasive breast tumours [91,226,299]. The researchers in [187] claimed that SCUBE2 plays a major role in suppressing breast-carcinoma-cell mobility and invasiveness. A high degree of correlation was observed in [10] between the expression levels of ESR1 and several markers, including SCUBE2 and it was found to be related to the ER expression. Herein, our study reveals evidence of a correlation between elevated SCUBE2 mRNA and high ER expression, as presented in Figure 6.5. This association was validated over multiple breast invasive carcinoma datasets that are collected from completely different studies. Consequently, further studies are required to investigate the potential usefulness of this biomarker in the early detection or evaluation of the progression of breast cancers.

- **{'EVL'}**. A recent study [225] has shown that EVL is up-regulated in ER+ tumours and suppresses invasion, and that EVL levels are reduced in tumours after anti-estrogenic hormone therapy. Similar funding were found earlier by Tavares et al. [286] that discovered EVL to be high in luminal breast tumours. Another study [140] recognised that the expression level of EVL was higher in breast tumours compared to normal tissues and its up-regulation was positively associated with the clinical stages of breast cancer. Moreover, it added that EVL may be implicated in invasion and/or metastasis of human breast cancer. In this thesis, evidence of a positive association was discovered between EVL mRNA expression and high ER levels, as shown in Figure 6.6, and across a wide range of independently generated breast cancer samples. This supports the potential of this gene to be a biomarker to ER+ breast cancer, thus further investigations are required to determine the biological role of the Ena/VASP protein, EVL in this heterogeneous disease.

- **{'PCP2'}**. Little information is available in the literature about this gene, particularly its relation to breast cancer. Only recently, the genome-wide association study [245] has identified genes associated with neuropathy in

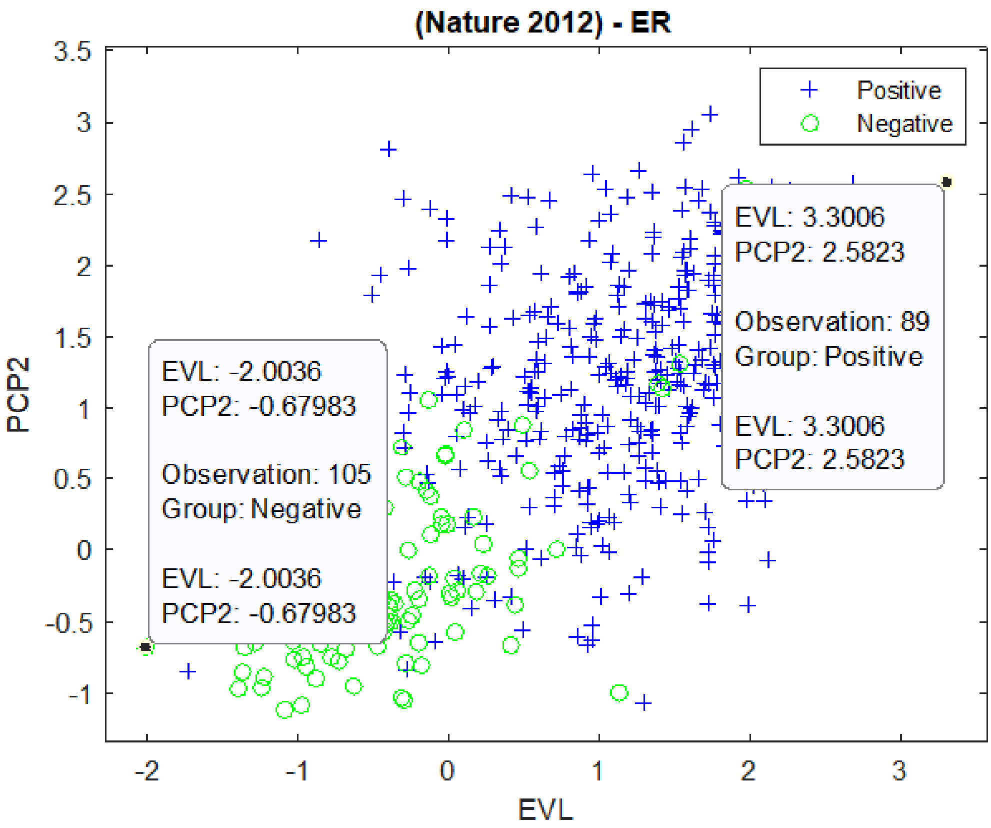


Figure 6.6: Scatter plot of EVL and PCP2 of (Nature 2012) dataset with ER groups, illustrating that the observation 89 from the ER+ group has high expression levels of (EVL, PCP2), which are $(3.3006, 2.5823)$ in comparison to the observation 105 from the ER- group, which has low expression levels of (EVL, PCP2), which are $(-2.0036, -0.67983)$.

patients with head and neck cancer, including PCP2. Therefore, this thesis is one of the first to show evidence of a positive correlation between PCP2 mRNA and ER, where highly expressed PCP2 contributes to ER positivity, as illustrated in Figure 6.6. Therefore, further investigations are required to indicate the clinical relevance of this gene to breast cancer.

- **{‘FSIP1’}**. FSIP1 is a cancer antigen expressed in the majority of breast cancer tissues and is associated with poor prognosis [190]. Several research studies have detected FSIP1 not only as a potential biomarker of breast cancer, but also as a potential therapeutic target. Liu et al. [196] identified

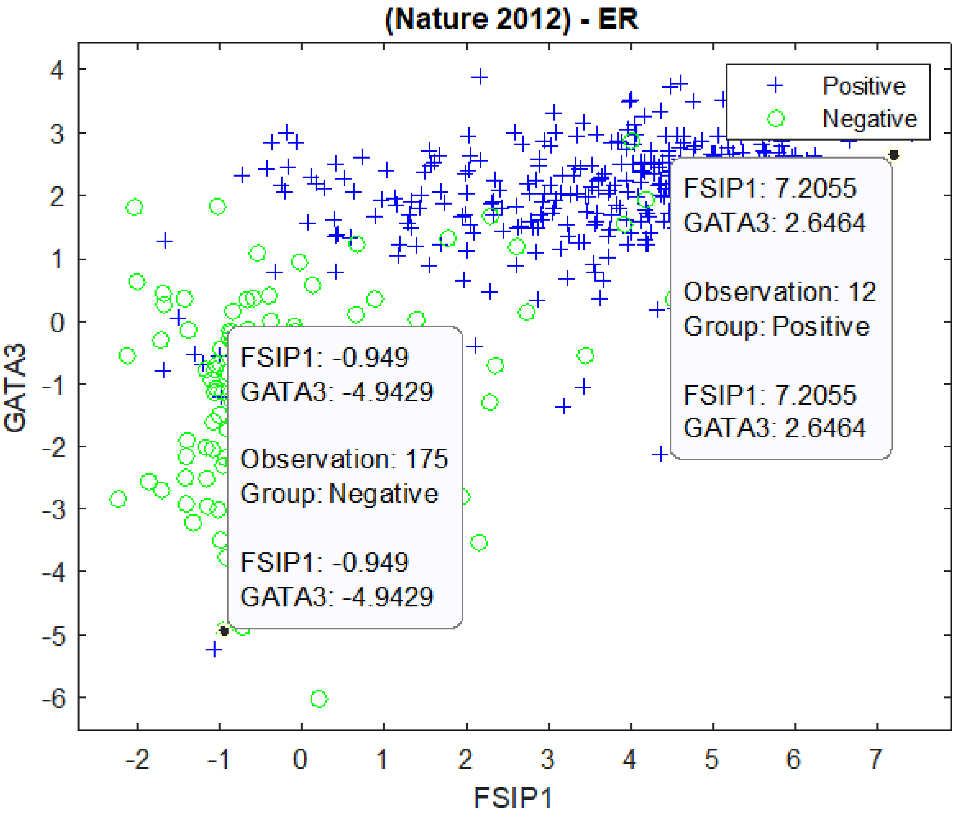


Figure 6.7: Scatter plot of FSIP1 and GATA3 of (Nature 2012) dataset with ER groups, illustrating that the observation 12 from the ER+ group has high expression levels of (FSIP1, GATA3), which are $(7.2055, 2.6464)$ in comparison to the observation 175 from the ER- group, which has low expression levels of (FSIP1, GATA3), which are $(-0.949, -4.9429)$.

FSIP1 as a signaling partner to HER2, and that FSIP inhibition reduces cell growth and invasiveness in HER2-positive breast cancer cells. A recent study [326] has shown that breast cancer cells and tissues consistently demonstrated elevated FSIP1 expressions, which correlated with poor overall survival. In this thesis, high levels of FSIP1 mRNA expression were detected mostly in patients with ER+ tumours, as explained in Figure 6.7, and across various groups of breast cancer samples that are collected from completely different studies. This provides strong evidence that this gene could act as a potential biomarker to human breast cancer.

- **{'GATA3'}**. GATA3 has been identified as one of the most frequently mutated genes in breast cancers. Takaku et al. [278] found that GATA3 zinc finger 2 mutations reprogram the breast cancer transcriptional network. Earlier in [279], they discussed the mutation of GATA3 in breast cancer, and the potential mechanisms by which mutation may lead to a growth advantage in cancer. The Significance and therapeutic potential of GATA3 expression and mutation in breast cancer were reviewed in [78]. In this thesis, GATA3 was found to be highly expressed in ER+ tumours compared to the samples from the ER- group, as presented in Figure 6.7, and across several independent datasets, which provides another evidence to the biological relevance of this gene to the positivity of the hormone receptor and breast cancer.

- **{'IGFALS'}**. Insulin like growth factor (IGF) has been implicated in the etiology and progression of breast and other cancers. A research study [71] found genetic variation in IGF1, IGF-1R, IGFALS, and IGFBP3 in breast cancer survival among Chinese women. The authors in [74] discovered that the lack of ALS proteins results in the disruption of the entire IGF circulating system. In our research, it has been shown that elevated levels of IGFALS mRNA expression were found to be in the patients with ER+ tumours compared to the ER- samples, as illustrated in Figure 6.8, and across variant independent subsets of breast cancer samples. Therefore, investigating the role IGFALS might play in ER+ breast cancer is necessary to provide insights and explanations to the biological and pathological processes of this complicated disease.

- **{'LRRC56'}**. Very limited information is available about this gene and its relevance to breast cancer and the hormone receptors. In this thesis, evidence of a positive correlation was detected between LRRC56 mRNA expression and high levels of ER, in which ER+ tumours are more likely characterised by high expression levels of LRRC56 compared to ER- samples, as shown in Figure 6.8, and over multiple genomic datasets. Therefore, targeting LRRC56 for further analysis could result in discovering its function in the underlying processes of ER+ breast cancer.

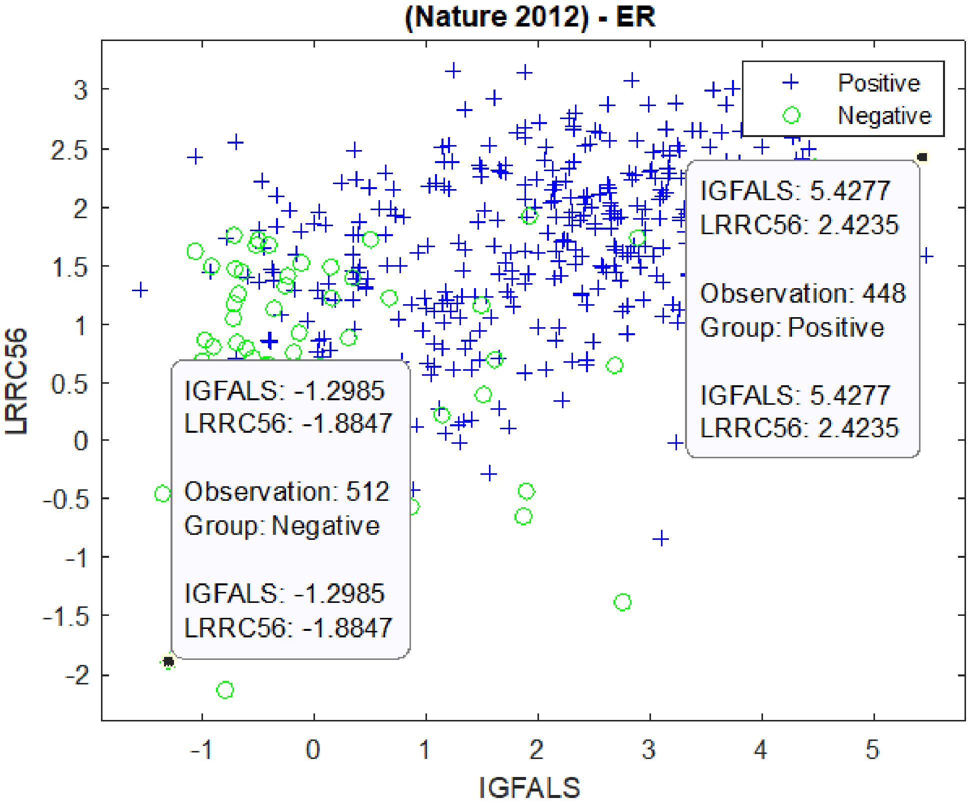


Figure 6.8: Scatter plot of IGFALS and LRRC56 of (Nature 2012) dataset with ER groups, illustrating that the observation 448 from the ER+ group has high expression levels of (IGFALS, LRRC56), which are $(5.4277, 2.4235)$ in comparison to the observation 512 from the ER- group, which has low expression levels of (IGFALS, LRRC56), which are $(-1.2985, -1.8847)$.

## 6.3 Discovered Biomarkers with HN Weight for ER

This section discusses the clinical relevance of the discovered biomarkers with HN weight with the oestrogen receptor recognised by the bioinformatics research in the literature. Furthermore, the identified association by this research between the expression levels of these biomarkers and ER will be explored to gain insight into the type of existent link. The relevance of the HN weighted mRNA markers to ER positivity will be investigated

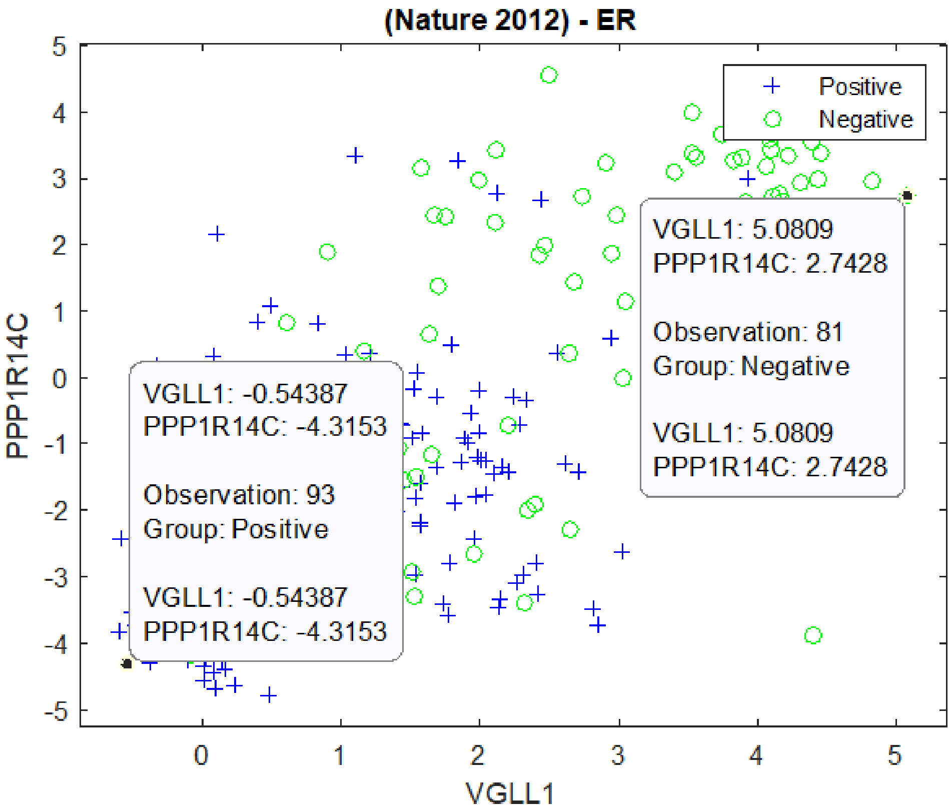


Figure 6.9: Scatter plot of VGLL1 and PPP1R14C of (Nature 2012) dataset with ER groups, illustrating that the observation 93 from the ER+ group has low expression levels of (VGLL1, PPP1R14C), which are $(-0.54387, -4.3153)$ in comparison to the observation 81 from the ER- group, which has high expression levels of (VGLL1, PPP1R14C), which are $(5.0809, 2.7428)$.

individually in the following points to provide conclusive evidence.

- **{'VGLL1'}**. Castilla et al. [47] found that VGLL1 expression is associated with a triple-negative basal-like phenotype in breast cancer. Li et al. [185] detected a set of genes including VGLL1 to be under-expressed in the ER-positive group and over-expressed in the ER-negative group. Recently, Segaert et al. [255] have identified 36 relevant genes to triple-negative breast cancer data including VGLL1. Lim et al. [186] recognised several a luminal progenitor signature including VGLL1 for basal tumour development. In this thesis, evidence of an inverse correlation was found between the expression patterns of VGLL1 and ER, as shown in Figure 6.9, and over a wide range of independently generated breast cancer samples. This negative association can be further researched to investigate the potential of the VGLL1 to be a biomarker to breast cancer, and particularly to triple-negative breast cancer as discussed above.

- **{'PPP1R14C'}**. As mentioned previously in VGLL1, Segaert et al. [255] have identified 36 genes, including PPP1R14C that are involved in triple-negative breast cancer. Castilla et al. [47] identified several genes, including PPP1R14C that are correlated with VGLL1 and miR-934 expression in a triple-negative basal-like phenotype in breast cancer. In this research, a negative correlation was detected between PPP1R14C mRNA expression and ER level, so that the declines in PPP1R14C expression values contribute to the positivity of ER, as illustrated in Figure 6.9. The inverse association between PPP1R14C and ER positivity was recognised over a wide range of breast cancer samples, therefore, the mechanism underlying that relationship should be explored further.

- **{'PROM1'}**. According to the NCBI website[1], *"Expression of this gene is associated with several types of cancer"*. The researchers in [321] found that that CD133 mRNA can be a suitable prognostic marker for human breast cancer. According to the research study [293], CD133 could act as a marker of breast cancer cells and stem cells. Recently, Zhang et al. [334] have found several differentially expressed genes including PROM1 that may play important roles in the process of bone metastasis from breast cancer. As mentioned previously in VGLL1 and PPP1R14C, Castilla et al. [47] identified several genes including PROM1 that are most correlated with VGLL1 and miR-934 expression in a triple-negative basal-like phenotype in breast cancer. As seen in VGLL1, Lim et al. [186] detected several a luminal progenitor signature including PROM1 for basal tumour development. Moreover, some studies have detected the correlation between PROM1 (CD133) and VGLL1 such as [30,156]. In this thesis, it has been shown that PROM1 mRNA expression is negatively correlated with ER+ tumours, as presented

[1] https://www.ncbi.nlm.nih.gov/gene/8842

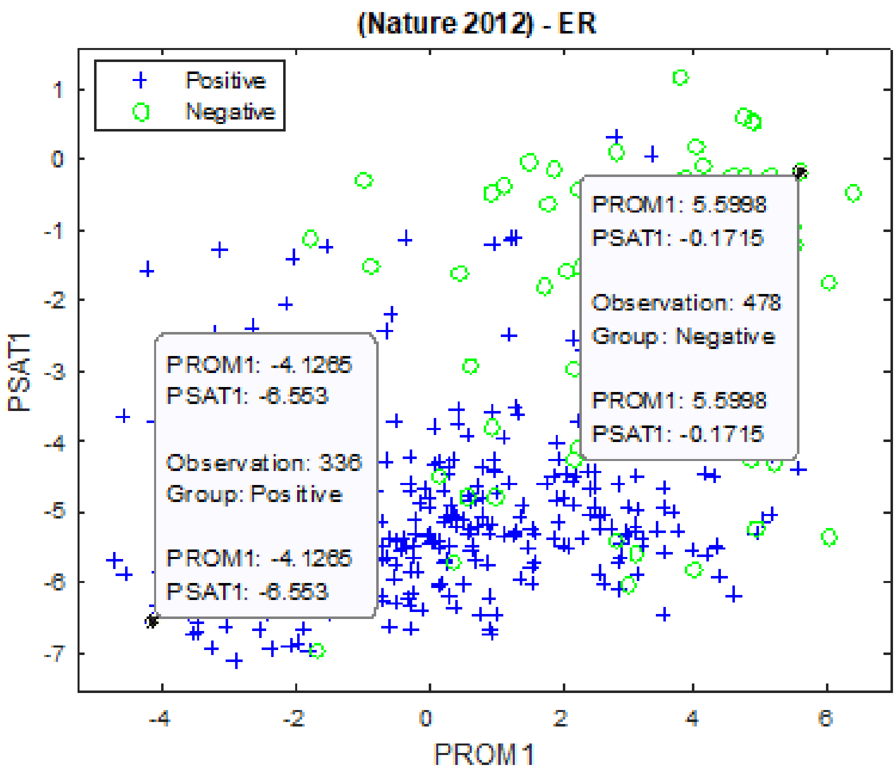


Figure 6.10: Scatter plot of PROM1 and PSAT1 of (Nature 2012) dataset with ER groups, illustrating that the observation 336 from the ER+ group has low expression levels of (PROM1, PSAT1), which are $(-4.1265, -6.553)$ in comparison to the observation 478 from the ER- group, which has high expression levels of (PROM1, PSAT1), which are $(5.5998, -0.1715)$.

in Figure 6.10. This inverse association was validated over a large number of variations in breast samples of (Nature 2012), (Cell 2015), (Provisional), (METABRIC) as well as the integrated datasets NCP1, NCP2, and NCP3. The obtained findings reveal that highly expressed PROM1 contribute to the phenotype associated with ER positivity, thus the examination of that potential relationship is imperative.

- **{‘PSAT1’}**. Possemato et al. [235] found that the inhibition of PSAT1 significantly decreased the proliferation of ER-negative breast cancer cells but not ER-positive breast cancer cells. Gao et al. [103] revealed that the expression of PSAT1 was significantly upregulated in ER-negative breast cancers compared with ER-positive breast cancers, and they added that PSAT1 up-regulation was correlated with tumour development and poor prognosis. PSAT1 hyper-methylation and mRNA levels were found in [206] to be significantly associated with the outcome to tamoxifen treatment in recurrent disease. In this research, evidence of a negative correlation was found between PSAT1 and ER, so that the declines in the expression values of PSAT1 lead to high ER levels, as explained in Figure 6.10. This potential association was validated across a wide range of breast cancer samples of (Nature 2012), (Cell 2015), (Provisional), and (METABRIC), as well as the integrated datasets with ER groups. The clinical relevance of PSAT1 mRNA expression to the positivity of ER requires further studying to declare whether this gene can act as a clinical indicator for breast cancer and the hormone receptor.

- **{'B3GNT5'}**. According to [236], glycolipids may play an important role in carcinogenesis of breast tumours that are shown by the association of B3GNT5 and UGCG genes to patient survival. The authors in [294] identified 12 markers, including B3GNT5 for detection of primary breast cancer. As mentioned previously in VGLL1 and PPP1R14C, Segaert et al. [255] have identified 36 relevant genes in triple-negative breast cancer data, including B3GNT5. Highly significant correlations were observed in [162] between cyclooxygenase 2 (COX2) with eight tumour-promoting genes, including B3GNT5, which are known to effectively increase the inflammogenesis of breast cancer. B3GNT5 levels were found in [344] to be higher in both ER-negative and PR-negative tumours. This research detects an inverse correlation between B3GNT5 mRNA expression and ER levels, in which B3GNT5 is highly expressed in the ER-negatives compared to the ER+ patient, as shown in Figure 6.11, and across a broad range of breast cancer samples. The mechanism underlying the relationship between B3GNT5 mRNA and ER levels can be further studied by domain experts to answer different biological questions of interest.

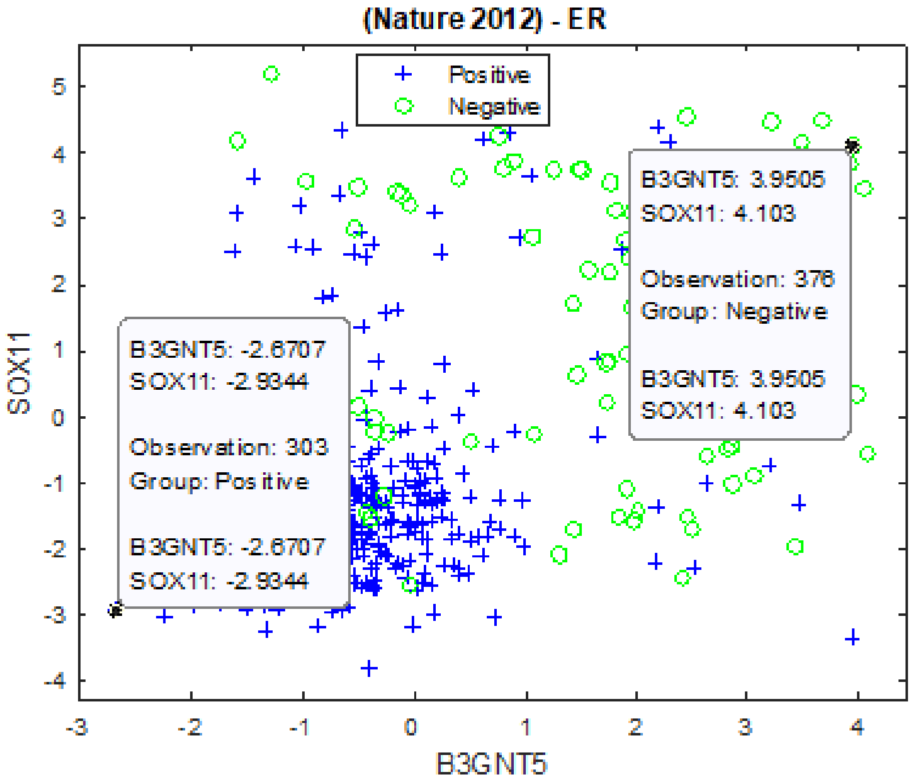


Figure 6.11: Scatter plot of B3GNT5 and SOX11 of (Nature 2012) dataset with ER groups, illustrating that the observation 303 from the ER+ group has low expression levels of (B3GNT5, SOX11), which are $(-2.6707, -2.9344)$ in comparison to the observation 378 from the ER- group, which has high expression levels of (B3GNT5, SOX11), which are $(3.9505, 4.103)$.

- **{'SOX11'}**. Shepherd et al. [257] found that SOX11 is a critical regulator of multiple phenotypes of Basal-like breast cancers such as growth, migration, and invasion. The research study [191] identified that Nuclear SOX11 was observed in (36.2%) and cytoplasmic SOX11 in (44.8%) of breast cancer samples. Recently, Wang et al. [308] have found that SOX11 expression was directly associated with breast cancer stem cell populations. In this thesis, it has been shown that SOX11 is highly expressed in ER- samples compared to the ER-positives, as illustrated in Figure 6.11, and across a large number of independent variations in the breast cancer samples. This gene exhibits great distinctions between the ER+ and ER- samples, thus investigating the potential of SOX11 to be a molecular marker to breast cancer is required to advance the move towards precision medicine.

- **{'KRT16'}**. The study [151] stated that breast tumours can alter the expression of certain keratins during the process of metastatic development and an association was found between KRT16 expression and shorter relapse-free survival in metastatic breast cancer. The study presented in [159] detected multiple autoimmune response signature associated with the development of triple negative breast cancer, involving KRT16. In this thesis, KRT16 was found to be negatively associated with ER levels, as illustrated in Figure 6.12, and over a wide range of independent breast cancer samples. This inverse association needs further researching to understand the biological role of the keratin 16 in human breast cancers and allow more innovative findings to be discovered.

- **{'TMEM40'}**. Little is known in the literature about this gene, particularly, its relevance to breast cancer and ER. Recently, a research study [333] has stated that TMEM40 gene encodes a protein of 233 amino acids and is located on chromosome 3p25.2. Moreover, it has found that high expression of TMEM40 contributes to progressive features of tongue squamous cell carcinoma. In another recent study [335], the role of TMEM40 in the tumorigenesis of bladder cancer has been identified and found that it was upregulated in bladder cancer tissues and cell lines, compared with their normal counterparts. The clinical relevance of TMEM40 mRNA expression to breast cancer and oestrogen receptor is not clear in the literature and evidence of a negative correlation was recognised and validated in this research between the expression pattern of TMEM40 and ER levels, as shown in Figure 6.12, and over multiple variant subsets of breast samples. These findings can motivate conducting further investigations to determine the mechanism underlying that association.

- **{'BBOX1'}**. In the literature, very limited information is available about BBOX1, especially its relevance to breast cancer and ER. The previously

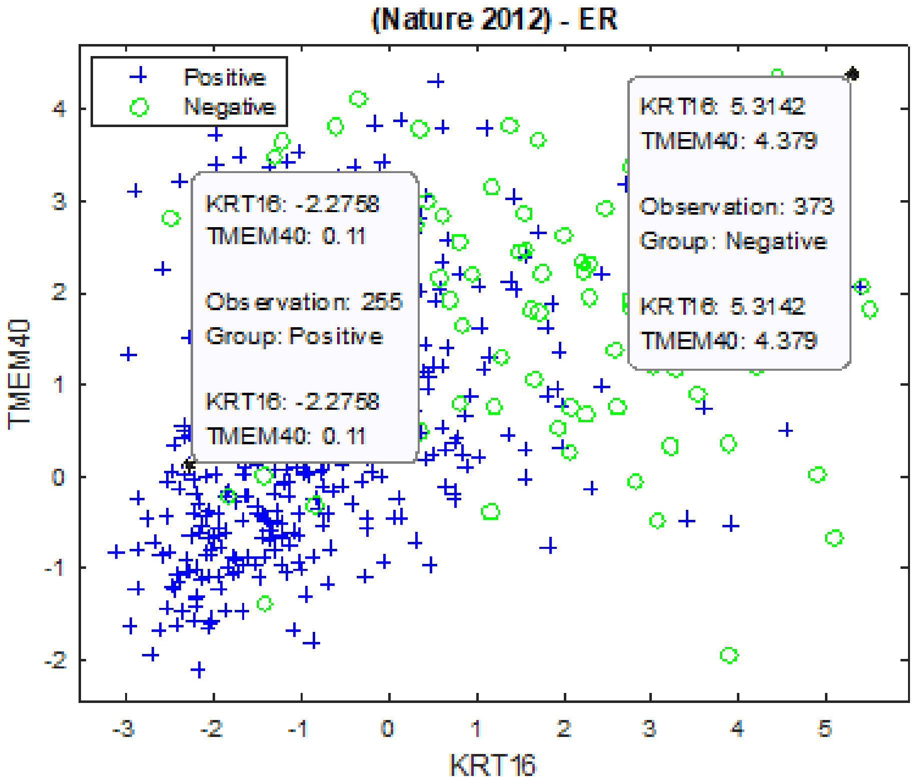


Figure 6.12: Scatter plot of KRT16 and TMEM40 of (Nature 2012) datasets with ER groups, illustrating that the observation 255 from the ER+ group has low expression levels of (KRT16, TMEM40), which are $(-2.2758, 0.11)$ in comparison to the observation 373 from the ER- group, which has high expression levels of (KRT16, TMEM40), which are $(5.3142, 4.379)$.

discussed study in VGLL1, PPP1R14C, and PROM1 [47], that mentioned the association of several genes, including BBOX1 to a triple-negative basal-like phenotype in breast cancer. This thesis is the first to report evidence of an inverse correlation to be found between BBOX1 mRNA and ER levels, in which the declines in BBOX1 expression values contribute to ER positivity, as introduced in Figure 6.13, and across a wide range of breast cancer samples. The behavior of BBOX1 differs significantly between ER+ tumours and ER-negative ones, which makes it a candidate indicator to high levels of ER. Therefore, further studies are required to examine the

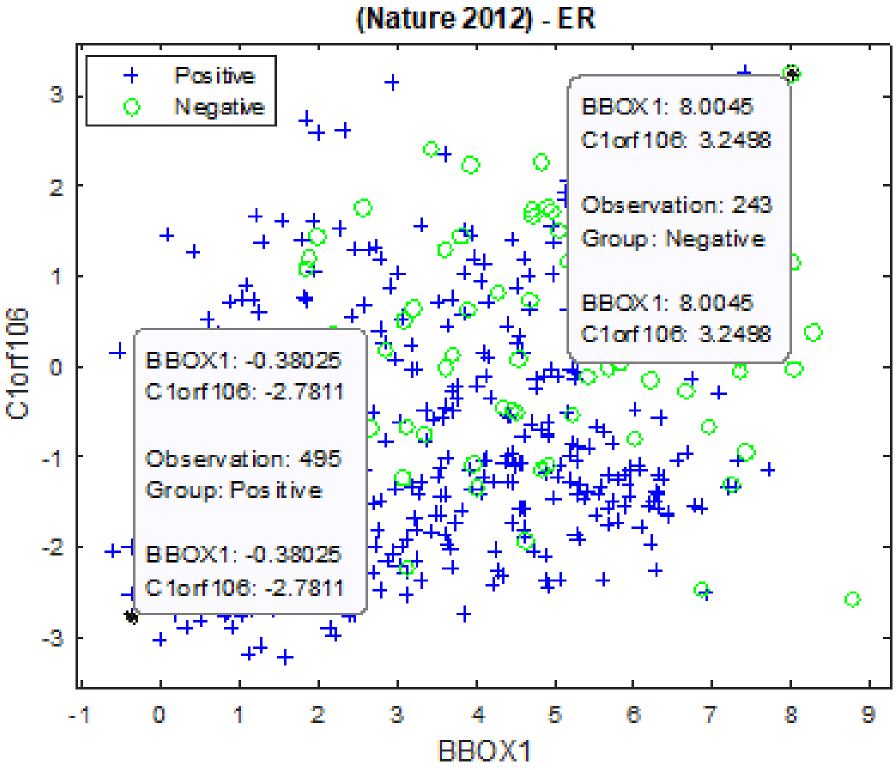


Figure 6.13: Scatter plot of BBOX1 and C1orf106 of (Nature 2012) dataset with ER groups, illustrating that the observation 495 from the ER+ group has low expression levels of (BBOX1, C1orf106), which are $(-0.38025, -2.7811)$ in comparison to the observation 243 from the ER- group, which has high expression levels of (BBOX1, C1orf106), which are $(8.0045, 3.2498)$.

clinical relevance of this gene to breast cancer.

- **{'C1orf106'}**. Recently, Yang et al. [328] have detected 61 differential expressed genes, including C1orf106 for basal-like breast cancer. Lemetre in the PhD thesis [179] detected C1orf106 as one of the relevant genes for breast cancer. In this thesis, a negative correlation was found between C1orf106 mRNA expression and ER levels, as presented in Figure 6.13. This inverse association was verified across the (Nature 2012), (Cell 2015), (Provisional), (METABRIC), NCP1, NCP2, and NCP3. Various biological questions could be addressed by examining the biological role of the

C1orf106 gene in the underlying process of the hormone receptor, ER.

- **{'DKK1'}**. It has been shown that DKK1 is involved in a variety of cancers. The authors in [157] claimed that Dkk1 might provide insights into the continued development of novel comprehensive and therapeutic strategies for breast cancer and its bone metastases. Recently, a research study [221] found that DKK1 over-expression dramatically inhibits breast cancer cell migration and invasion, where knockdown of DKK1 promotes migration and invasion of breast cancer cells. Forget et al. [99] identified DKK1 as a potential prognostic and diagnostic marker for cohorts of breast cancer patients with poor prognosis. In this thesis, an inverse correlation between DKK1 mRNA expression and the positivity of ER was discovered, in which low expression levels of DKK1 can be found mostly in ER+ tumours compared to ER- samples, as shown in Figure 6.14, and over a wide range of independently generated breast cancer samples. Therefore, DKK1 can be investigated further to be utilised for developing various diagnosis and prognosis systems for detecting and monitoring breast cancer.

- **{'KRT81'}**. According to [37], KRT81 is expressed in the human breast cancer cell line SKBR3, and in metastatic lymph nodes of breast carcinomas according to [290]. A study [218] found that KRT81 is expressed in normal breast epithelial cells and breast cancer cells, and suggested that KRT81 contributes to the migration and invasion of breast cancer cells. In this thesis, KRT81 was found to be negatively associated with the positivity of ER, in which high expression levels of KRT81 can be found mostly in the ER- samples compared to the ER-positives, as illustrated in Figure 6.14. This inverse association was detected across a wide range of independent breast cancer samples, thus targeting this gene for further researching could contribute to understanding the role KRT81 might play in the heterogeneity of breast cancer and the positivity of ER.

- **{'RARRES1'}**. Among five breast cancer subtypes, the authors in [59] found that RARRES1 expression is greatest in basal-like TNBCs, and they revealed that RARRES1 is a tumour suppressor in TNBC. Coyle et al. [60]

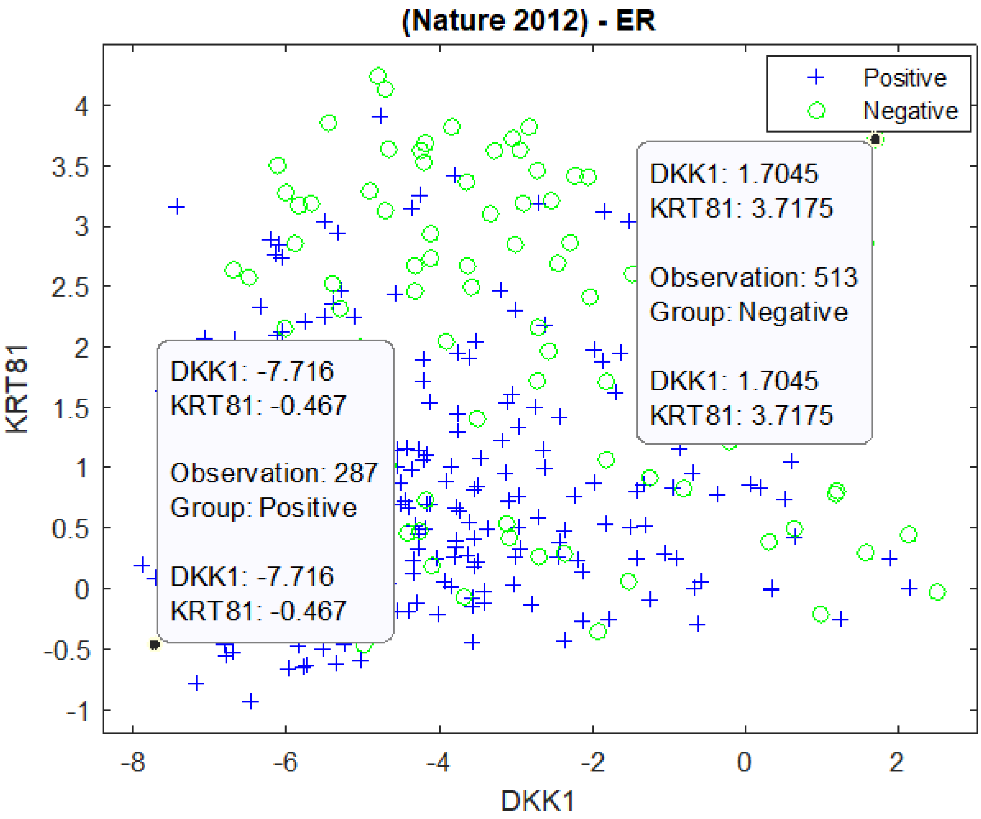


Figure 6.14: Scatter plot of DKK1 and KRT81 of (Nature 2012) dataset with ER groups, illustrating that the observation 287 from the ER+ group has low expression levels of (DKK1, KRT81), which are $(-7.716, -0.467)$ in comparison to the observation 513 from the ER- group, which has high expression levels of (DKK1, KRT81), which are $(1.7045, 3.7175)$.

identified RARRES1 as a tumour suppressor in triple-negative breast cancer cell lines. In this thesis, it has been shown that RARRES1 exhibits low expression levels for the patients with high ER levels in comparison to the ER-negative samples, as presented in Figure 6.15, and across multiple datasets that are collected from different studies. As a result, it is relevant to examining the potential of RARRES1 to be a target for breast cancer and ER positivity.

- **{'S100A8'}**. S100A8/A9 was detected to be associated with ER loss in breast cancer in [15]. Zhong et al. [339] found that S100A8 may be asso-

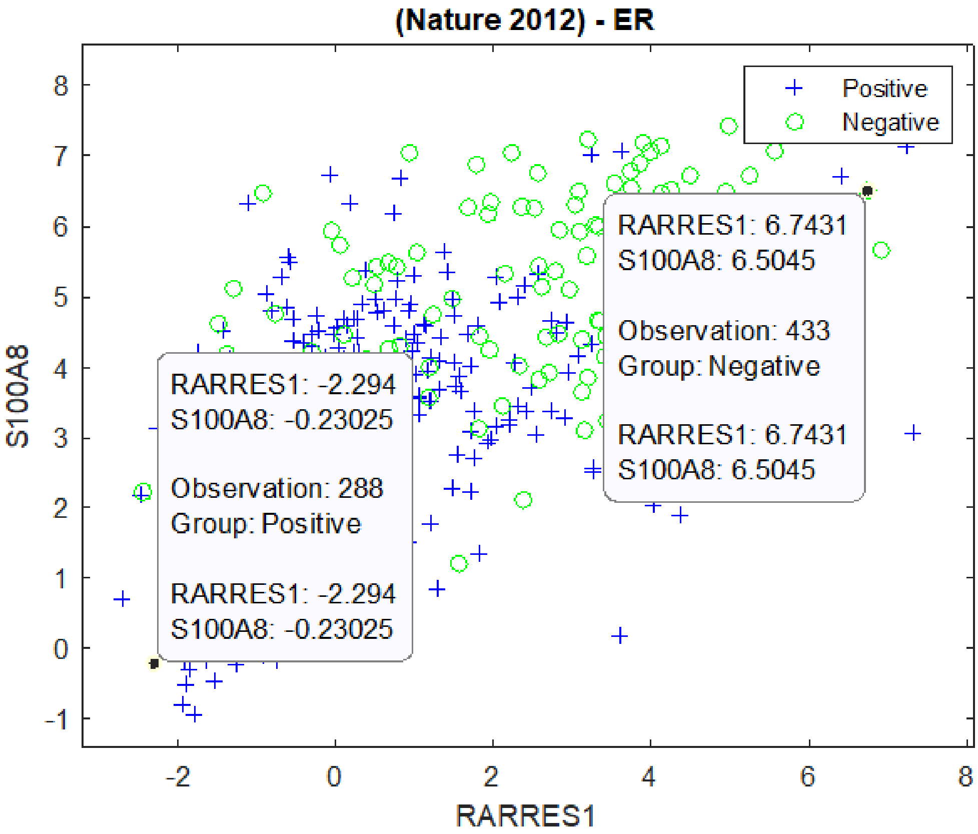


Figure 6.15: Scatter plot of RARRES1 and S100A8 of (Nature 2012) dataset with ER groups, illustrating that the observation 288 from the ER+ group has low expression levels of (RARRES1, S100A8), which are $(-2.294, -0.23025)$ in comparison to the observation 433 from the ER- group, which has high expression levels of (RARRES1, S100A8), which are $(6.7431, 6.5045)$.

ciated with lymph nodes metastasis of breast cancer and be a marker for progression of breast cancer. The authors in [307] detected that ER- and triple-negative breast cancer samples has significantly higher expression of S100A8 than samples with other subtypes, thus they suggested S100A8 as a potential biomarker for relapse in breast cancer patients. In this thesis, evidence of a negative correlation was found between the expression patterns of S100A8 and ER, as clarified in Figure 6.15, and over different groups of independent breast cancer samples. Thus, considering this gene in future studies could contribute to understand its role in breast cancer and triple-negative breast cancer as discussed above.

- **{'S100A9'}**. As mentioned in S100A8, S100A9 was detected to be associated with ER loss in breast cancer in [15]. The researchers in [182] identified S100A9 as a novel OM-regulated gene and indicated its involvement in the growth regulation of breast cancer cells. S100A9 has been recognised in [27] as a novel therapeutic target for patients with ERPgR breast cancers. This thesis has shown that the patients with ER+ tumours exhibit low expression levels of S100A9 mRNA compared to the ER- samples, as explained in Figure 6.16, and across a wide range of breast cancer samples. As a result, S100A9 should be investigated further to be employed as a potential target for ER- patients.

- **{'TRPV6'}**. It has been shown that TRPV6 is involved in colon cancer, breast cancer, prostate cancer, parathyroid cancer and thyroid cancer [265]. TRPV6 was identified in [232] as a novel therapeutic strategy for the treatment of ER- breast cancers. A very recent study [269] revealed TRPV6 as a promising drug target in a variety of cancers, including breast, ovarian, prostate and pancreatic tissues. Herein, we demonstrated a negative correlation between the expression levels of ER and TRPV6, in which the drops in TRPV6 mRNA expression could contribute to the positivity of ER, as shown in Figure 6.16, and over several independent genomic datasets. Therefore, TRPV6 can be further researched to indicate its contributions to breast cancer.

- **{'HRASLS'}**. Mardine et al. [205] discussed the role of each of the HRASLS enzymes in cancer, as well as their biochemical function, and then they concluded that reduced expression of these enzymes can be found mostly in cancer cells. This thesis reveals a negative association between the expression patterns of HRASLS mRNA and ER positivity, in which highly expressed HRASLS can be found mostly in ER-negative samples than ER-positives, as clarified in Figure 6.17, and across a wide range of breast cancer samples that are generated independently. Further investigations are required to indicate its function in breast cancer.

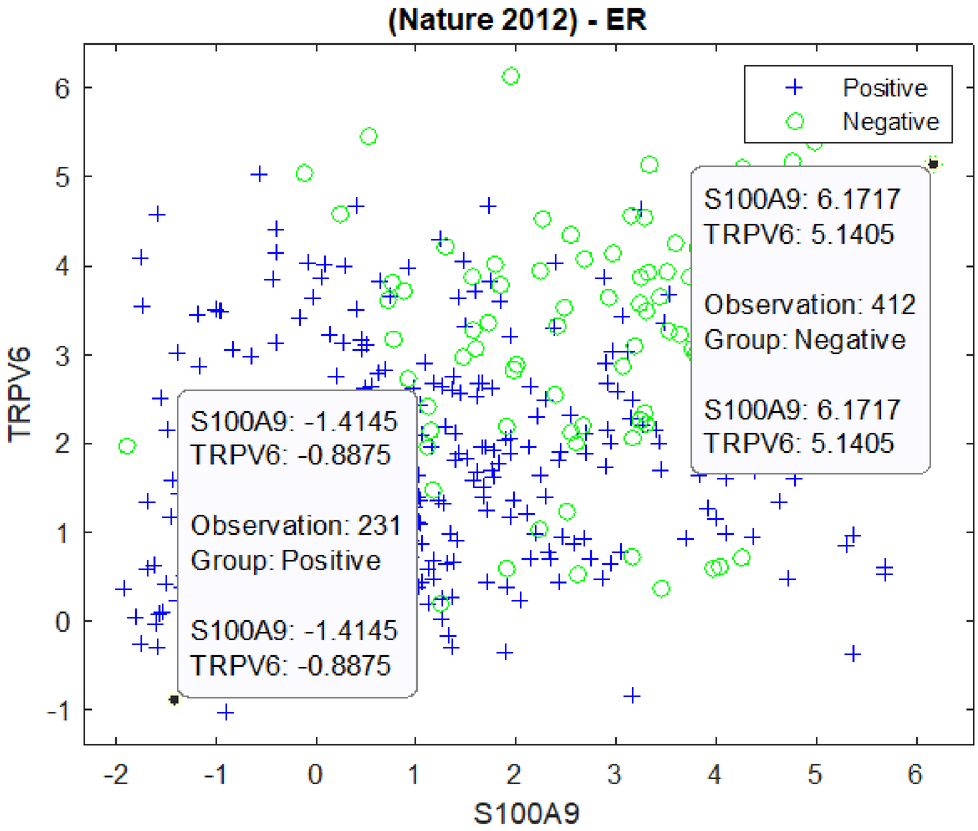


Figure 6.16: Scatter plot of S100A9 and TRPV6 of (Nature 2012) dataset with ER groups, illustrating that the observation 231 from the ER+ group has low expression levels of (S100A9, TRPV6), which are $(-1.4145, -0.8875)$ in comparison to the observation 412 from the ER- group, which has high expression levels of (S100A9, TRPV6), which are $(6.1717, 5.1405)$.

- **{'PPP1R1A'}**. The expression of PPP1R1A in lung, colorectal, and gastric cancer cell lines was different from that of the normal tissues [280]. This thesis is first to report that PPP1R1A mRNA expression is negatively associated with ER+ breast cancer samples, in which the drops in the expression levels of this gene could lead to high levels of ER expression, as shown in Figure 6.17, and across independent cancer genomic datasets. Considering PPP1R1A in future studies could contribute to revealing its biological role in breast cancer.

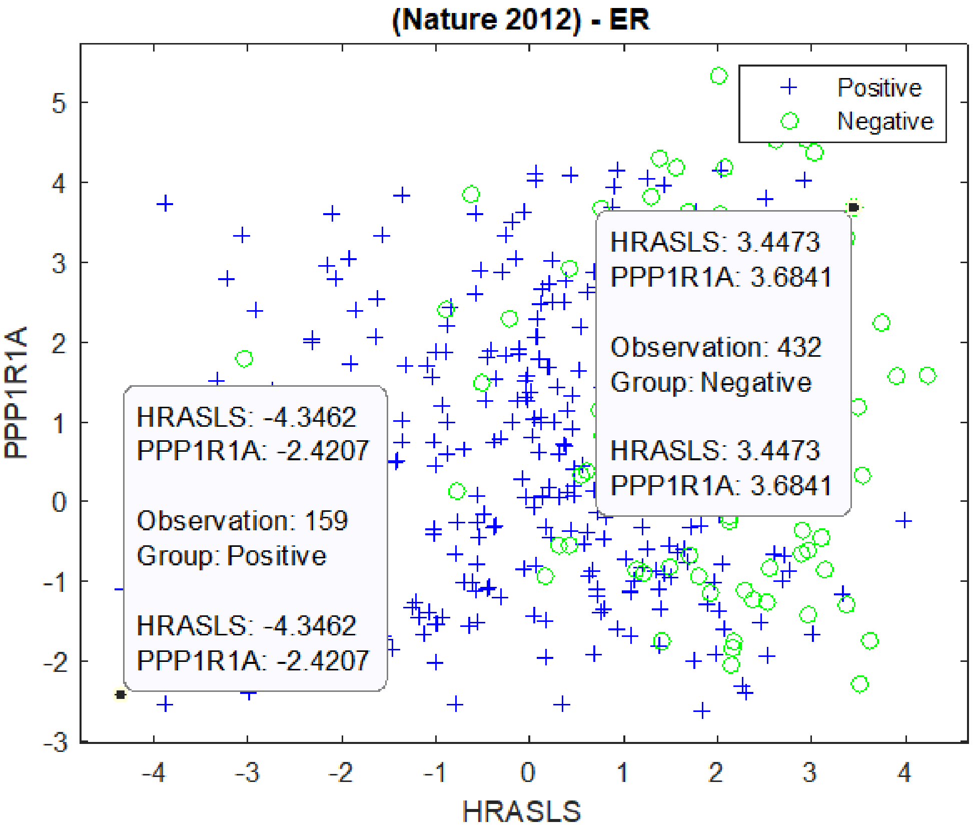


Figure 6.17: Scatter plot of HRASLS and PPP1R1A of (Nature 2012) dataset with ER groups, illustrating that the observation 159 from the ER+ group has low expression levels of (HRASLS, PPP1R1A), which are $(-4.3462, -2.4207)$ in comparison to the observation 432 from the ER- group, which has high expression levels of (HRASLS, PPP1R1A), which are $(3.4473, 3.6841)$.

## 6.4 Discovered Biomarkers with HP Weight for PR

The relevance of the discovered biomarkers with HP weight to the progesterone receptor is discussed in this section with respect to current state-of-the-art bioinformatics research found in the literature. The recognised relationship in our research between each mRNA marker and PR will be discussed to provide conclusive evidence about the type of existent associ-

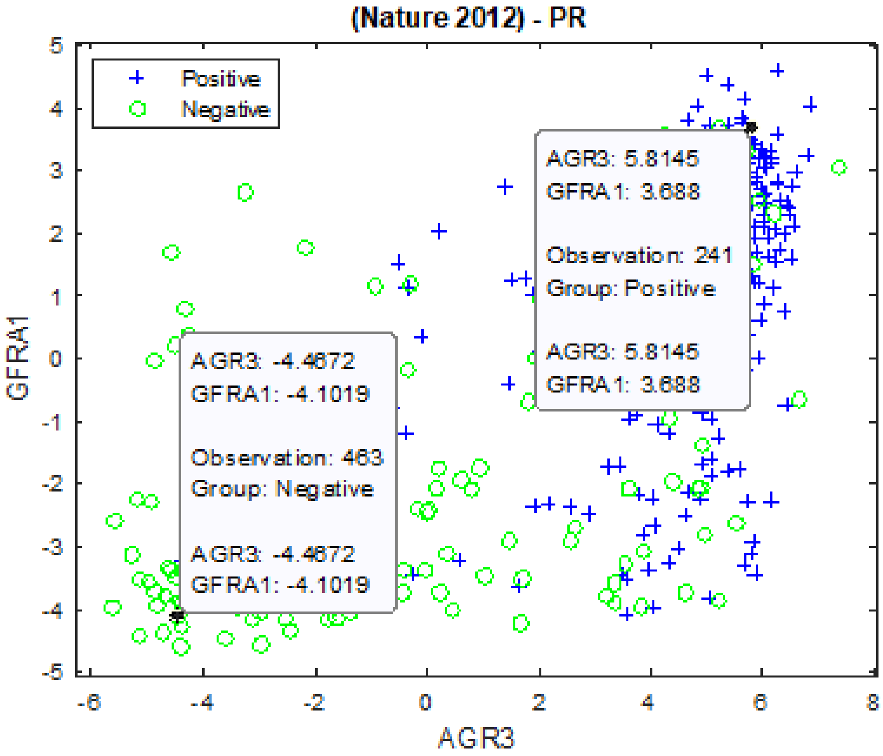


Figure 6.18: Scatter plot of AGR3 and GFRA1 of (Nature 2012) dataset with PR groups, illustrating that the observation 241 from the PR+ group has high expression levels of (AGR3, GFRA1), which are $(5.8145, 3.688)$ in comparison to the observation 463 from the PR- group, which has low expression levels of (AGR3, GFRA1), which are $(-4.4672, -4.1019)$.

ation.

- **{'AGR3'}**. In this thesis, evidence of a positive correlation was detected between the AGR3 mRNA expression and PR levels, so that the gains in AGR3 lead to high PR levels, as shown in Figure 6.18, and across a broad range of breast cancer samples. The relevance of AGR3 to breast cancer and the positivity of the hormone receptors in terms of what has been revealed in the literature by bioinformatics analysis research was discussed earlier in Section 6.2. Biomarker discovery in bodily fluids can advance the move towards a new generation of diagnosis and prognosis models.

- **{'GFRA1'}**. A potential relationship between GFRA1 mRNA expression and PR positivity was recognised in this research, as illustrated in Figure 6.18, and over a large number of variant breast cancer samples. Our findings reveal GFRA1 mRNA as a strong candidate biomarker for breast cancer that is positively correlated to the hormone receptor ER and PR, thus further investigations from biomedical experts are necessitated to indicate the mechanism underlying the association.

- **{'SCUBE2'}**. SCUBE2 mRNA was found to be positively correlated to PR expression, as presented in Figure 6.19, and over the independent variations in breast cancer samples. The positive relationship between the expression patterns of SCUBE2 mRNA and the hormone receptors ER and PR recognised in this thesis necessities conducting further experiments to determine the potential of this gene to be a biomarker for human breast cancers.

- **{'SIAH2'}**. Our omics data analysis study detected a positive correlation between SIAH2 mRNA and PR expression levels, as explained in Figure 6.19, and across a wide range of breast cancer samples. Furthermore, SIAH2 was also detected in this research to have a positive correlation with ER, and this is discussed in Section 6.2. SIAH2 mRNA and its relevancy to the hormone receptors ER and PR can be further examined to indicate its potential in the early detection and management of breast cancers.

- **{'FGD3'}**. FGD3 was identified by the attractor metagene methodology [53] applied to the 2,000 breast cancer sample of METABRIC dataset [62] and won the Sage Bionetworks-DREAM Breast Cancer Prognosis Challenge. Willis et al. [318] found that FGD3 mRNA is regulated by ESR1 and it is an important clinical biomarker. FGD3 was detected in [273] to be a potential prognostic biomarker for breast cancer. Yao et al. [330] identified nine genes, including FGD3 for breast cancer grading and staging. In this thesis, FGD3 was detected to be positively correlated to PR expression levels, as illustrated in Figure 6.20, and across a wide range of breast cancer samples, thus targeting these mRNAs in further investigations is required

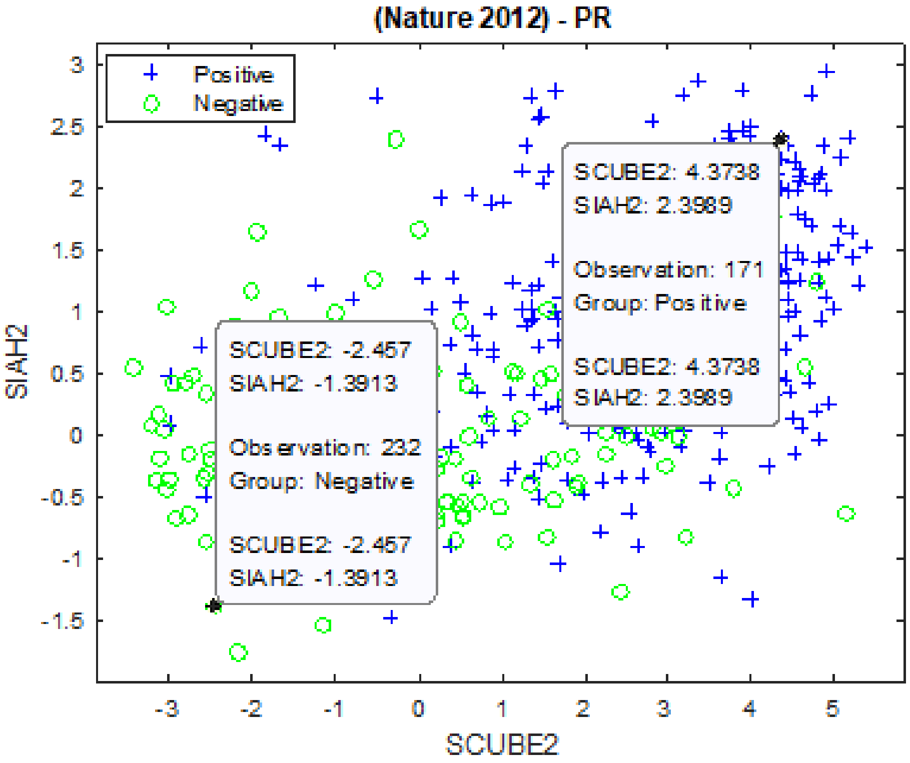


Figure 6.19: Scatter plot of SCUBE2 and SIAH2 of (Nature 2012) dataset with PR groups, illustrating that the observation 171 from the PR+ group has high expression levels of (SCUBE2, SIAH2), which are $(4.3738, 2.3989)$ in comparison to the observation 232 from the PR- group, which has low expression levels of (SCUBE2, SIAH2), which are $(-2.457, -1.3913)$.

to approach personalised and precision medicine for breast cancer.

- **{'SUSD3'}**. As mentioned previously in FGD3, SUSD3 was also identified by the attractor metagene methodology [53] as a potential biomarker for breast cancer. SUSD3 was found in [216] to be a significantly discriminative gene, and a novel promoter of estrogen-dependent cell proliferation. Zhao et al. [336] found that the expression of Insulin-like Growth Factor-I Receptor (IGF-IR) and SUSD3 may be associated with the occurrence and progression of breast cancer. In this thesis, it has been shown that SUSD3 exhibits a positive association with PR expression levels so that PR+ tumours are

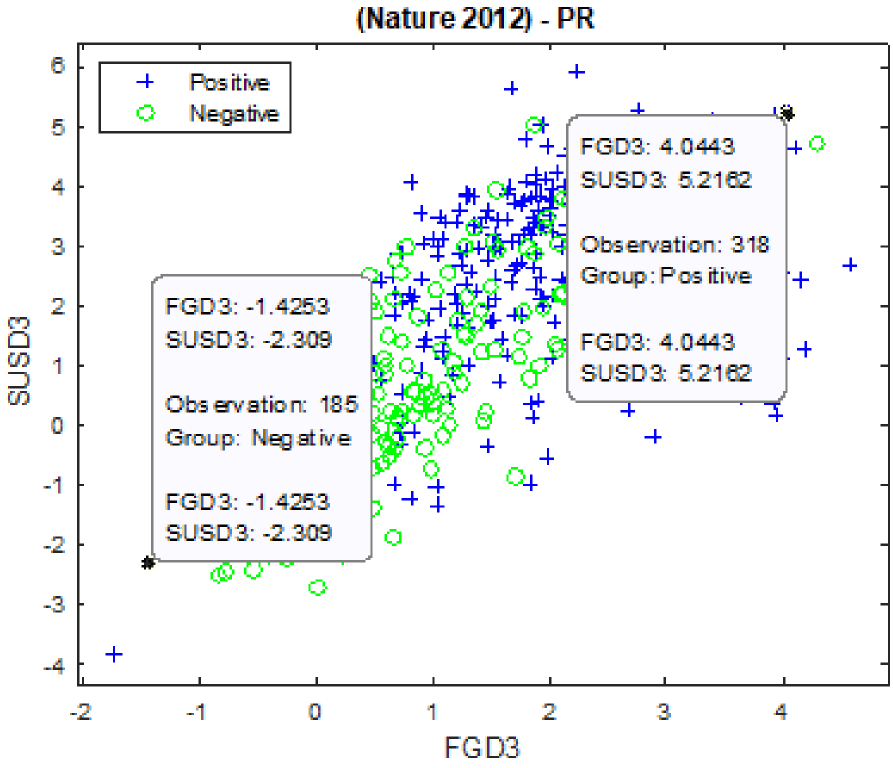


Figure 6.20: Scatter plot of FGD3 and SUSD3 of (Nature 2012) dataset with PR groups, illustrating that the observation 318 from the PR+ group has high expression levels of (FGD3, SUSD3), which are $(4.0443, 5.2162)$ in comparison to the observation 185 from the PR- group, which has low expression levels of (FGD3, SUSD3), which are $(-1.4253, -2.309)$.

characterised by a high expression levels of SUSD3 mRNA compared to PR- samples, as clarified in Figure 6.20, and across a wide range of breast cancer samples. The obtained findings provide strong evidence that SUSD3 could act as potential diagnostic and prognostic biomarkers to breast cancer and PR, thus targeting this gene in further investigations is required to understand the role it might play in PR+ breast cancer.

- **{'GRPR'}**. According To NCBI website[1], GRPR *"regulates numerous functions of the gastrointestinal and central nervous systems. The recep-*

[1] https://www.ncbi.nlm.nih.gov/gene/2925

*tor is aberrantly expressed in numerous cancers such as lung, colon, and prostate"*. Morgat et al. [213] found that GRPR is over-expressed in 83% of ER-positive tumours, and this over-expression was also found in lymph node metastases in 94.6% of cases. Results of their recent study [214] on breast cancer samples have also shown that GRP-R targeting is highly relevant in breast cancer, specifically in ER-positive tumours. High GRPR mRNA levels were detected in [65] to be more frequent in samples with positivity for ER mRNA ESR1, or PR mRNA. Dalm et al. [64] recognised GRPR to be over-expressed on primary breast cancer and thus, they have investigated the possibility of integrating it with other candidate genes for receptor-mediated nuclear imaging and therapy. GRPR mRNA was discovered in this thesis to be positively related to PR, as shown in Figure 6.21, and across independently generated breast cancer samples. Therefore, the mechanism underlying that association can be further studied to allow more innovative findings.

- **{'PGLYRP2'}**. Shanle et al. [256] identified several ER target genes, including PGLYRP2 in triple negative breast cancer cells. On the other hand, the research study [13] recognised that the absence of PGLYRP2 leads to alterations in the expression of the autism risk gene c-Met, and sex-dependent changes in social behavior, similar to mice with manipulated microbiota. Herein, the expression of PGLYRP2 mRNA was detected to be positively associated with the PR levels, as presented in Figure 6.21, and across independent subsets of breast cancer samples. Thus, further studies can be conducted to determine the role of PGLYRP2 in the biological or pathological process of breast cancers.

- **{'GREB1'}**. According to NCBI website[1], *"this gene is an estrogen-responsive gene that is an early response gene in the estrogen receptor-regulated pathway. It is thought to play an important role in hormone-responsive tissues and cancer"*. The researchers in [240] found GREB 1 to be critically involved in the estrogen induced growth of breast cancer cells and has the potential of being a clinical marker for response to endocrine ther-

[1] https://www.ncbi.nlm.nih.gov/gene/9687

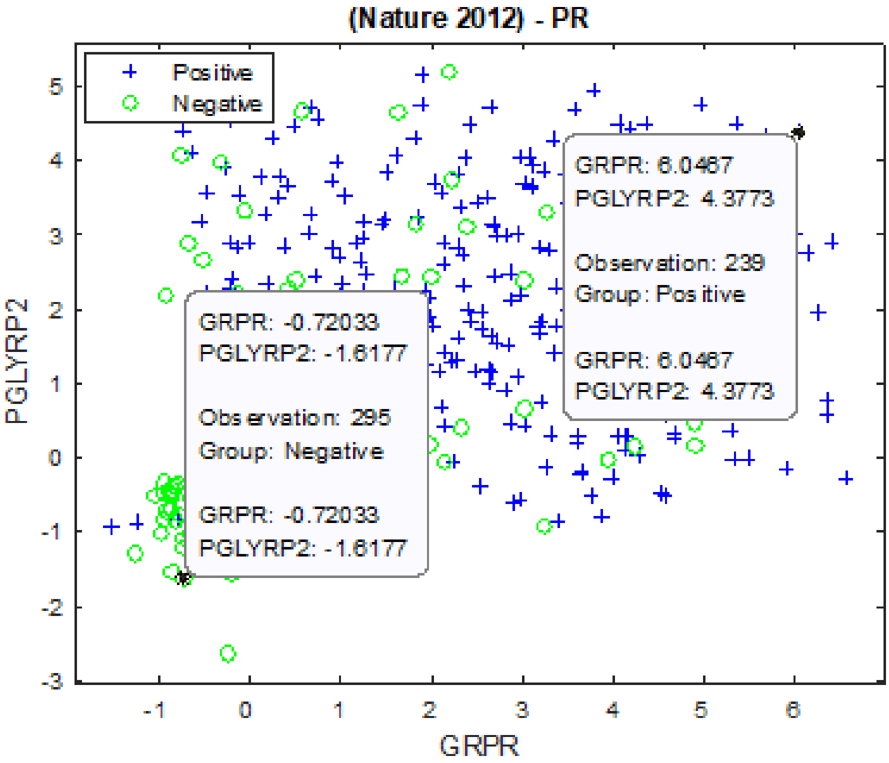


Figure 6.21: Scatter plot of GRPR and PGLYRP2 of (Nature 2012) dataset with PR groups, illustrating that the observation 239 from the PR+ group has high expression levels of (GRPR, PGLYRP2), which are (6.0467, 4.3773) in comparison to the observation 295 from the PR- group, which has low expression levels of (GRPR, PGLYRP2), which are (−0.72033, −1.16177).

apy as well as a potential therapeutic target. Camden et al. [147] found that GREB1 is a novel progesterone-responsive gene required for progesterone-driven human endometrial stromal cell (HESC) decidualization. The recent review study [52] has examined evidence that GREB1 participates in several hormone-dependent cancers and could be targeted to treat these cancers and concluded that the hormone-responsive gene GREB1 plays important roles in the initiation and progression of some sex hormone-driven cancers. Similar findings have also been shown recently by the study [133], which detected GREB1 to be an estrogen receptor-regulated tumour promoter that

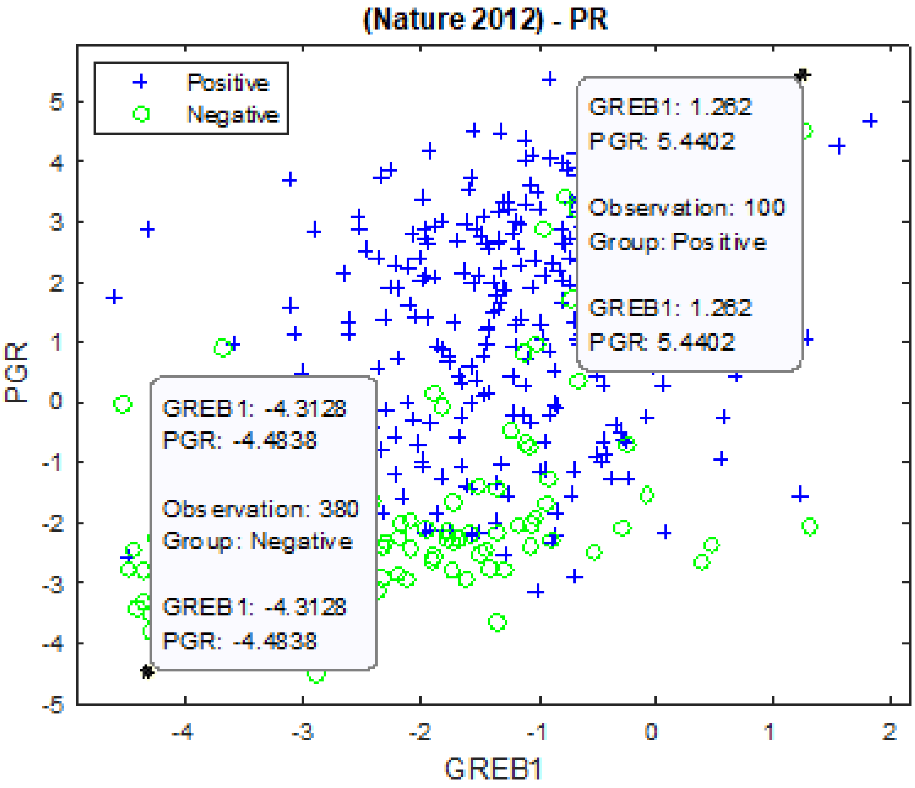


Figure 6.22: Scatter plot of GREB1 and PGR of (Nature 2012) dataset with PR groups, illustrating that the observation 100 from the PR+ group has high expression levels of (GREB1, PGR), which are (1.262, 5.4402) in comparison to the observation 380 from the PR- group, which has low expression levels of (GREB1, PGR), which are (−4.3128, −4.4838).

is frequently expressed in ovarian cancer. In our research, a positive relationship between GREB1 mRNA expression and PR levels was detected, as clarified in Figure 6.22, and across a large number of variations in breast cancer samples. Our findings revealed GREB1 mRNA as a potential indicator to the positivity of PR, therefore, this gene and its relevance to breast cancer should be investigated further.

- **{'PGR'}**. Progesterone receptor. According to the NCBI website[1], PGR *"The encoded protein mediates the physiological effects of progesterone, which plays a central role in reproductive events associated with the establishment and maintenance of pregnancy"*. The authors in [127] found that the expression of ESR1, the gene encoding ER$\alpha$ and that SNPs in the PGR gene predict tumour PGR/PgR expression, and they concluded that ESR1 and PGR polymorphisms are associated with estrogen and progesterone receptor expression in breast tumours. PGR was identified in this thesis to be a potential biomarker for the status of PR, as clarified in Figure 6.22, and over a wide range of variant and independent breast samples. The obtained findings reveal the potential of PGR gene to be a biomarker to breast cancer and PR positivity, thus further investigations are required.

## 6.5 Discovered Biomarkers with HN Weight for PR

The clinical relevance between the discovered biomarkers with HN weight and the progesterone receptor will be discussed in regards to bioinformatics analysis studies in the literature. Furthermore, the association of each mRNA marker to the hormone receptor PR identified in this thesis will be explored in details in the following points to provide conclusive evidence.

- **{'ATP6V0A4'}**. Recently, the authors in [254] have identified fourteen differentially expressed genes, involving ATP6V0A4 for visceral organ metastasis in breast cancer. Misra et al. [210] found that ZAR2 transcriptionally represses the ATPase ATP6V0A4 to negatively regulate invasiveness of breast cancer cells. This thesis observes the ATP6V0A4 gene to be negatively associated with PR status, as clarified in Figure 6.23, and over the (Nature 2012), (Cell 2015), (Provisional), and (METABRIC) in addition to the integrated datasets with PR groups. The findings of this research

[1] https://www.ncbi.nlm.nih.gov/gene/5241

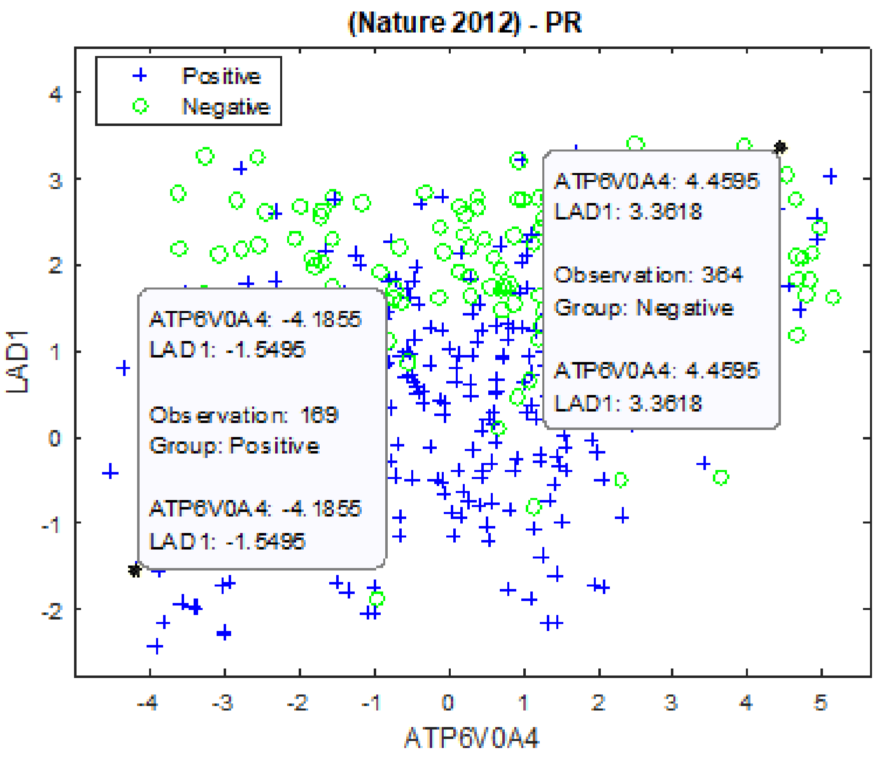


Figure 6.23: Scatter plot of ATP6V0A4 and LAD1 of (Nature 2012) dataset with PR groups, illustrating that the observation 169 from the PR+ group has low expression levels of (ATP6V0A4, LAD1), which are $(-4.1855, -1.5495)$ in comparison to the observation 364 from the PR- group, which has high expression levels of (ATP6V0A4, LAD1), which are $(4.459, 3.3618)$.

recommend considering the ATP6V0A4 gene and its link to breast cancer and PR in future studies to allow unexplored knowledge to be discovered.

- **{'LAD1'}**. Recently, Roth et al. [247] have identified LAD1 as a filamin-binding regulator of actin dynamics in response to the epidermal growth factor receptor (EGFR) and a marker of aggressive breast tumours. Groger et al. [115] found several genes, including LAD1 that are correlated significantly with impaired pathological complete response (pCR) in breast cancer patients. The researchers in [118] identified 50 genes, including LAD1 to be associated with ER in breast cancer. Several biomarkers including LAD1 were identified in [6] for Luminal A and Basal. In this thesis, evidence of an inverse correlation was detected between the expression patterns of LAD1 and PR, as illustrated in Figure 6.23, and over a large number of breast cancer samples of (Nature 2012), (Cell 2015), (Provisional), and (METABRIC) as well as the integrated datasets with PR groups. Therefore, this thesis supports investigating the potential of the LAD1 to be a biomarker for breast cancers and PR status.

- **{'C9orf58'}**. Chromosome 9 open reading frame 58, which is also known as allograft inflammatory factor 1 like (AIF1L). Recently, Liu et al. [194] have stated that AIF1L plays a key role in mammary tumorigenesis, and their findings have suggested AIF1L to be a potential prognostic marker that plays a vital role in regulating the cytoskeleton in breast cancer. Excluding this recent study, little is presented in the literature about the clinical relevance of this gene to breast cancer and the hormone receptors. In this thesis, C9orf58 was found to be a potential biomarker to breast cancer that is negatively associated with PR expression levels, as shown in Figure 6.24, and over a wide range of variant breast cancer samples. Therefore, more studies can be conducted to understand the biological mechanism underlying the inverse association between the expression patterns of C9orf58 mRNA and PR.

- **{'NXPH1'}**. A study of genome-wide methylation screen [93] in low-grade breast cancer identifies several epigenetically altered genes, including NXPH1 as potential biomarkers for tumour diagnosis. This thesis detected an inverse relationship between NXPH1 mRNA and PR levels, as presented in Figure 6.24, and across a wide range of breast cancer samples. The potential of NXPH1 gene to be a biomarker to breast cancer and PR can be further researched by biomedical studies to indicate its role in human breast cancers.

- **{'CLCA2'}**. According to NBCI website[1], " *In breast cancer, expression of this gene is down-regulated and the encoded protein may inhibit migra-*

[1] https://www.ncbi.nlm.nih.gov/gene/9635

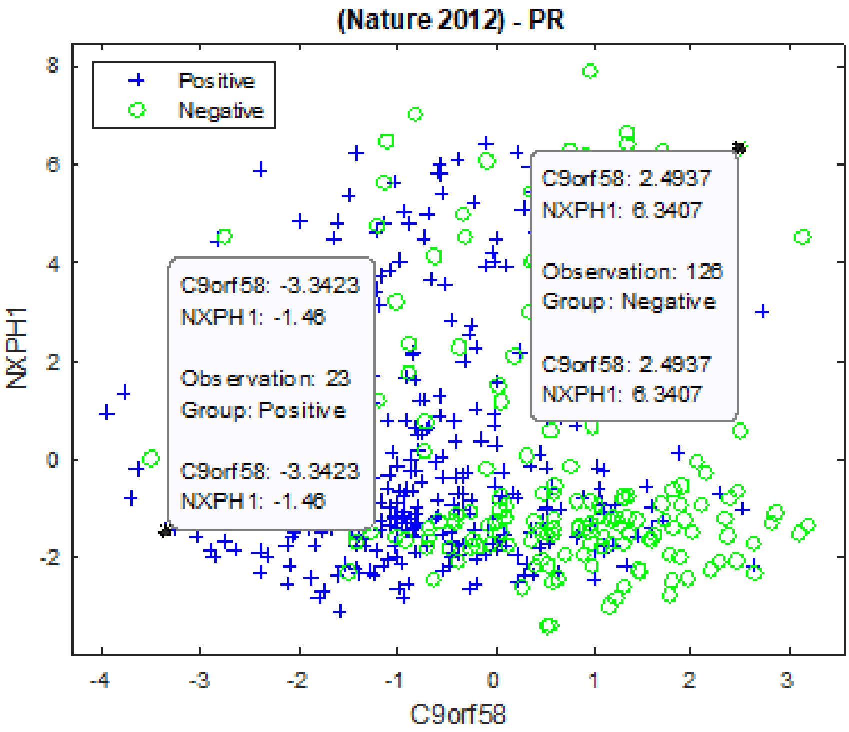


Figure 6.24: Scatter plot of C9orf58 and NXPH1 of (Nature 2012) dataset with PR groups, illustrating that the observation 23 from the PR+ group has low expression levels of (C9orf58, NXPH1), which are $(-3.3423, -1.46)$ in comparison to the observation 126 from the PR- group, which has high expression levels of (C9orf58, NXPH1), which are $(2.4937, 6.3407)$.

*tion and invasion while promoting mesenchymal-to-epithelial transition in cancer cell lines".* Sasaki et al. [252] found CLCA2 to be involved in the p53 tumour suppressor network and it significantly impacts cancer cell migration and invasion. CLCA2 was found in [184] to be frequently inactivated in breast cancer, which makes it a strong candidate for the 1p31 breast cancer tumour suppressor gene. Similar findings were found by [117], which stated that CLCA2 might act as a tumour suppressor in breast cancer. CLCA2 mRNA was recognised in this thesis as one of the biomarkers that are associated negatively with PR levels, as presented in Figure 6.26, and over a wide range of breast samples. Targeting CLCA2 gene for further studies

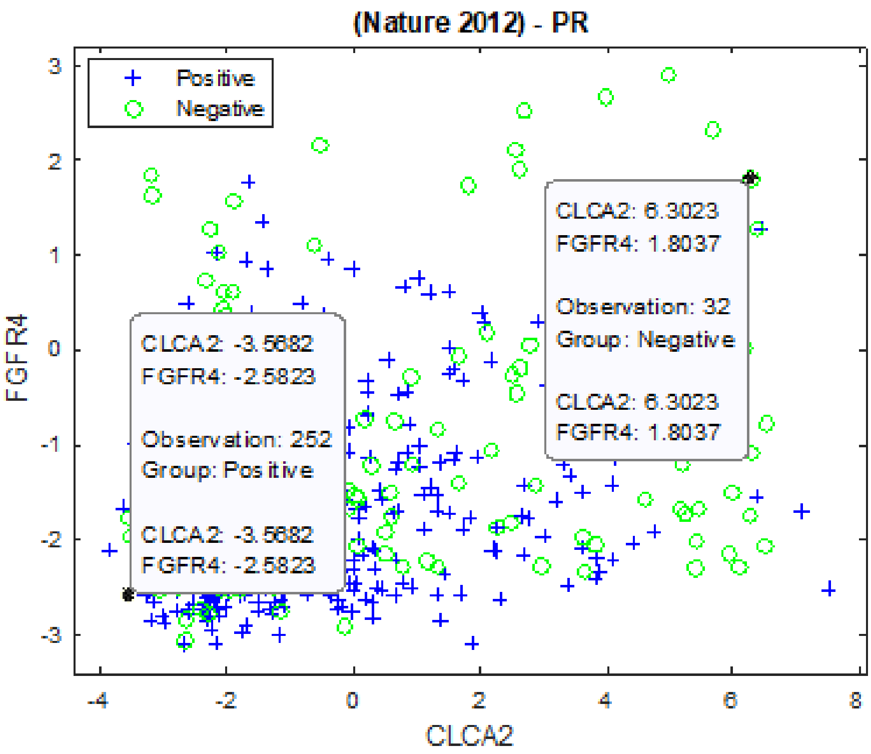


Figure 6.25: Scatter plot of CLCA2 and FGFR4 of (Nature 2012) dataset with PR groups, illustrating that the observation 252 from the PR+ group has low expression levels of (CLCA2, FGFR4), which are $(-3.5682, -2.5823)$ in comparison to the observation 32 from the PR- group, which has high expression levels of (CLCA2, FGFR4), which are $(6.3023, 1.8037)$.

can help to determine its role in breast cancer and PR positivity.

- **{'FGFR4'}**. According to NBCI website[1] website, *"The encoded protein is involved in the regulation of several pathways, including cell proliferation, cell differentiation, cell migration, lipid metabolism, bile acid biosynthesis, vitamin D metabolism, glucose uptake, and phosphate homeostasis"*. The recent review study [285] on the role of FGFR4 in cancers has stated that information on the involvement of FGFR4 in cancers has significantly increased in recent years and concluded targeting FGFR4 as a potential thera-

[1] https://www.ncbi.nlm.nih.gov/gene/2264

peutic strategy. Recently, Zhao et al. [337] have found that FGFR4 provides the conduit to facilitate FGF19 signaling in the progression of breast cancer. Another recent study [323] has suggested targeting FGFR4 as a therapeutic opportunity for chemoresistant tumours because it increases glucose metabolism and leads to chemoresistance in breast cancer. Recently, another study [180] has found that FGFR4 is a novel druggable target for recurrent ER-positive breast cancers. In this PhD study, evidence of an inverse association was observed between the expression patterns of FGFR4 and PR, as illustrated in Figure 6.26, and over variant and independent breast cancer samples of (Nature 2012), (Cell 2015), (Provisional), and (METABRIC) as well as across the variations in the integrated datasets with PR groups. More investigations are required to determine the role of FGFR4 mRNA in breast cancers and more specifically, in the progesterone receptor.

- **{'PPP1R1A'}**. As mentioned previously in Section 6.3, a research study [280] found that the expression of PPP1R1A in lung, colorectal, and gastric cancer cell lines was different from that of the normal tissues. This thesis is the first to report a negative correlation between PPP1R1A mRNA expression and high levels of ER, as well as PR positivity, as shown in Figure 6.26, and across independent cancer genomic datasets. Therefore, the potential of PPP1R1A mRNA expression to be a diagnostic or prognostic markers to breast cancer and hormone receptors can be further studied in future research.

- **{'TRPV6'}**. As mentioned previously in Section 6.3, TRPV6 was identified to be negatively associated with ER+ tumours. Moreover, it has been shown in this research that TRPV6 is lowly expressed in PR+ tumours, in which the drops in TRPV6 could contribute to the positivity of PR, as explained in Figure 6.26, and this has been demonstrated over several independent breast cancer genomic datasets. Therefore, TRPV6 can be further studied to indicate its contributions to breast cancer.

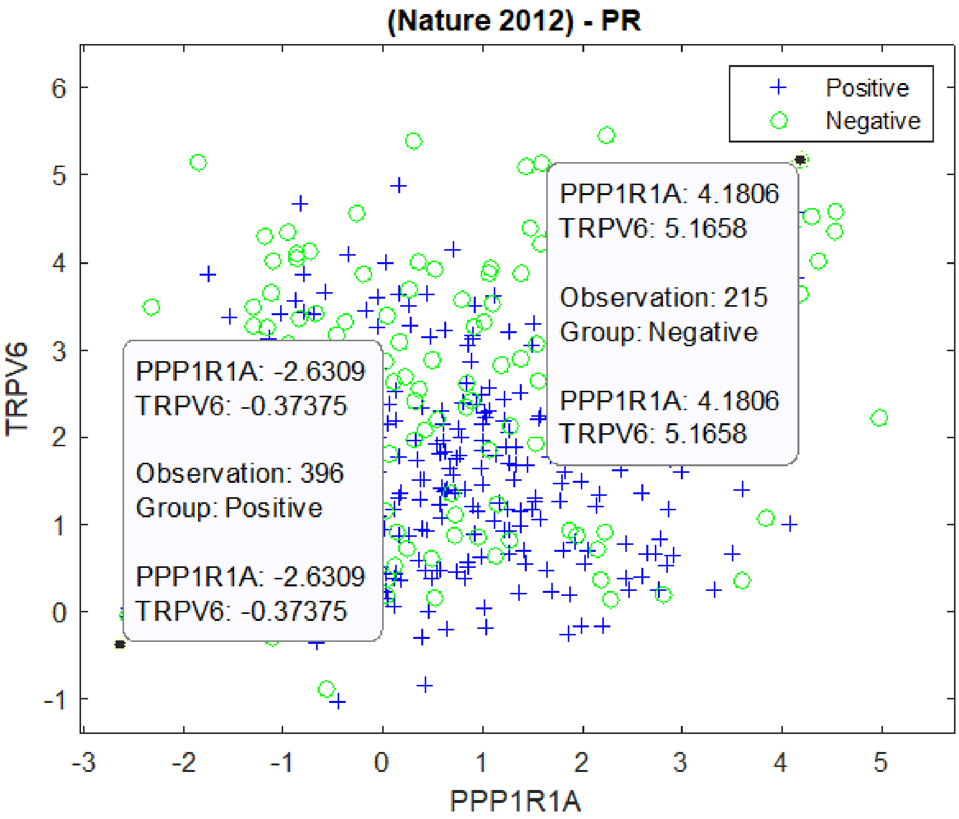


Figure 6.26: Scatter plot of PPP1R1A and TRPV6 of (Nature 2012) dataset with PR groups, illustrating that the observation 396 from the PR+ group has low expression levels of (PPP1R1A, TRPV6), which are $(-2.6309, -0.37375)$ in comparison to the observation 215 from the PR- group, which has high expression levels of (PPP1R1A, TRPV6), which are $(4.18067, 5.1658)$.

- **{'C1orf115'}**. Very limited information is available in the literature about this gene, particularly its relevance to breast cancers and the hormone receptors. This thesis is the first to show evidence of an inverse association between the expression patterns of Clorf115 and PR, in which the PR+ tumours are characterised by low expression levels of Clorf115 mRNA compared to PR- samples, as introduced in Figure 6.27, and over multiple independent datasets. Therefore, further investigations are necessitated to identify the biological mechanism underlying that association.

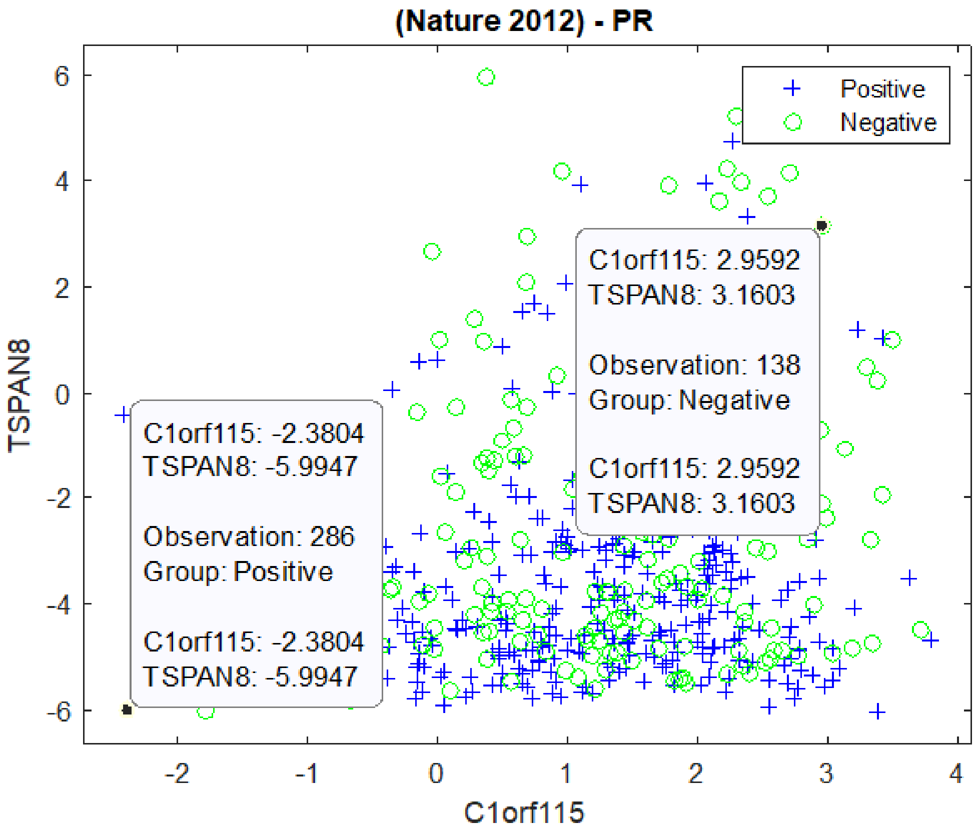


Figure 6.27: Scatter plot of C1orf115 and TSPAN8 of (Nature 2012) dataset with PR groups, illustrating that the observation 286 from the PR+ group has low expression levels of (C1orf115, TSPAN8), which are $(-2.3804, -5.9947)$ in comparison to the observation 138 from the PR- group, which has high expression levels of (C1orf115, TSPAN8), which are $(2.9592, 3.1603)$.

- **{'TSPAN8'}**. Growing evidence in the literature suggests that TSPAN8 promotes tumour cell migration, invasion, and metastasis in multiple types of human cancers [108,227]. Zhu et al. [342] have revealed that several genes, including TSPAN8 are positively correlated in human breast cancer, and high expression levels of TSPAN8 correlate with poor prognosis. This thesis demonstrated a negative association between TSPAN8 and PR positivity, as clarified in Figure 6.27, and across a wide range of breast cancer samples, thus targeting this gene for further researching could contribute to advance our knowledge about the role TSPAN8 might play in PR+ breast cancer.

## 6.6 Discussion

Previous chapters discussed the fundamental concepts, design, implementation of the proposed feature mining models as well as validation and evaluation of the discovered biomarkers. This chapter covered the clinical relevance of the recognised biomarkers to breast cancer and the hormone receptor ER and PR according to current bioinformatics research in the literature. Furthermore, conclusive evidence of a positive or negative association between each single mRNA marker and the hormone receptor ER or PR was introduced and verified.

The positive association was observed and validated between the discovered biomarkers with HP weight and ER, as discussed in Section 6.2, and the HP weighted biomarkers and PR, as mentioned in Section 6.4. The positive association corresponds to the gains in the expression levels of these biomarkers and its contribution to ER/PR positivity. The inverse correlation was recognised and demonstrated between the HN weighted biomarkers and ER, as explained in Section 6.3, as well as the identified generic biomarkers with HN weight and PR, as presented in Section 6.5. The negative association refers to the declines in the expression levels of these mRNA markers and its contribution to high ER/PR expression level. The emphasis was made in this chapter on conducting further investigations to examine the potential of the identified groups of biomarkers to improve breast cancer patient's health and survival or develop more cost-effective therapies.

The computational models reported in this thesis effectively extracted knowledge from omics data using a systematic approach to data modelling, analysis and validation. The discovered mRNA markers could answer different biological questions of interest and provide insights and explanations to the biological, pathological and pharmacological process underlying human breast cancers and the hormone receptors ER and PR. The assessment of the clinical relevance of the discovered biomarkers by bioinformatics researchers is essential for inferring conclusive evidence that enables to use the recognised molecular markers to construct diagnostic and prognostic models that can be accepted in clinical practice.

# Chapter 7

# Conclusion and Future Work

## 7.1 Introduction

This thesis discusses the fundamental concepts, research problems and directions for the discovery of robust biomarkers for diseases, such as cancers, from HDSSS omics data. The detailed and critical discussions presented considered relevant current state-of-the-art research to establish key limitations and challenges of employing them in bioinformatics and computational biology. Therefore, the limitations of existing approaches established by the literature review led to the proposal of two models, based on computational intelligence and deep learning, for the extraction, analysis, interpretation and validation of reliable biomarkers from human molecular data. Furthermore, the availability of data repositories and portals for different types of cancer genomic data, main requirements for effective genome analysis and feature mining, and critical aspects for assessing the outputs of omics data analysis models are all covered and discussed in details. Then, this thesis introduces the modelling design, analysis, implementation and application of the novel feature mining model that integrates traditional statistical techniques and computational intelligence methods for the goal of the discovery of the underlying structure of the genomic and proteomic data. It also proposes a general framework for deep feature learning together with an explanatory technique that can be used for discovering robustly high-level abstract representations from such datasets and reveal key determinants underlying these latent representations. The outputs of our computational models were validated using effective evaluation metrics and independent validations, thus relevant, robust, and reproducible biomarkers to breast cancer were discovered and verified. Moreover, the clinical relevance of the discovered biomarkers to human breast cancers and ER/PR positivity was also discussed in order to provide conclusive evidence about the type of underlying association.

## 7.2 Concluding Remarks

Extracting knowledge from omics datasets is a serious challenge for the research community interested in understanding the cancer genotype and phenotype. Such datasets are characterised by high dimensionality and relatively small sample sizes with small signal-to-noise ratios. This significantly challenges existing machine learning-based solutions due the curse of dimensionality issues, where the addition of new input features typically requires an exponential number of input observations (which are commonly unavailable) to discover the underlying structure of the data that allows these models to generalise well to unseen cases. This also puts great pressure on data mining models that attempt to separate the signal from the noise in a bid to discover robust determinants. Increasing the awareness of the key research challenges allow us to introduce more potential solutions, by understanding the required computational and statistical resources. The potential solutions introduced in this thesis to tackle the challenges of knowledge discovery from omics data are concluded in the following sections.

### 7.2.1 Filtering Methods

It is a well-known that much more accurate machine-learning methods are required to specify and measure phenotypes of complex diseases such as cancer. In particular, our focus has specifically been to reduce the amount of spurious positive associations within sophisticated classifier-based systems by proposing an intelligent feature mining model. One of the key challenges related to deriving knowledge from such high dimensional biomedical data is the amount of noise and the experimental variability. Filtering methods based on the variation and entropy criterion were employed to exclude genes that exhibit a variance or entropy value less than the 10th percentile from further analysis. The resulting outcomes of filtering methods were that the genomic datasets had less noise or less unreliable genes that are not expressed at biologically significant levels. The preliminary quality assessment can be considered an essential step in the process of biomarker discovery from omics data in order to increase the quality of the data prior to the modelling and analysis stages. Therefore, the practical impact of filtering out the least reliably expressed genes led to increasing the potential of detecting differentially expressed genes.

### 7.2.2 Evolutionary Mining Model

As mentioned earlier, extracting knowledge from omics datasets is a serious challenge for machine learning-based solutions due to the curse of dimensionality issues. To alleviate these limitations , the evolutionary mining model was proposed based on ad-hoc traditional statistical techniques with the computational evolutionary method to effectively handle the size and complexity of omics data. Therefore, the proposed model comprises of three main phases, based on different selection paradigms, which are univariate, multivariate, and ensemble. The univariate selection phase is utilised first to eliminate the least promising features for the next optimisation phase. The multivariate selection phase, based on the evolutionary method, is used to optimise the search process for finding the best possible combination of features in the reduced feature space. The ensemble phase is utilised to enhance the robustness of the finally selected subset of candidate predictors. Evaluating the performance of the evolutionary mining model reveals that this multi-staged model was successful to some extents in recognising robust biomarkers from several HDSSS microarray and mass spectrometry datasets. The discovered biomarkers exhibited computational and biological relevance and were capable of developing highly accurate and reliable prediction systems. However, it has been shown through this research that the number of identified biomarkers by this mining model that are generic over independently generated breast cancer samples was small compared to the outputs of the deep mining model.

Therefore, the requirement for multiple levels of feature learning is necessitated to exploit the unknown structure of HDSSS omics data efficaciously for capturing enough of its underlying relevant variations. This has motivated us to explore another direction of research that focuses on unsupervised feature extraction, rather than supervised feature selection, based on state of the art deep learning for the goal of automatic deep self-taught learning so that the interesting complexity can be uncovered adequately from the raw high dimensional genomic and proteomic data. However, the extraction process based on deep learning methods lacks the transparency and interpretability found in the feature selection based on the evolutionary mining model. To address this issue, our research introduces a novel weight interpretation technique that helps to open the black box of such deep learning models to reveal key determinants underlying its latent representations.

### 7.2.3 Deep Mining Model

In this thesis, we critically evaluate the usefulness of state-of-the-art deep neural network methods for the problem of knowledge discovery from high throughput biomedical data. The key requirements for automated deep feature learning model that is able to handle the crucial challenges underlying the problem of inferring knowledge from HDSSS omics data are discussed and established. Therefore, the proposed deep feature learning model was introduced, based on a set of non-linear sparse Auto-encoders, that are deliberately constructed in an under-complete manner to force the network to find progressively the complex featural representations necessary to capture the important variations underlying the biological samples.

As discussed previously, the dimensionality of omics data is high, which means that there is an exponential number of possible input configurations. Therefore, the available biological samples become increasingly sparse, making the process of discovering plausible and robust input configurations a very difficult task. Moreover, very few genes are expressed reliably at biologically significant levels and distinguishably from noise and measurement variation [32]. Therefore, the Compressed Auto-encoder attempts to reduce the number of biological samples required to find a small proportion of molecules that can recover a large proportion of variations underlying the data. The Sparse Compressed Auto-Encoder endeavours to promote the notion that different aspects are characterised by different features. Adding the sparsity penalty to the under-complete layers leads to rendering the hidden neurons to be inactive (i.e. at or near zero) so that a small set of different groups of hidden neurons allocated to different subsets of features. As a result, a small proportion of potentially relevant and insensitive determinants is utilised to represent various inputs through multiple levels of feature transformations of the Stacked Sparse Compressed Auto-Encoder. Consequently, the learning process proceeds successfully using the available samples addressing the problem of small sample sizes of omics, where the number of molecules vastly exceeds the number of observations. Furthermore, the computational and statistical challenges arising from handling the high dimensional spaces of genomic and proteomic data are tackled, and a high level of efficiency is achieved.

Consequently, the proposed deep feature learning model was very successful when applying to the individual cancer datasets, as well as the composite datasets, which are derived from the individual breast invasive carcinoma datasets. The characteristics of the presented deep feature learning model were very effective in handling the challenges of cancer genomic datasets and discovering highly non-linear generic features, capturing high degrees of variation amongst breast cancer samples. Given its ability to learn complex functional relationships of varying degrees of abstraction, it is expected that the proposed deep feature learning model will detect any salient high-level generic features that are latent yet pervasive across the data.

### 7.2.4 The weight Interpretation Method

The weight interpretation method was introduced in this thesis to add explanatory power to the deep feature learning model by determining the candidate input features that force the different biomarker classification behaviours. The detailed evaluation of the deep mining model demonstrates its capacity to open the black-box of the deep feature learning model by finding robustly the deferentially expressed genes or proteins that exhibit HP or HN weight scores over the depth of the network. Fundamentally, two types of outcomes were revealed by our deep mining model, both indicating strong likelihoods of a patient having cancer. The first outcome indicated a subset of highly positively-weighted genes whereby the amplifications and gains in the gene expression levels were associated with the likelihood of a patient having cancer. Conversely, the second outcome revealed another subset of genes that were highly negatively-weighted and coincided with significant downregulation in the gene expression levels, and again indicated the strong likelihood of a patient having cancer.

This mechanism explains the internal state of the proposed deep feature learning model, which relies mainly on allocating HP weight to the features that are highly expressed for the positives in comparison to most of the negatives. In contrast, HN weights are allocated by the proposed deep feature learning model to the features that are lowly expressed for most of the positives in comparison to the negatives. The detailed evaluation of the proposed weight interpretation method provides significant evidence that this explanatory technique was very effective in offering explainability to the deep learning model and detect key determinants underlying its latent representations. As our deep learning model is problem-independent and data-driven, it provides a general framework for knowledge discovery applications based on deep feature learning to omics data characterised by high dimensionality and relatively small sample size.

### 7.2.5 Discovered Biomarkers

The application of the deep feature learning model together with the deep mining interpretation method to the breast invasive carcinoma datasets results in detecting *relevant*, *robust*, and *reproducible* biomarkers over a wide range of independently generated breast cancer samples. The clinical relevance of these molecular indicators to breast cancer is discussed, in regards to the bioinformatics analysis studies in the literature, where the type of relationship between each mRNA and the hormone receptor ER and PR recognised in this research is revealed and proved. Some of these biomarkers have been explored individually by other bioinformatics studies in the literature, in addition to our research. Therefore, there is growing evidence that the gains or declines in the expression levels of these biomarkers contribute to human breast cancers and the positivity of the hormone receptors ER and PR.

The biomarkers that are detected in this PhD work to have a positive association with ER levels are: {‘AGR3‘, ‘ESR1’, ‘GFRA1’, ‘SIAH2’, ‘SLC39A6’, ‘SCUBE2’, ‘C6orf97’, ‘ANXA9’, ‘CA12’, ‘NAT1’, ‘GATA3’, ‘PCP2’, ‘FSIP1’, ‘EVL’, ‘LRRC56’, ‘IGFALS’}. In the literature, there is increasing evidence about the existence of some of these biomarkers in breast cancers like {‘ESR1’, ‘GFRA1’, ‘AGR3’, ‘SIAH2’, ‘NAT1’, ‘SCUBE2’, ‘GATA3’}, while {‘C6orf97’, ‘SLC39A6’, ‘ANXA9’, ‘CA12’, ‘EVL’, ‘FSIP1’, ‘IGFALS’, } have been detected by few studies, where limited information is available about {‘PCP2’, ‘LRRC56’}, which are recognised for the first time to be positively correlated with ER levels.

The biomarkers that are found in this research to have an inverse correlation with ER positivity are: {‘PSAT1’, ‘PPP1R14C’, ‘TMEM40’, ‘VGLL1’, ‘C1orf106’, ‘BBOX1’, ‘SOX11’, ‘PROM1’, ‘DKK1’, ‘PARRES1’, ‘S100A8’, ‘S100A9’, ‘TRPV6’, ‘B3GNT5’, ‘KRT16’, ‘KRT81’, ‘HRASLS’, ‘PPP1R1A’}. There is growing evidence in the literature that some of these biomarkers are frequent events in breast cancer and triple-negative breast cancer, such as {‘VGLL1’, ‘PROM1’, ‘PSAT1’}, while {‘PPP1R14C’, ‘SOX11’, ‘B3GNT5’, ‘KRT16’, ‘DKK1’, ‘ S100A8’, ‘ S100A9’, ‘TRPV6’}, have been discovered by a few studies, where little is known about {‘TMEM40’, ‘C1orf106’, ‘BBOX1’, ‘KRT81’, ‘RARRES1’, ‘HRASLS’, ‘PPP1R1A’} and their inverse association with the hormone receptor ER has not yet been recognised.

For the hormone receptor PR, the biomarkers that are discovered in this PhD study to have a positive association with PR expression levels are: {‘FGD3’, ‘GFRA1’, ‘GRPR’, ‘PGR’, ‘SUSD3’, ‘GREB1’, ‘SIAH2’, ‘SCUBE2’, ‘AGR3’, ‘PGLYRP2’}. In the literature, there is growing evidence that demonstrates the role of some of these biomarkers in breast cancers and PR positivity such as {‘FGD3’, ‘SUSD3’, ‘GRPR’, ‘PGR’, ‘GREB1’}. While limited information is available about the role of {‘AGR3’, ‘GFRA1’, ‘SIAH2’, ‘SCUBE2’, ‘PGLYRP2’} in high PR levels. The biomarkers that are discovered in this study to be negatively associated with PR expression levels are: {‘LAD1’, ‘ATP6V0A4’, ‘NXPH1’, ‘C9orf58’, ‘CLCA2’, ‘FGFR4’, ‘PPP1R1A’, ‘TRPV6’, ‘C1orf115’, ‘TSPAN8’}. These biomarkers have also been detected by a few studies in the literature, whereas the inverse correlation between their expression patterns and the hormone receptor PR has not yet been recognised.

Furthermore, four subsets of biomarkers from the METABRIC breast cancer dataset were discovered robustly to be relevant to breast cancers and the hormone receptors ER and PR. Most of these mRNA markers were also discovered from the (Nature 2012), (Cell 2015) and (Provisional) datasets, as discussed in Chapter 6. The clinical relevance of these biomarkers is investigated in the literature individually by bioinformatics research. For example, a very recent state of the art finding that is congruent with the results of our research is the biomarker CD24 that is shown in our research to be negatively associated with ER levels. A very recent study from Stanford University [16] has revealed CD24 as 'don't eat me' signal, which stops immune cells engulfing and destroying the cancer cell. The recent study has found CD24 to be present in high quantities on the surface of both ovarian and triple negative breast cancer cells and was investigated to see if blocking this could lead to tumour shrinkage.

As a result, this thesis identifies diverse biomarkers that are positively and negatively associated with breast cancer and the hormone receptor ER and PR. The new risk determinants can be further investigated by domain experts to examine the potential of these genes to be clinical markers for the presence and progression of this heterogeneous disease, in addition to the predictivity, they can add to the diagnosis and prognosis models. Moreover, more personalised treatments or monitoring plannings for breast cancer could be developed by scouting the mechanism underlying the association of the expression patterns of these molecular markers and ER/PR positivity.

Herein, it is important to mention *a significant obstacle for biomarker discovery research, which is the need for more effective interdisciplinary research environments. There are relatively few examples of situations where novel molecular markers originating from the cancer research community has found its way into routine clinical practice. Effective inter-disciplinary research is therefore paramount if findings from state-of-the-art machine learning research is to be truly exploited and brought into the service of precision medicine.*

## 7.3 Directions for Future Work

Our deep feature learning and mining models drive novel molecular markers for breast cancer. The discovered biomarkers are reported after the rigorous and detailed designation of each step of the biomarker discovery process, which allow reproducing the molecular markers across independently generated cancer genomic datasets. Furthermore, it has been shown through this thesis that the discovered biomarkers are exhibiting clinical relevance potential to the hormone receptors ER and PR. The breast cancer datasets used in this thesis are publically available, thus investigating the findings of this thesis by biomedical experts can contribute in developing approaches for personalised cancer medicine that can prevent, screen, manage and treat this complicated disease and enhance the breast cancer patient's life and survival.

Moving forward, we will investigate the capacity of our deep mining model to detect generic biomarkers for selected cancers across a range of independent high-quality genomic samples collected from different studies. This will indicate which of these the academic and wider biomedical community should explore further. Furthermore, research studies have shown that complex diseases like cancers are extremely heterogeneous and caused by the complex interaction of various underlying factors, including genetic, genomic, behavioural and environmental effects and factors. The rise of high quality integrated and multi-modal omics data, such as the TCGA database which contains a combination of genomic, epigenomic, proteomic, imaging and clinical data for matched patient groups, will enable us to develop sophisticated 'integrative models' that may reveal even more valuable indicators of disease. We feel this will provide a sound basis for the development of more effective diagnostic and prognostic systems in the future.

Several studies have constructed various integrative analysis models to investigate the integration gain of diverse biomedical data. Most of these integrative studies have adopted clinicogenomic models that rely on combining clinical and genomic datasets. Clinicogenomic integrative models focus on addressing the challenges of integrating disparate dimensionalities of clinical and high dimensional genomic datasets. In terms of biological problems, most clinicogenomic studies use gene expression data from widely available public genomic datasets, despite the fact that each of these datasets provides variant aspects about the cellular activity. Therefore, the next generation of diagnosis and prognosis systems is to develop approaches for examining the potential interactions between a range of diversified cancer data.

The development of integrative prediction models from HDSSS genomic data poses a range of challenging issues that arise due to experimental, computational, and statistical complexities. All of these challenges were already discussed in our research paper: *"Challenges in Developing Prediction Models for Multi-Modal High-Throughput Biomedical Data"* [9]. The various challenges encountered are based on the characteristics of the data, the aim of the integration, and the level of the integration. Furthermore, three integration levels namely the Early, Intermediate, and Late were illustrated in this paper and the emphasis was made that the appropriate integration stage can be identified based on the aim of the analysis model and the characteristics of the datasets. The directions are introduced briefly in this paper to address these challenges and some possibilities for future work are discussed.

Therefore, as a direction for future research, we are aiming to design and implement an integrative analysis model that can leverage various cancer datasets for answering diverse systems biology questions. The integration gain and the differences in performances between multi-modal and uni-modal approaches using multiple datasets and diverse biomedical modalities will be investigated.

# Appendix A

Table 1: The sizes of the training-validation sets of ovarian cancer dataset and METABRIC dataset with ER and PR groups.

| Dataset | Training Sets | Validation sets |
|---|---|---|
| Ovarian Cancer | [173, 172, 173, 173, 173] | [43, 44, 43, 43, 43] |
| METABRIC | [1524 1523 1523 1523 1523] | [380, 381, 381, 381, 381] |

Table 2: The average MSE of each SCAE of ovarian cancer dataset and METABRIC datasets with ER and PR groups.

| Dataset | $L^1$ | $L^2$ | $L^3$ | $L^4$ |
|---|---|---|---|---|
| Ovarian Cancer | 0.0016 | 0.0010 | 0.0024 | 0.0024 |
| METABRIC with ER | 0.0356 | 0.0130 | 0.0074 | 0.0037 |
| METABRIC with PR | 0.0373 | 0.0125 | 0.0079 | 0.0036 |

Table 3: The sizes of the training-validation sets of the breast invasive carcinoma datasets with ER groups.

| Dataset | Training Sets | Validation sets |
|---|---|---|
| (Nature 2012) | [416, 415, 415, 415, 415] | [103, 104, 104, 104, 104] |
| (Cell 2015) | [332, 332, 332, 332, 332] | [83, 83, 83, 83, 83] |
| (Provisional) | [416 415 415 415 415] | [103, 104, 104, 104, 104] |

Table 4: The sizes of the training-validation sets of the breast invasive carcinoma datasets with PR groups.

| Dataset | Training Sets | Validation sets |
|---|---|---|
| (Nature 2012) | [415, 414, 414, 414, 415] | [103, 104, 104, 104, 103] |
| (Cell 2015) | [332, 332, 332, 332, 332] | [83, 83, 83, 83, 83] |
| (Provisional) | [415, 414, 414, 414, 415] | [103, 104, 104, 104, 103] |

Table 5: The average MSE of each SCAE of the breast invasive carcinoma datasets with ER groups.

| Layer | (Nature 2012) | (Cell 2015) | (Provisional) |
|---|---|---|---|
| $L^1$ | 0.1446 | 0.1363 | 0.1368 |
| $L^2$ | 0.0160 | 0.0147 | 0.0158 |
| $L^3$ | 0.0087 | 0.0084 | 0.0087 |
| $L^4$ | 0.0037 | 0.0041 | 0.0040 |

Table 6: The average MSE of each SCAE of the breast invasive carcinoma datasets with PR groups.

| Layer | (Nature 2012) | (Cell 2015) | (Provisional) |
|---|---|---|---|
| $L^1$ | 0.1334 | 0.1351 | 0.1356 |
| $L^2$ | 0.0158 | 0.0147 | 0.0161 |
| $L^3$ | 0.0089 | 0.0088 | 0.0088 |
| $L^4$ | 0.0038 | 0.0039 | 0.0039 |

Table 7: The sizes of the training-validation sets of the integrated datasets with ER groups.

| Dataset | Training Sets | Validation sets |
|---|---|---|
| NCP1 | [583, 582, 582, 582, 583] | [145, 146, 146, 146, 145] |
| NCP2 | [520, 520, 520, 520, 520] | [130, 130, 130, 130, 130] |
| NCP3 | [584, 583, 583, 583, 583] | [145, 146, 146, 146, 146] |

Table 8: The sizes of the training-validation sets of the integrated datasets with PR groups.

| Dataset | Training Sets | Validation sets |
|---|---|---|
| NC | [528, 527, 527, 527, 527] | [131, 132, 132, 132, 132] |
| CN | [474, 473, 473, 474, 474] | [118, 119, 119, 118, 118] |
| NP | [556, 556, 556, 556, 556] | [139, 139, 139, 139, 139] |
| PN | [557, 556, 557, 557, 557] | [139, 140, 139, 139, 139] |
| CP | [473, 472, 473, 473, 473] | [118, 119, 118, 118, 118] |
| PC | [528, 527, 527, 527, 527] | [131, 132, 132, 132, 132] |

Table 9: The average MSE of each SCAE of the integrated datasets with ER groups.

| Layer | NCP1 | NCP2 | NCP3 |
|---|---|---|---|
| $L^1$ | 0.0827 | 0.0770 | 0.0824 |
| $L^2$ | 0.0194 | 0.0198 | 0.0197 |
| $L^3$ | 0.0096 | 0.0097 | 0.0098 |
| $L^4$ | 0.0045 | 0.0047 | 0.0044 |

Table 10: The average MSE of each SCAE of the integrated datasets with PR groups.

| Layer | NC | CN | NP | PN | CP | PC |
|---|---|---|---|---|---|---|
| $L^1$ | 0.1049 | 0.0962 | 0.0931 | 0.0846 | 0.0922 | 0.1000 |
| $L^2$ | 0.0188 | 0.0191 | 0.0199 | 0.0198 | 0.0194 | 0.0189 |
| $L^3$ | 0.0091 | 0.0094 | 0.0095 | 0.0094 | 0.0098 | 0.0093 |
| $L^4$ | 0.0041 | 0.0049 | 0.0043 | 0.0043 | 0.0044 | 0.0042 |

# Appendix B

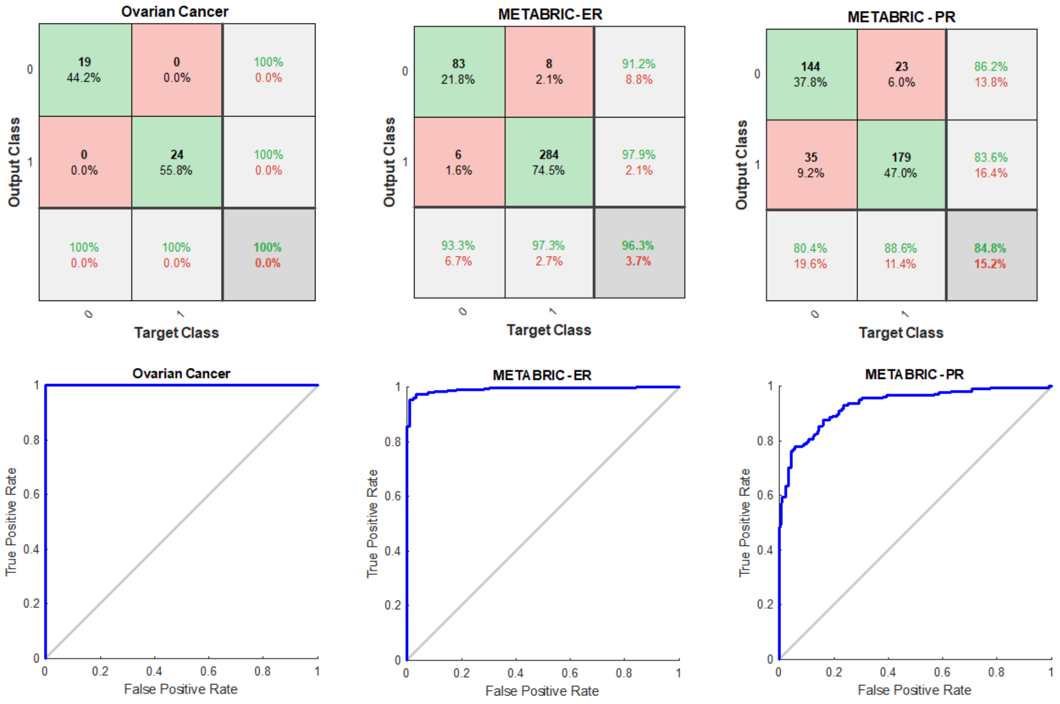


Figure 1: The performance of the SSCAE at the final iteration of ovarian cancer dataset and METABRIC dataset with ER and PR groups.

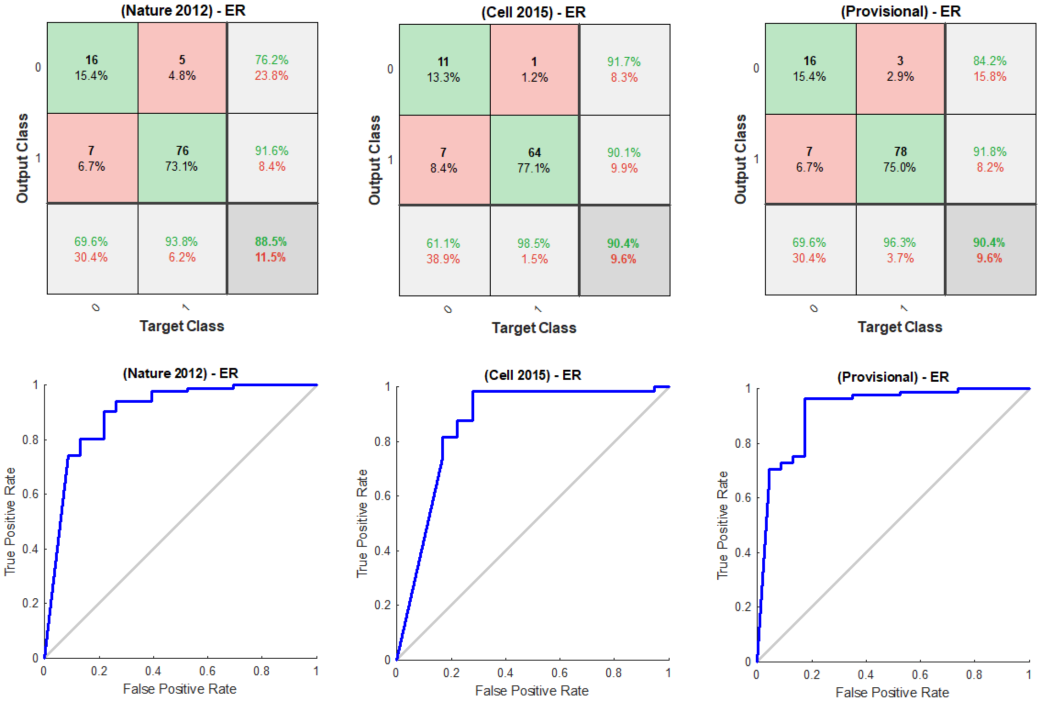


Figure 2: The performance of the SSCAE at the final iteration of the breast invasive carcinoma datasets with ER groups.

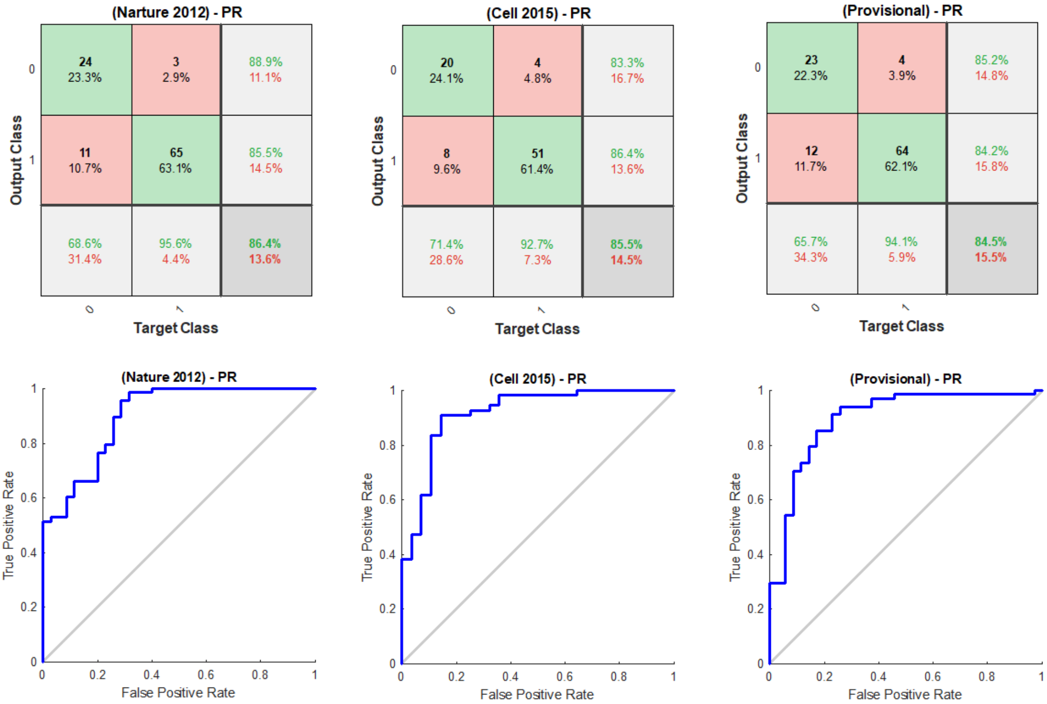


Figure 3: The performance of the SSCAE at the final iteration of the breast invasive carcinoma datasets with PR groups.

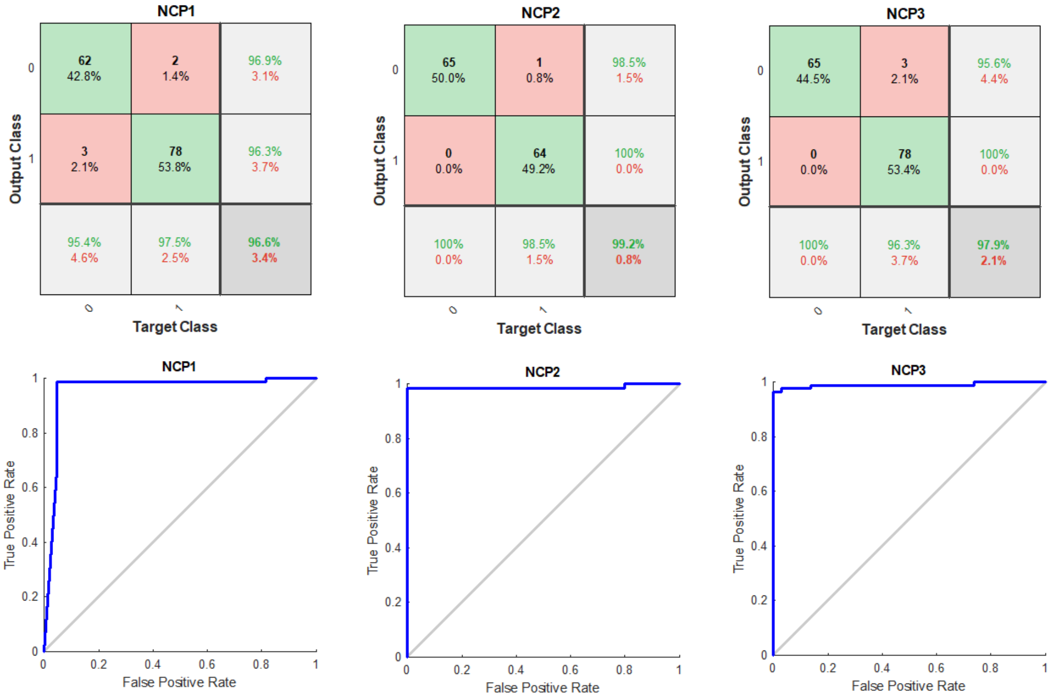


Figure 4: The performance of the SSCAE at the final iteration of the integrated datasets with ER groups.

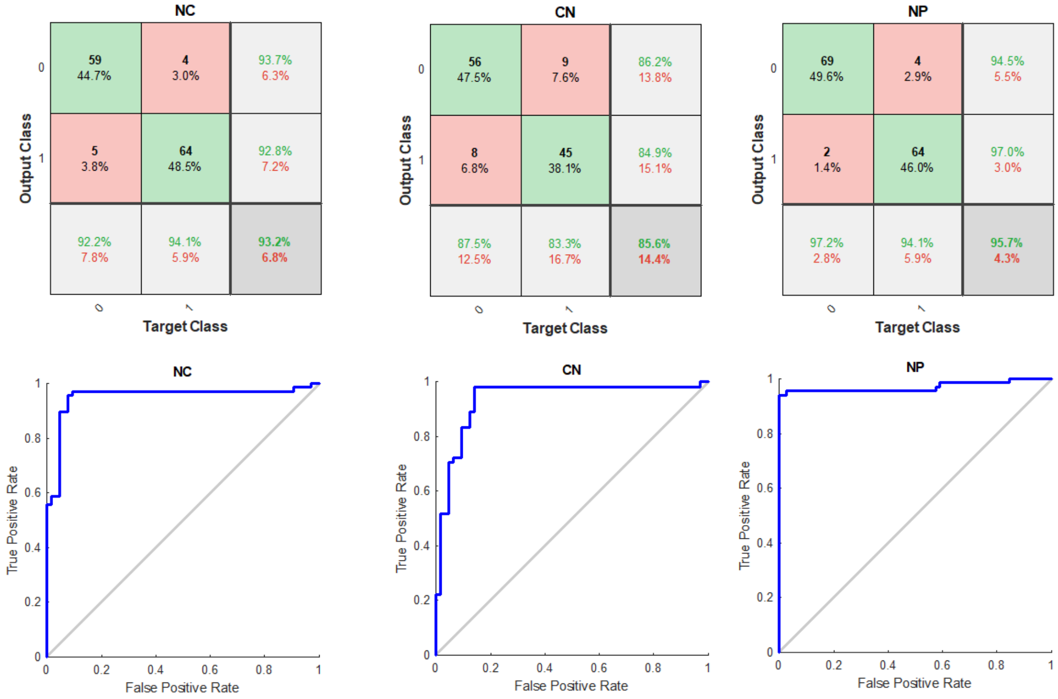


Figure 5: The performance of the SSCAE at the final iteration of the integrated datasets with PR groups.

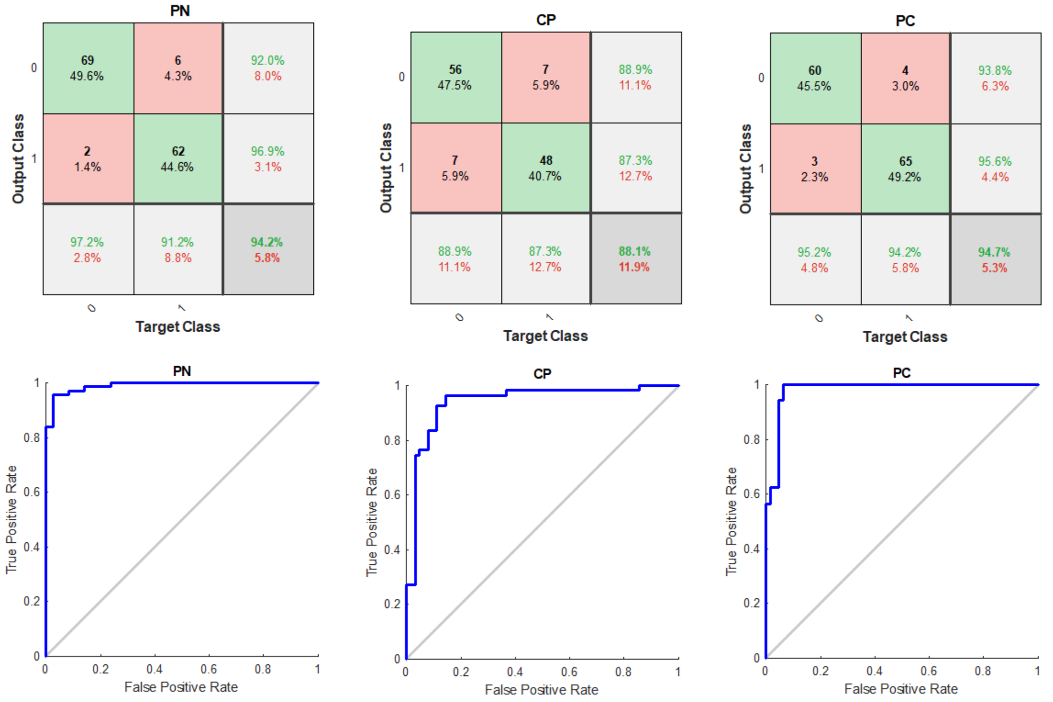


Figure 6: The performance of the SSCAE at the final iteration of the integrated datasets with PR groups.

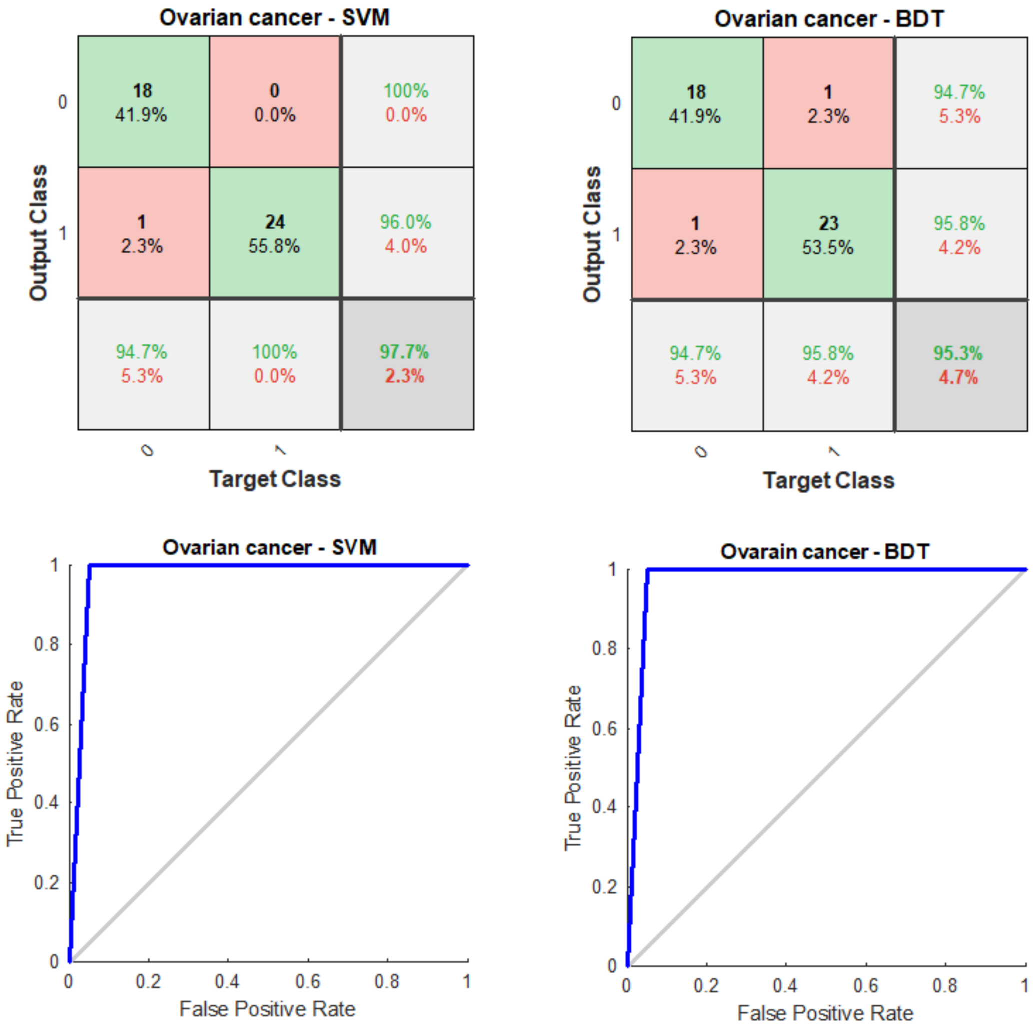


Figure 7: The performance of the SVM and BDT models at the final iteration of ovarian cancer dataset.

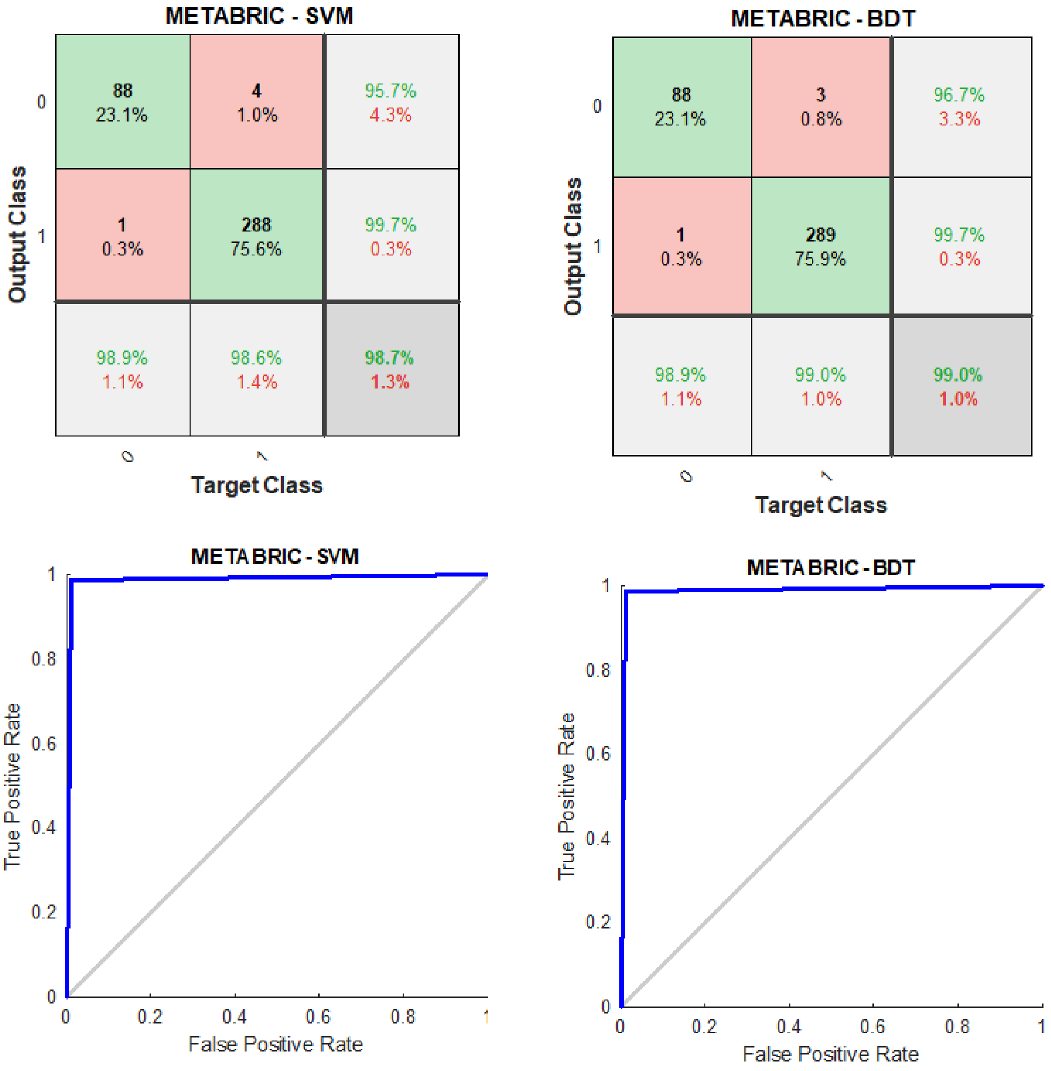


Figure 8: The performance of the SVM and BDT models at the final iteration of METABRIC dataset with ER groups.

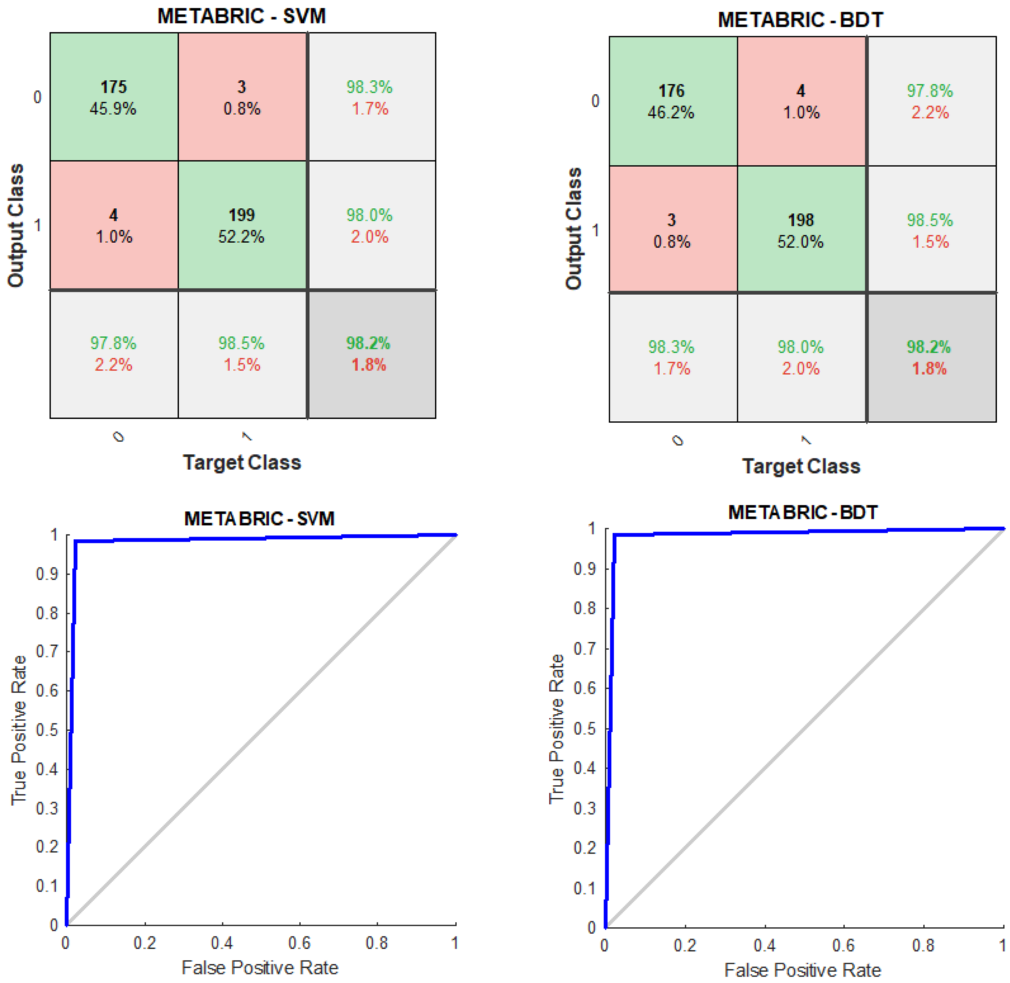


Figure 9: The performance of the SVM and BDT models at the final iteration of METABRIC dataset with PR groups.

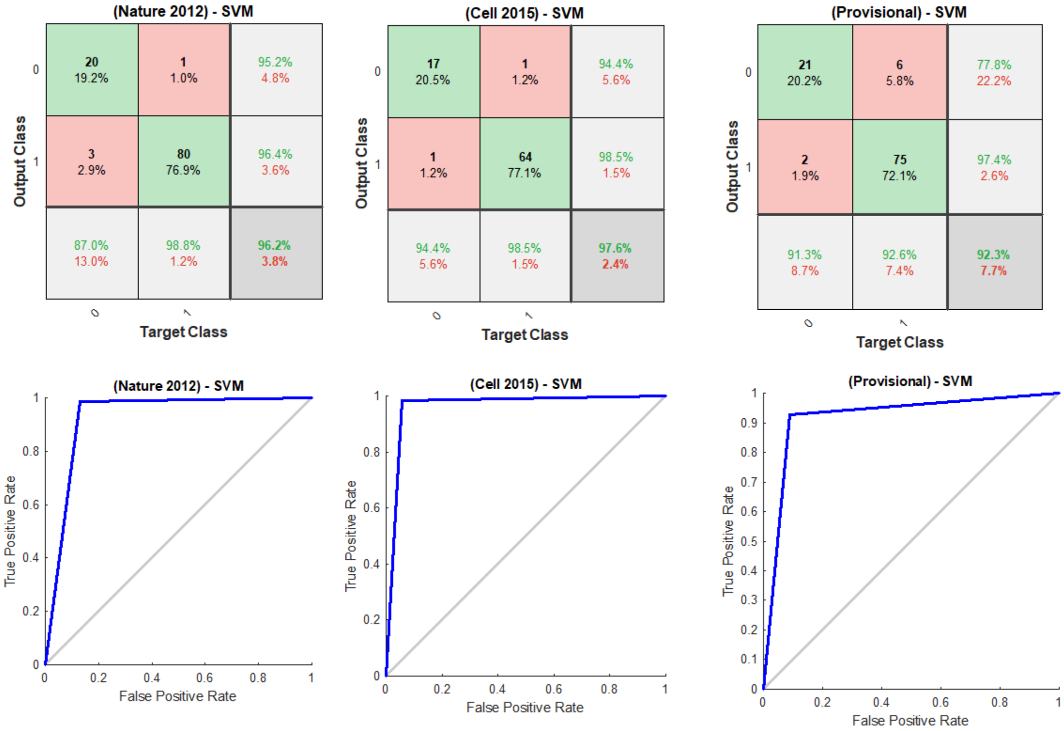


Figure 10: The performance of the SVM model at the final iteration of breast invasive carcinoma dataset with ER groups.

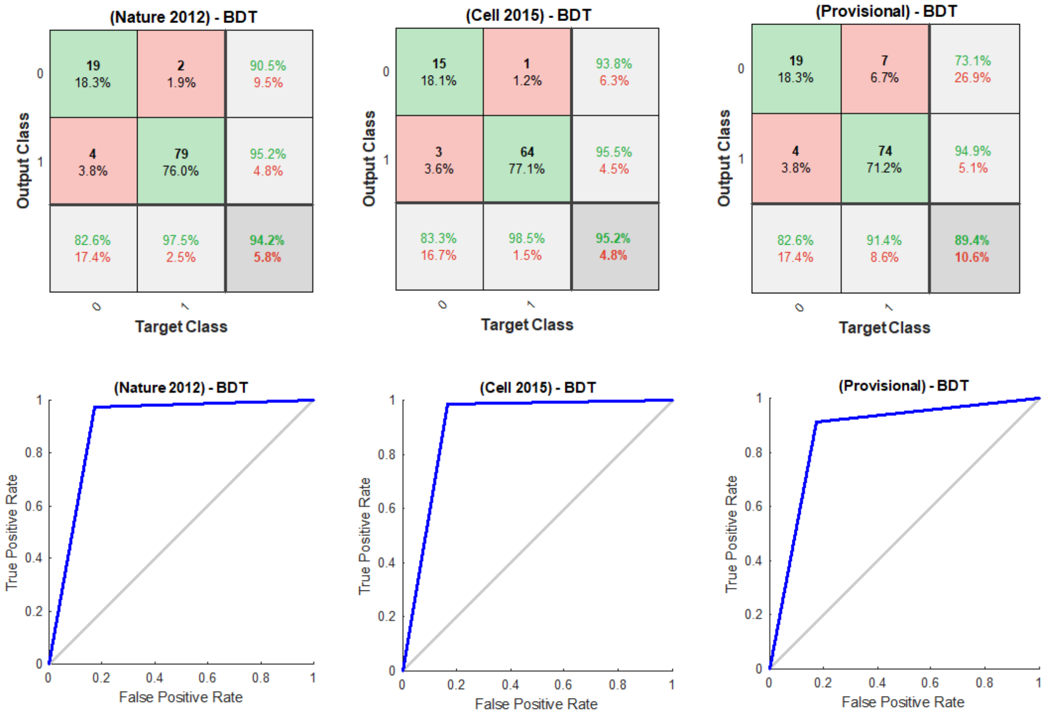


Figure 11: The performance of the BDT model at the final iteration of breast invasive carcinoma dataset with ER groups.

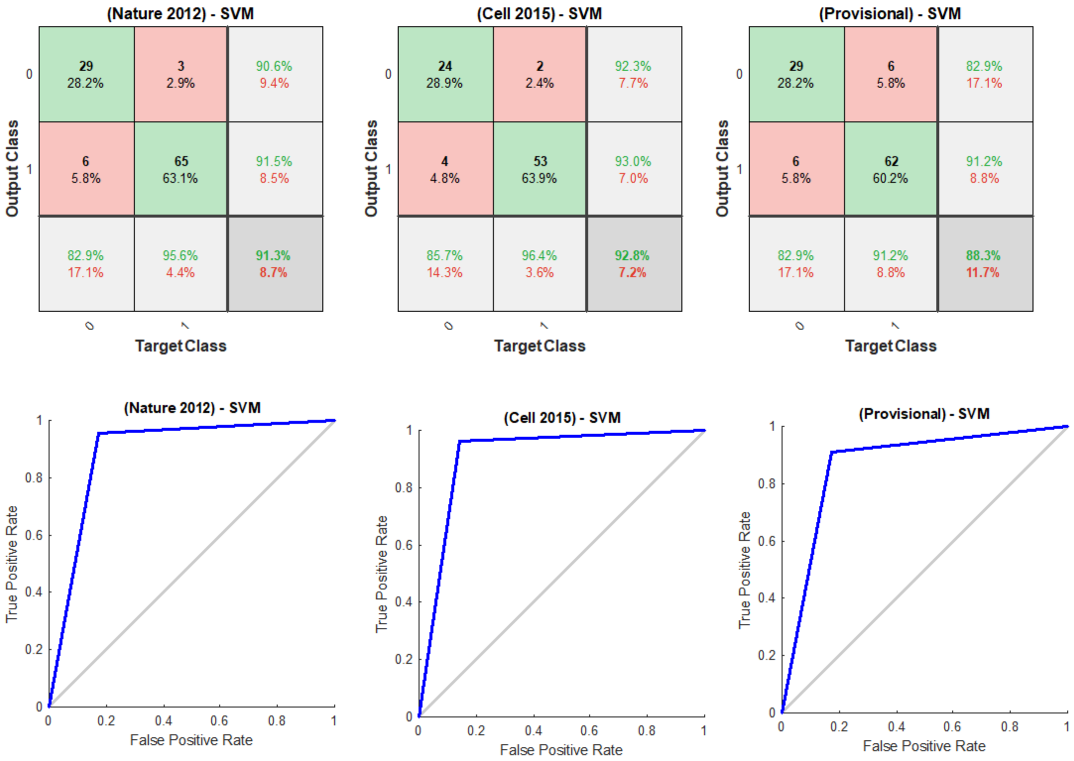


Figure 12: The performance of the SVM model at the final iteration of breast invasive carcinoma dataset with PR groups.

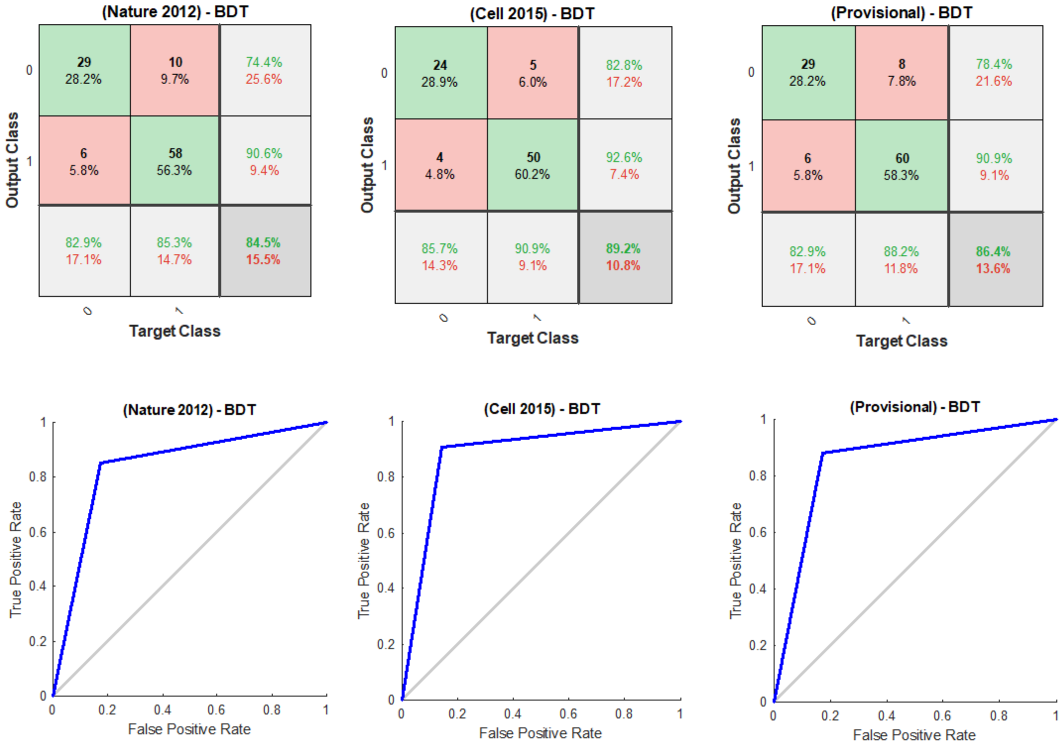


Figure 13: The performance of the BDT model at the final iteration of breast invasive carcinoma dataset with PR groups.

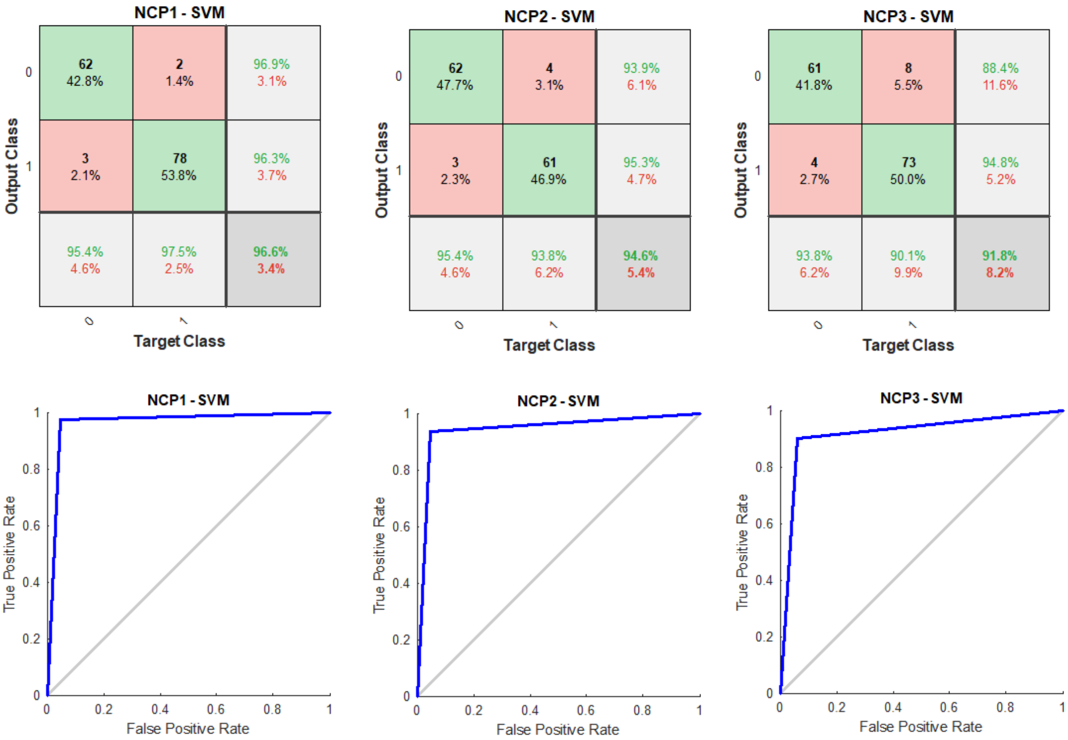


Figure 14: The performance of the SVM model at the final iteration of the integrated dataset with ER groups.

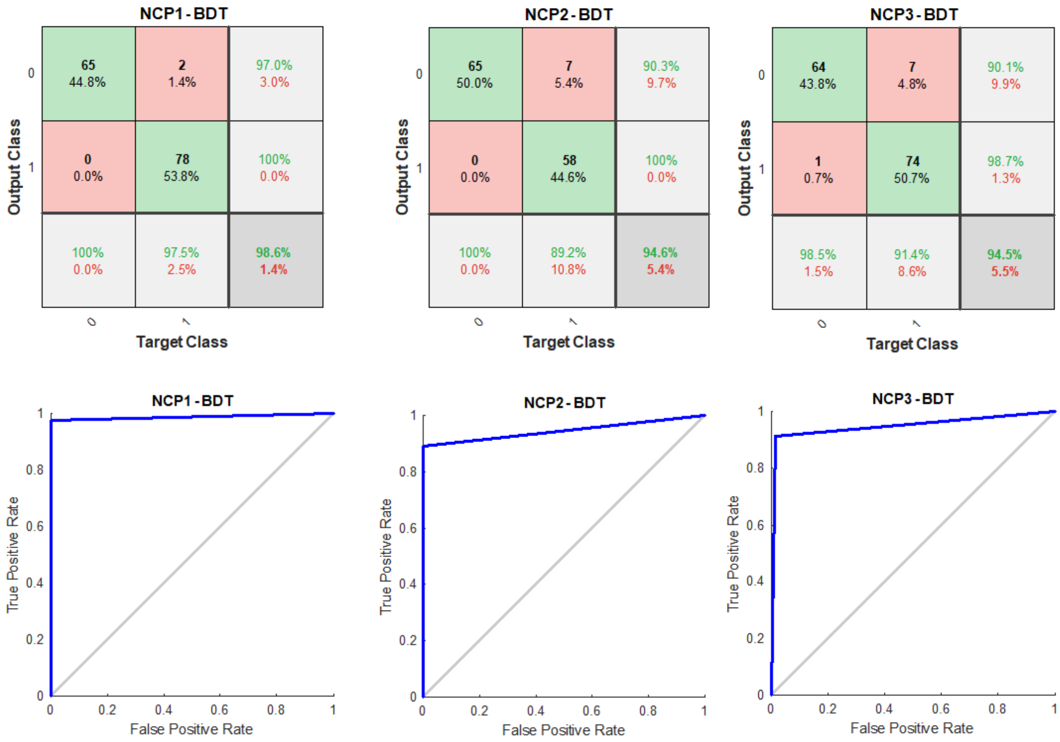


Figure 15: The performance of the BDT model at the final iteration of the integrated dataset with ER groups.

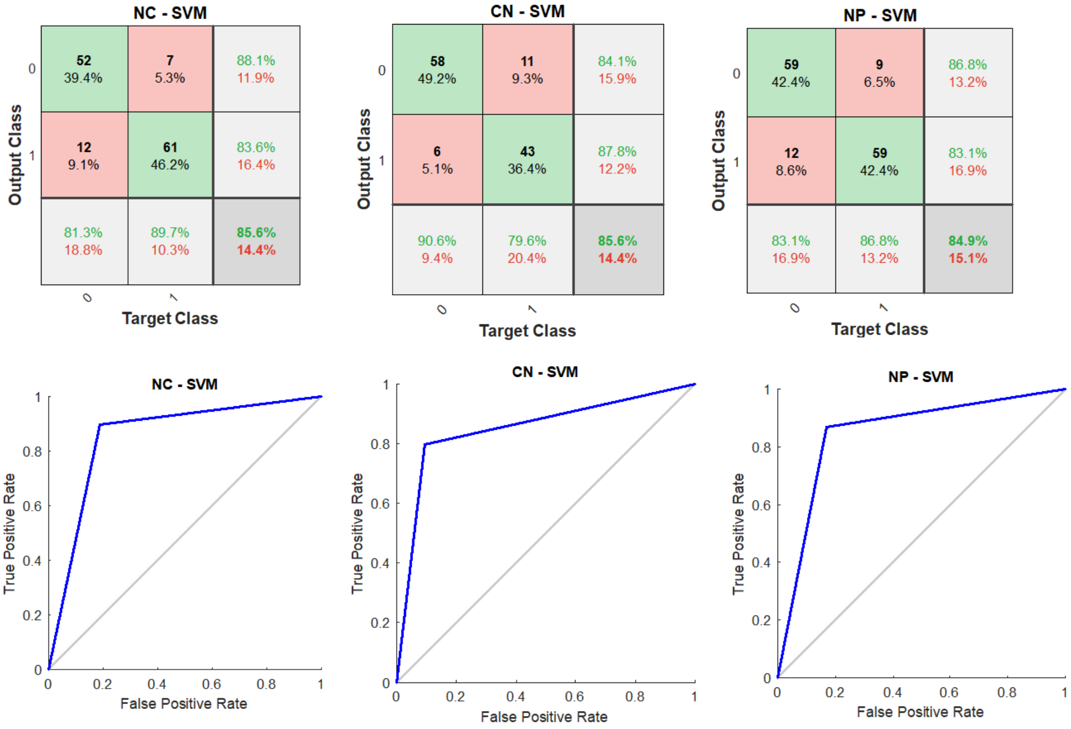
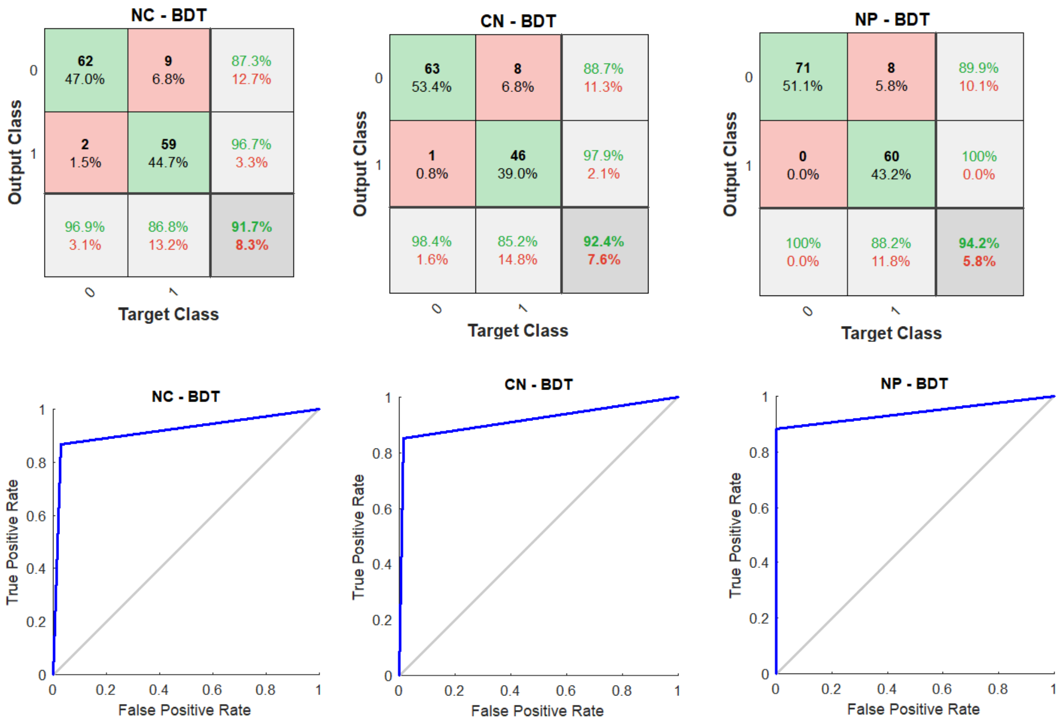


Figure 16: The performance of the SVM model at the final iteration for the integrated dataset with PR groups.



NC - BDT

| Output Class \ Target Class | 0 | 1 | |
|---|---|---|---|
| 0 | 62<br>47.0% | 9<br>6.8% | 87.3%<br>12.7% |
| 1 | 2<br>1.5% | 59<br>44.7% | 96.7%<br>3.3% |
| | 96.9%<br>3.1% | 86.8%<br>13.2% | 91.7%<br>8.3% |

CN - BDT

| Output Class \ Target Class | 0 | 1 | |
|---|---|---|---|
| 0 | 63<br>53.4% | 8<br>6.8% | 88.7%<br>11.3% |
| 1 | 1<br>0.8% | 46<br>39.0% | 97.9%<br>2.1% |
| | 98.4%<br>1.6% | 85.2%<br>14.8% | 92.4%<br>7.6% |

NP - BDT

| Output Class \ Target Class | 0 | 1 | |
|---|---|---|---|
| 0 | 71<br>51.1% | 8<br>5.8% | 89.9%<br>10.1% |
| 1 | 0<br>0.0% | 60<br>43.2% | 100%<br>0.0% |
| | 100%<br>0.0% | 88.2%<br>11.8% | 94.2%<br>5.8% |

Figure 17: The performance of the BDT model at the final iteration of the integrated dataset with PR groups.

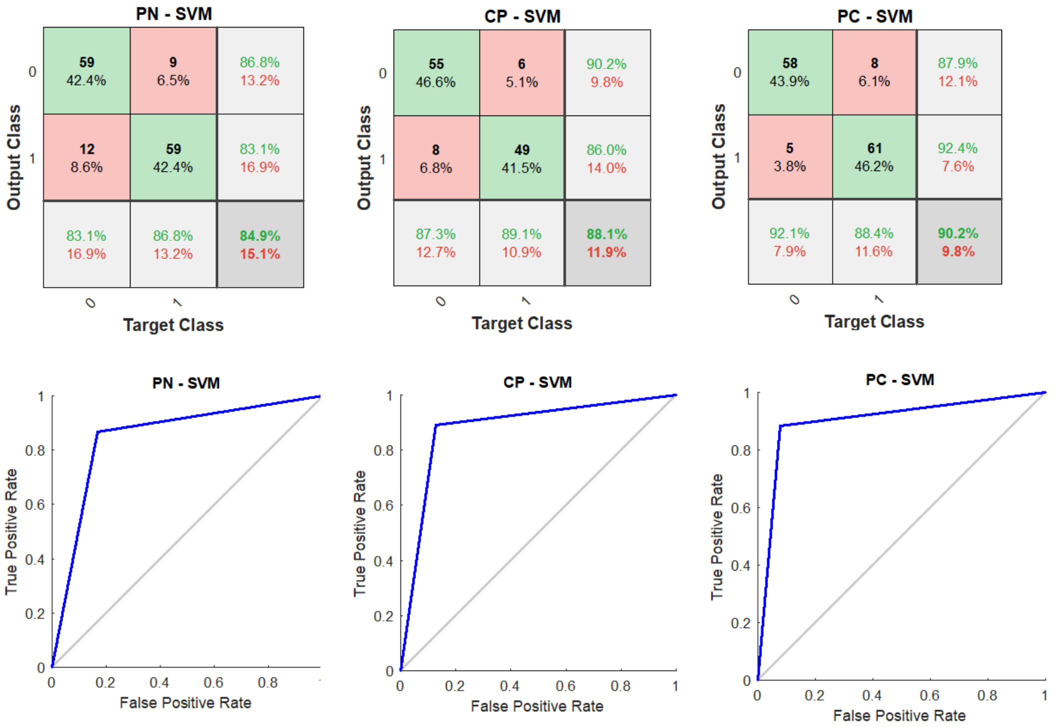


Figure 18: The performance of the SVM model at the final iteration of the integrated dataset with PR groups.

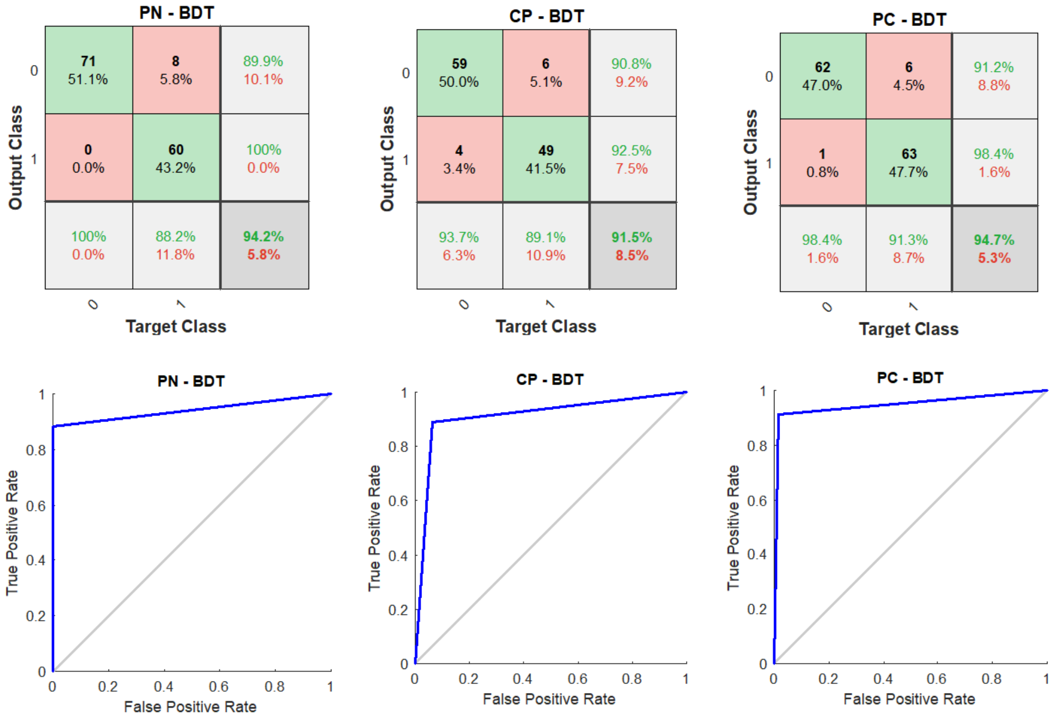


Figure 19: The performance of the BDT model at the final iteration of the integrated dataset with PR groups.

# References

[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, **22**[10]:1533–1545, 2014. 35

[2] Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, **26**[3]:392–398, 2009. 30, 32

[3] M Gordian Adam, Sonja Matt, Sven Christian, Holger Hess-Stumpp, Andrea Haegebarth, Thomas G Hofmann, and Carolyn Algire. Siah ubiquitin ligases regulate breast cancer cell migration and invasion independent of the oxygen status. *Cell Cycle*, **14**[23]:3734–3747, 2015. 135

[4] Paul J Adam, Joanne Berry, Julie A Loader, Kerry L Tyson, Graham Craggs, Paul Smith, Jackie De Belin, Graham Steers, Francesco Pezzella, Kris F Sachsenmeir, et al. Arylamine n-acetyltransferase-1 is highly expressed in breast cancers and conveys enhanced growth and resistance to etoposide in vitro. *Molecular cancer research*, **1**[11]:826–835, 2003. 138

[5] Heli I Alanen, Richard A Williamson, Mark J Howard, Anna-Kaisa Lappi, Heli P Jäntti, Sini M Rautio, Sakari Kellokumpu, and Lloyd W Ruddock. Functional characterization of erp18, a new endoplasmic reticulum-located thioredoxin superfamily member. *Journal of Biological Chemistry*, **278**[31]:28912–28920, 2003. 134

[6] Gabriela Alexe, G. S. Dalgin, Ramakrishna Ramaswamy, Charles DeLisi, and Gyan Bhanot. Data perturbation independent diagnosis and validation of breast cancer subtypes using clustering and patterns. *Cancer Informatics*, **2**:243 – 274, 2006. 169

[7] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, **33**[8]:831, 2015. 34

[8] Prasanna G Alluri, Corey Speers, and Arul M Chinnaiyan. Estrogen receptor mutations and their role in breast cancer progression. *Breast Cancer Research*, **16**[6]:494, 2014. 133

[9] Abeer Alzubaidi. Challenges in developing prediction models for multimodal high-throughput biomedical data. In *Proceedings of SAI Intelligent Systems Conference*, pages 1056–1069. Springer, 2018. 185

[10] Sarah A Andres and James L Wittliff. Co-expression of genes with estrogen receptor-$\alpha$ and progesterone receptor in human breast carcinoma tissue. *Hormone molecular biology and clinical investigation*, **12**[1]:377–390, 2012. 141

[11] Christof Angermueller, Heather Lee, Wolf Reik, and Oliver Stegle. Accurate prediction of single-cell dna methylation states using deep learning. *BioRxiv*, page 055715, 2017. 34

[12] Lindsay Angus, Nick Beije, Agnes Jager, John WM Martens, and Stefan Sleijfer. Esr1 mutations: Moving towards guiding treatment decision-making in metastatic breast cancer patients. *Cancer treatment reviews*, **52**:33–40, 2017. 133

[13] T Arentsen, Y Qian, S Gkotzis, T Femenia, T Wang, K Udekwu, H Forssberg, and R Diaz Heijtz. The bacterial peptidoglycan-sensing molecule Pglyrp2 modulates brain development and behavior. *Molecular Psychiatry*, **22**:257, nov 2016. 164

[14] Francisco Azuaje. *Bioinformatics and biomarker discovery.* Wiley Online Library, 2010. 1, 19

[15] YI Bao, Antao Wang, and Juanfen Mo. S100a8/a9 is associated with estrogen receptor loss in breast cancer. *Oncology letters*, **11**[3]:1936–1942, 2016. 155, 157

[16] Amira A Barkal, Rachel E Brewer, Maxim Markovic, Mark Kowarsky, Sammy A Barkal, Balyn W Zaro, Venkatesh Krishnan, Jason Hatakeyama, Oliver Dorigo, Layla J Barkal, et al. Cd24 signalling through macrophage siglec-10 is a target for cancer immunotherapy. *Nature*, page 1, 2019. 183

[17] Daniel H Barnett, Shubin Sheng, Tze Howe Charn, Abdul Waheed, William S Sly, Chin-Yo Lin, Edison T Liu, and Benita S Katzenellenbogen. Estrogen receptor regulation of carbonic anhydrase xii through a distal enhancer in breast cancer. *Cancer Research*, **68**[9]:3505–3515, 2008. 140

[18] Hamid Behravan, Jaana M Hartikainen, Maria Tengström, Katri Pylkäs, Robert Winqvist, Veli-Matti Kosma, and Arto Mannermaa. Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in finnish cases and controls. *Scientific reports*, **8**[1]:13149, 2018. 14

[19] Richard Bellman. *Dynamic Programming.* Princeton University Press, Princeton, NJ, USA, 1 edition, 1957. 2

[20] Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. Support vector machines and kernels for computational biology. *PLoS computational biology*, **4**[10]:e1000173, 2008. 24

[21] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36, 2012. 92

[22] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, **35**[8]:1798–1828, 2013. 34

[23] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, **2**[1]:1–127, 2009. 21

[24] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. pages 153–160, 2007. 29, 36, 93

[25] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995. 19

[26] Wendie A Berg. Tailored supplemental screening for breast cancer: what now and what next? *American Journal of Roentgenology*, **192**[2]:390–399, 2009. 15

[27] Caroline Bergenfelz, Alexander Gaber, Roni Allaoui, Meliha Mehmeti, Karin Jirström, Tomas Leanderson, and Karin Leandersson. S100a9 expressed in er- pgr- breast cancers induces inflammatory cytokines and is associated with an impaired overall survival. *British journal of cancer*, **113**[8]:1234–1243, 2015. 157

[28] Pablo Bermejo, Luis de la Ossa, José A Gámez, and José M Puerta. Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. *Knowledge-Based Systems*, **25**[1]:35–44, 2012. 31

[29] Sarah M Bernhardt, Pallave Dasari, David Walsh, Amanda R Townsend, Timothy J Price, and Wendy V Ingman. Hormonal modulation of breast cancer gene expression: implications for intrinsic subtyping in premenopausal women. *Frontiers in oncology*, **6**:241, 2016. 16

[30] Philippe Bertheau, Elisabeth Turpin, David S Rickman, Marc Espié, Aurélien De Reyniès, Jean-Paul Feugeas, Louis-François Plassa, Hany Soliman, Mariana Varna, Anne

De Roquancourt, et al. Exquisite sensitivity of tp53 mutant and basal breast cancers to a dose-dense epirubicin- cyclophosphamide regimen. *PLoS medicine*, **4**[3]:e90, 2007. 147

[31] Sunil Bhakta, Lisa M Crocker, Yvonne Chen, Meredith Hazen, Melissa M Schutten, Dongwei Li, Coenraad Kuijl, Rachana Ohri, Fiona Zhong, Kirsten A Poon, et al. An anti-gdnf family receptor alpha 1 (gfra1) antibody–drug conjugate for the treatment of hormone receptor–positive breast cancer. *Molecular cancer therapeutics*, **17**[3]:638–649, 2018. 134

[32] Verena E Bichsel, Lance A Liotta, et al. Cancer proteomics: from biomarker discovery to signal pathway profiling. *Cancer journal (Sudbury, Mass.)*, **7**[1]:69–78, 2001. 49, 93, 179

[33] SE Bleeker, HA Moll, EW Steyerberg, ART Donders, Gerarda Derksen-Lubsen, DE Grobbee, and KGM Moons. External validation is necessary in prediction research:: A clinical example. *Journal of clinical epidemiology*, **56**[9]:826–832, 2003. 45

[34] Emily E Bosco, R James Christie, Rosa Carrasco, Darrin Sabol, Jiping Zha, Karma DaCosta, Lee Brown, Maureen Kennedy, John Meekin, Sandrina Phipps, et al. Preclinical evaluation of a gfra1 targeted antibody-drug conjugate in breast cancer. *Oncotarget*, **9**[33]:22960, 2018. 134

[35] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992. 24

[36] Anne Boulay, Madlaina Breuleux, Christine Stephan, Caroline Fux, Cathrin Brisken, Maryse Fiche, Markus Wartmann, Michael Stumm, Heidi A Lane, and Nancy E Hynes. The ret receptor tyrosine kinase pathway functionally interacts with the er$\alpha$ pathway in breast cancer. *Cancer research*, **68**[10]:3743–3751, 2008. 134

[37] Anne Boulay, Catherine H Régnier, Patrick Anglard, Isabelle Stoll, Catherine Tomasetto, and Marie-Christine Rio. Transcription regulation and protein subcellular localization of the truncated basic hair keratin hhb1-$\delta$n in human breast cancer cells. *Journal of Biological Chemistry*, **276**[25]:22954–22964, 2001. 154

[38] Y-lan Boureau, Yann L Cun, et al. Sparse feature learning for deep belief networks. In *Advances in neural information processing systems*, pages 1185–1192, 2008. 93

[39] George EP Box and R Daniel Meyer. An analysis for unreplicated fractional factorials. *Technometrics*, **28**[1]:11–18, 1986. 29

[40] Norman F Boyd, Helen Guo, Lisa J Martin, Limei Sun, Jennifer Stone, Eve Fishell, Roberta A Jong, Greg Hislop, Anna Chiarelli, Salomon Minkin, et al. Mammographic density and the risk and detection of breast cancer. *New England Journal of Medicine*, **356**[3]:227–236, 2007. 15

[41] Ulisses M Braga-Neto and Edward R Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, **20**[3]:374–380, 2004. 42

[42] Leo Breiman. *Classification and regression trees*. Routledge, 2017. 28

[43] Frank Z Brill, Donald E Brown, and Worthy N Martin. Fast generic selection of features for neural network classifiers. *IEEE Transactions on Neural Networks*, **3**[2]:324–328, 1992. 33

[44] Tom Brosch, Roger Tam, Alzheimers Disease Neuroimaging Initiative, et al. Manifold learning of brain mris by deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 633–640. Springer, 2013. 34

[45] Samantha M Carlisle and David W Hein. Retrospective analysis of estrogen receptor 1 and n-acetyltransferase gene expression in normal breast tissue, primary breast tumors, and established breast cancer cell lines. *International journal of oncology*, **53**[2]:694–702, 2018. 138

[46] J S Carroll. Mechanisms of oestrogen receptor (ER) gene regulation in breast cancer. *European journal of endocrinology*, **175**[1]:R41–9, jul 2016. 16

[47] María Ángeles Castilla, María Ángeles López-García, María Reina Atienza, Juan Manuel Rosa-Rosa, Juan Díaz-Martín, María Luisa Pecero, Begona Vieites, Laura Romero-Pérez, Javier Benítez, Annarica Calcabrini, et al. Vgll1 expression is associated with a triple-negative basal-like phenotype in breast cancer. *Endocrine-related cancer*, **21**[4]:587–599, 2014. 146, 147, 152

[48] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, 2012. 51

[49] Peter Chan, Andreas Möller, Mira CP Liu, Jaclyn E Sceneay, Christina SF Wong, Nic Waddell, Katie T Huang, Alexander Dobrovic, Ewan KA Millar, Sandra A O'Toole, et al. The expression of the ubiquitin ligase siah2 (seven in absentia homolog 2) is mediated through gene copy number in breast cancer and is associated with a basal-like phenotype and p53 expression. *Breast Cancer Research*, **13**[1]:R19, 2011. 136

[50] Dingiun Chen, Keith CC Chan, and Xindong Wu. Gene expression analyses using genetic algorithm based hybrid approaches. In *Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on*, pages 963–969. IEEE, 2008. 33

[51] Jie-Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen, and Chung-Ming Chen. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. *Scientific reports*, **6**:24454, 2016. 34

[52] Meng Cheng, Stephanie Michalski, and Ramakrishna Kommagani. Role for growth regulation by estrogen in breast cancer 1 (greb1) in hormone-dependent cancers. *International journal of molecular sciences*, **19**[9]:2543, 2018. 165

[53] Wei-Yi Cheng, Tai-Hsien Ou Yang, and Dimitris Anastassiou. Biomolecular events in cancer revealed by attractor metagenes. *PLoS computational biology*, **9**[2]:e1002920, 2013. 161, 162

[54] Aaron Ciechanover. Proteolysis: from the lysosome to ubiquitin and the proteasome. *Nature reviews Molecular cell biology*, **6**[1]:79, 2005. 135

[55] Giovanni Ciriello, Michael L Gatza, Andrew H Beck, Matthew D Wilkerson, Suhn K Rhie, Alessandro Pastore, Hailei Zhang, Michael McLellan, Christina Yau, Cyriac Kandoth, et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, **163**[2]:506–519, 2015. 53

[56] Dearbhaile C Collins, Raghav Sundar, Joline SJ Lim, and Timothy A Yap. Towards precision medicine in the clinic: from biomarker discovery to novel therapeutics. *Trends in pharmacological sciences*, **38**[1]:25–40, 2017. 2

[57] Francis S Collins and Lawrence A Tabak. Nih plans to enhance reproducibility. *Nature*, **505**[7485]:612–613, 2014. 50

[58] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, **12**[Aug]:2493–2537, 2011. 35

[59] Krysta M Coyle, J Patrick Murphy, Dejan Vidovic, Ahmad Vaghar-Kashani, Cheryl A Dean, Mohammad Sultan, Derek Clements, Melissa Wallace, Margaret L Thomas, Amos Hundert, et al. Breast cancer subtype dictates dna methylation and aldh1a3-mediated expression of tumor suppressor rarres1. *Oncotarget*, **7**[28]:44096, 2016. 154

[60] Krysta M Coyle, Ahmad Vaghar-Kashani, Florence Wong, Cheryl Dean, Carman Giacomantonio, and Paola Marcato. Rarres1 is a tumor suppressor in triple-negative breast cancer cell lines, 2014. 154

[61] F H CRICK. On protein synthesis. *Symposia of the Society for Experimental Biology*, **12**:138–163, 1958. 17

[62] 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, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**[7403]:346, 2012. 62, 161

[63] SéRgio Francisco Da Silva, Marcela Xavier Ribeiro, João do ES Batista Neto, Caetano Traina-Jr, and Agma JM Traina. Improving the ranking quality of medical image retrieval using a genetic feature selection method. *Decision support systems*, **51**[4]:810–820, 2011. 33

[64] Simone U Dalm, Willemijne AME Schrijver, Anieta M Sieuwerts, Maxime P Look, Angelique CJ Ziel-Van Der Made, Vanja De Weerd, John W Martens, Paul J Van Diest, Marion De Jong, and Carolien HM Van Deurzen. Prospects of targeting the gastrin releasing peptide receptor and somatostatin receptor 2 for nuclear imaging and therapy in metastatic breast cancer. *PloS one*, **12**[1]:e0170536, 2017. 164

[65] Simone U Dalm, Anieta M Sieuwerts, Maxime P Look, Marleen Melis, Carolien HM van Deurzen, John A Foekens, Marion de Jong, and John WM Martens. Clinical relevance of targeting the gastrin-releasing peptide receptor, somatostatin receptor 2, or chemokine cxc motif receptor 4 in breast cancer for imaging and therapy. *Journal of Nuclear Medicine*, **56**[10]:1487–1493, 2015. 164

[66] Padideh Danaee, Reza Ghaeini, and David A Hendrix. A deep learning approach for cancer detection and relevant gene identification.

In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*, pages 219–229. World Scientific, 2017. 38

[67] Chad A Davis, Fabian Gerick, Volker Hintermair, Caroline C Friedel, Katrin Fundel, Robert Küffner, and Ralf Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, **22**[19]:2356–2363, 2006. 45

[68] Sarah-Jane Dawson, Oscar M Rueda, Samuel Aparicio, and Carlos Caldas. A new genome-driven integrated classification of breast cancer and its implications. *The EMBO journal*, **32**[5]:617–628, 2013. 62

[69] Kenneth A De Jong. *Evolutionary computation: a unified approach*. MIT press, 2006. 33, 69

[70] Beatriz De La Iglesia. Evolutionary computation for feature selection in classification problems. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **3**[6]:381–407, 2013. 33

[71] Sandra L Deming, Zefang Ren, Wanqing Wen, Xiao Ou Shu, Qiuyin Cai, Yu-Tang Gao, and Wei Zheng. Genetic variation in igf1, igf-1r, igfals, and igfbp3 in breast cancer survival among chinese women: a report from the shanghai breast cancer study. *Breast cancer research and treatment*, **104**[3]:309–319, 2007. 144

[72] Meaza Demissie, Barbara Mascialino, Stefano Calza, and Yudi Pawitan. Unequal group variances in microarray data analyses. *Bioinformatics*, **24**[9]:1168–1174, 2008. 20

[73] Li Deng and Xiao Li. Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, **21**[5]:1060–1089, 2013. 35

[74] Horacio M Domené, Vivian Hwa, Jesús Argente, Jan M Wit, Cecilia Camacho-Hübner, Héctor G Jasper, Jesús Pozo, Hermine A Van Duyvenvoorde, Shoshana Yakar, Olga V Fofanova-Gambetti, et al. Human acid-labile subunit deficiency: clinical, en-

docrine and metabolic consequences. *Hormone Research in Paediatrics*, **72**[3]:129–141, 2009. 144

[75] David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, **1**[2000]:32, 2000. 20

[76] Marco Dorigo and Mauro Birattari. Ant colony optimization. In *Encyclopedia of machine learning*, pages 36–39. Springer, 2011. 33

[77] Marco Dorigo and Gianni Di Caro. Ant colony optimization: a new meta-heuristic. In *Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406)*, **2**, pages 1470–1477. IEEE, 1999. 33

[78] Feng Du, Peng Yuan, Teng Wang, Jiuda Zhao, Zitong Zhao, Yang Luo, and Binghe Xu. The significance and therapeutic potential of gata3 expression and mutation in breast cancer: a systematic review. *Medicinal research reviews*, **35**[6]:1300–1315, 2015. 144

[79] 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, 2002. 74

[80] Anita K Dunbier, Helen Anderson, Zara Ghazoui, Elena Lopez-Knowles, Sunil Pancholi, Ricardo Ribas, Suzanne Drury, Kally Sidhu, Alexandra Leary, Lesley-Ann Martin, et al. Esr1 is co-expressed with closely adjacent uncharacterised genes spanning a breast cancer susceptibility locus at 6q25. 1. *PLoS genetics*, **7**[4]:e1001382, 2011. 136

[81] Kevin Dunne, Padraig Cunningham, and Francisco Azuaje. Solutions to instability problems with sequential wrapper-based approaches to feature selection. *Journal of Machine Learning Research*, pages 1–22, 2002. 45

[82] Béatrice Duval and Jin-Kao Hao. Advances in metaheuristics for gene selection and classification of microarray data. *Briefings in bioinformatics*, **11**[1]:127–141, 2009. 30

[83] Darius M Dziuda. *Data mining for genomics and proteomics: analysis of gene and protein expression data*, **1**. John Wiley & Sons, 2010. 18

[84] Russell Eberhart and James Kennedy. Particle swarm optimization. In *Proceedings of the IEEE international conference on neural networks*, **4**, pages 1942–1948. Citeseer, 1995. 33

[85] Emad Elbeltagi, Tarek Hegazy, and Donald Grierson. Comparison among five evolutionary-based optimization algorithms. *Advanced engineering informatics*, **19**[1]:43–53, 2005. 33

[86] Yumi Endo, Hiroko Yamashita, Satoru Takahashi, Shinya Sato, Nobuyasu Yoshimoto, Tomoko Asano, Yukari Hato, Yu Dong, Yoshitaka Fujii, and Tatsuya Toyama. Immunohistochemical determination of the mir-1290 target arylamine n-acetyltransferase 1 (nat1) as a prognostic biomarker in breast cancer. *Bmc Cancer*, **14**[1]:990, 2014. 139

[87] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, **11**[Feb]:625–660, 2010. 35, 93, 100

[88] S Esseghir, JS Reis-Filho, A Kennedy, M James, MJ O'hare, R Jeffery, R Poulsom, and CM Isacke. Identification of transmembrane proteins as potential prognostic markers and therapeutic targets in breast cancer by a screen for signal sequence encoding transcripts. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, **210**[4]:420–430, 2006. 134

[89] Selma Esseghir, S Katrina Todd, Toby Hunt, Richard Poulsom, Ivan Plaza-Menacho, Jorge S Reis-Filho, and Clare M Isacke. A role for glial cell–derived neurotrophic factor–induced expression by inflammatory cytokines and ret/gfr$\alpha$1 receptor up-regulation in breast cancer. *Cancer research*, **67**[24]:11732–11741, 2007. 134

[90] Ruth Etzioni, Nicole Urban, Scott Ramsey, Martin McIntosh, Stephen Schwartz, Brian Reid, Jerald Radich, Garnet Anderson, and Leland Hartwell. Early detection: The case for early detection. *Nature Reviews Cancer*, **3**[4]:243, 2003. 15

[91] Cheng Fan, Daniel S Oh, Lodewyk Wessels, Britta Weigelt, Dimitry SA Nuyten, Andrew B Nobel, Laura J Van't Veer, and Charles M Perou. Concordance among gene-expression–based predictors for breast cancer. *New England Journal of Medicine*, **355**[6]:560–569, 2006. 141

[92] Jianqing Fan and Runze Li. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133*, 2006. 2

[93] Marta Faryna, Carolin Konermann, Sebastian Aulmann, Justo Lorenzo Bermejo, Markus Brugger, Sven Diederichs, Joachim Rom, Dieter Weichenhan, Rainer Claus, Michael Rehli, et al. Genome-wide methylation screen in low-grade breast cancer identifies novel epigenetically altered genes as potential biomarkers for tumor diagnosis. *The FASEB Journal*, **26**[12]:4937–4950, 2012. 169

[94] Artur J Ferreira and MáRio AT Figueiredo. Efficient feature selection filters for high-dimensional data. *Pattern Recognition Letters*, **33**[13]:1794–1804, 2012. 30

[95] Graeme C Fielder, Teresa Wen-Shan Yang, Mahalakshmi Razdan, Yan Li, Jun Lu, Jo K Perry, Peter E Lobie, and Dong-Xu Liu. The gdnf family: A role in cancer? *Neoplasia*, **20**[1]:99–117, 2018. 134

[96] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, **7**[2]:179–188, 1936. 20, 73

[97] Patrick L Fitzgibbons, Douglas A Murphy, M Elizabeth H Hammond, D Craig Allred, and Paul N Valenstein. Recommendations for validating estrogen and progesterone receptor immunohistochem-

istry assays. *Archives of pathology & laboratory medicine*, **134**[6]:930–935, 2010. 16

[98] Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013. 97

[99] MA Forget, S Turcotte, D Beauseigle, J Godin-Ethier, S Pelletier, J Martin, S Tanguay, and R Lapointe. The wnt pathway regulator dkk1 is preferentially expressed in hormone-resistant breast tumours and in some common cancer types. *British journal of cancer*, **96**[4]:646–653, 2007. 154

[100] Vittorio Fortino, Pia Kinaret, Nanna Fyhrquist, Harri Alenius, and Dario Greco. A robust and accurate method for feature selection and prioritization from multi-class omics data. *PloS one*, **9**[9]:e107801, 2014. 10, 32, 45

[101] Alex A Freitas. *Data mining and knowledge discovery with evolutionary algorithms*. Springer Science & Business Media, 2013. 33

[102] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci. Signal.*, **6**[269]:pl1–pl1, 2013. 51

[103] Song Gao, Anqi Ge, Shouping Xu, Zilong You, Shipeng Ning, Yashuang Zhao, and Da Pang. Psat1 is regulated by atf4 and enhances cell proliferation via the gsk3$\beta$/$\beta$-catenin/cyclin d1 signaling pathway in er-negative breast cancer. *Journal of Experimental & Clinical Cancer Research*, **36**[1]:179, 2017. 148

[104] Montserrat Garcia-Closas and Stephen Chanock. Genetic susceptibility loci for breast cancer by estrogen receptor status. *Clinical Cancer Research*, **14**[24]:8000–8009, 2008. 16

[105] Stefan Garczyk, Saskia von Stillfried, Wiebke Antonopoulos, Arndt Hartmann, Michael G Schrauder, Peter A Fasching,

Tobias Anzeneder, Andrea Tannapfel, Yavuz Ergönenc, Ruth Knüchel, et al. Agr3 in breast cancer: Prognostic impact and suitable serum-based biomarker for early cancer detection. *PloS one*, **10**[4]:e0122106, 2015. 134

[106] Robert Gentleman, Vincent Carey, Wolfgang Huber, Rafael Irizarry, and Sandrine Dudoit. *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer Science & Business Media, 2006. 49

[107] Filippo Geraci, Marco Pellegrini, and M Elena Renda. Amic@: all microarray clusterings@ once. *Nucleic acids research*, **36**[suppl_2]:W315–W319, 2008. 21

[108] Sabine Gesierich, Claudia Paret, Dagmar Hildebrand, Jürgen Weitz, Kaspar Zgraggen, Friedrich H Schmitz-Winnenthal, Vaclav Horejsi, Osamu Yoshie, Dorothee Herlyn, Leonie K Ashman, et al. Colocalization of the tetraspanins, co-029 and cd151, with integrins in human pancreatic adenocarcinoma: impact on cell motility. *Clinical cancer research*, **11**[8]:2840–2852, 2005. 174

[109] Philip E Gill and Walter Murray. Safeguarded steplength algorithms for optimization using descent methods. 1974. 97

[110] Wilson Wen Bin Goh and Limsoon Wong. Evaluating feature-selection stability in next-generation proteomics. *Journal of bioinformatics and computational biology*, **14**[05]:1650029, 2016. 44

[111] David E Goldberg and John H Holland. Genetic algorithms and machine learning. *Machine learning*, **3**[2]:95–99, 1988. 33

[112] Abhishek Golugula, George Lee, and Anant Madabhushi. Evaluating feature selection strategies for high dimensional, small sample size datasets. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 949–952. IEEE, 2011. 45

[113] Royston Goodacre, Seetharaman Vaidyanathan, Warwick B Dunn, George G Harrigan, and Douglas B Kell. Metabolomics by numbers: acquiring and understanding global metabolite data. *TRENDS in Biotechnology*, **22**[5]:245–252, 2004. 18

[114] Peter C Gøtzsche and Margrethe Nielsen. Screening for breast cancer with mammography. *Cochrane database of systematic reviews*, [1], 2011. 14

[115] Christian J Gröger, Markus Grubinger, Thomas Waldhör, Klemens Vierlinger, and Wolfgang Mikulits. Meta-analysis of gene expression signatures defining the epithelial to mesenchymal transition during cancer progression. *PloS one*, **7**[12]:e51136, 2012. 168

[116] Biomarkers Definitions Working Group, Arthur J Atkinson Jr, Wayne A Colburn, Victor G DeGruttola, David L DeMets, Gregory J Downing, Daniel F Hoth, John A Oates, Carl C Peck, Robert T Schooley, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, **69**[3]:89–95, 2001. 1

[117] Achim D. Gruber and Bendicht U. Pauli. Tumorigenicity of human breast cancer is associated with loss of the ca2+-activated chloride channel clca2. **59**[21]:5488–5491, 1999. 170

[118] Sofia Gruvberger, Markus Ringnér, Yidong Chen, Sujatha Panavally, Lao H. Saal, Åke Borg, Mårten Fernö, Carsten Peterson, and Paul S. Meltzer. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Research*. 168

[119] Martin Gutlein, Eibe Frank, Mark Hall, and Andreas Karwath. Large-scale attribute selection using wrappers. In *2009 IEEE symposium on computational intelligence and data mining*, pages 332–339. IEEE, 2009. 31

[120] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, **3**[Mar]:1157–1182, 2003. 30

[121] Francis H. C. Crick. Central dogma of molecular biology. **227**:561–3, 09 1970. xi, 17, 18

[122] Nils Y Hammerla, Shane Halloran, and Thomas Ploetz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016. 35

[123] M Elizabeth H Hammond, Daniel F Hayes, Mitch Dowsett, D Craig Allred, Karen L Hagerty, Sunil Badve, Patrick L Fitzgibbons, Glenn Francis, Neil S Goldstein, Malcolm Hayes, et al. American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version). *Archives of pathology & laboratory medicine*, **134**[7]:e48–e72, 2010. 16

[124] Xianlin Han. Neurolipidomics: challenges and developments. *Frontiers in bioscience: a journal and virtual library*, **12**:2601, 2007. 19

[125] David J Hand. Classifier technology and the illusion of progress. *Statistical science*, pages 1–14, 2006. 20

[126] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one*, **6**[12]:e28210, 2011. 31

[127] Daniel L Hertz, N Lynn Henry, Kelley M Kidwell, Dafydd Thomas, Audrey Goddard, Faouzi Azzouz, Kelly Speth, Lang Li, Mousumi Banerjee, Jacklyn N Thibert, et al. Esr1 and pgr polymorphisms affect estrogen and progesterone receptor expression in breast tumors. *American Journal of Physiology-Heart and Circulatory Physiology*, 2016. 167

[128] Magnus Rudolph Hestenes. *Conjugate direction methods in optimization*, **12**. Springer Science & Business Media, 2012. 97

[129] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, **29**[6]:82–97, 2012. 35

[130] Geoffrey E Hinton. Connectionist learning procedures. In *Machine Learning, Volume III*, pages 555–610. Elsevier, 1990. 34

[131] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, **18**[7]:1527–1554, 2006. 36

[132] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, **313**[5786]:504–507, 2006. 37

[133] Kendra Hodgkinson, Laura A Forrest, Nhung Vuong, Kenneth Garson, Bojana Djordjevic, and Barbara C Vanderhyden. Greb1 is an estrogen receptor-regulated tumour promoter that is frequently expressed in ovarian cancer. *Oncogene*, **37**[44]:5873, 2018. 165

[134] John H Holland. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, 1975. 33, 69

[135] John Henry Holland. Nonlinear environments permitting efficient adaptation. 1967. 33

[136] Frederik Holst, Cathy B Moelans, Martin Filipits, Christian F Singer, Ronald Simon, and Paul J Van Diest. On the evidence for esr1 amplification in breast cancer. *Nature Reviews Cancer*, **12**[2]:149, 2012. 132

[137] Regina J Hooley, Liva Andrejeva, and Leslie M Scoutt. Breast cancer screening and problem solving using mammography, ultrasound, and magnetic resonance imaging. *Ultrasound quarterly*, **27**[1]:23–47, 2011. 15

[138] Richard P Horgan and Louise C Kenny. omictechnologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, **13**[3]:189–195, 2011. 17

[139] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, **24**[6]:417, 1933. 20

[140] Li-De Hu, Hua-Fei Zou, Shu-Xuan Zhan, and Kai-Ming Cao. Evl (ena/vasp-like) expression is up-regulated in human breast cancer and its relative expression level is correlated with clinical stages. *Oncology reports*, **19**[4]:1015–1020, 2008. 141

[141] Zhi Hu, Richard Neve, Yinghui Guan, and Joe Gray. Identification of new therapeutic targets of breast cancer using sirna technology, 2007. 138

[142] Jianping Hua, Waibhav D Tembe, and Edward R Dougherty. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, **42**[3]:409–424, 2009. 30

[143] Jinjie Huang, Yunze Cai, and Xiaoming Xu. A hybrid genetic algorithm for feature selection wrapper based on mutual information. *Pattern Recognition Letters*, **28**[13]:1825–1844, 2007. 33

[144] Iñaki Inza, Borja Calvo, Rubén Armañanzas, Endika Bengoetxea, Pedro Larrañaga, and José A Lozano. Machine learning: an indispensable tool in bioinformatics. In *Bioinformatics methods in clinical research*, pages 25–48. Springer, 2010. 30

[145] John PA Ioannidis. Microarrays and molecular research: noise discovery? *The Lancet*, **365**[9458]:454–455, 2005. 44

[146] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031*, 2019. 8

[147] Alison J Camden, Maria M Szwarc, Sangappa Chadchan, Francesco Demayo, Bert W O 'malley, John P Lydon, and Ramakrishna Kommagani. Growth regulation by estrogen in breast cancer 1 (greb1) is a novel progesterone-responsive gene required for human endometrial stromal decidualization. *Molecular Human Reproduction*, **23**:1–8, 08 2017. 165

[148] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, **112**. Springer, 2013. 42

[149] Maurice PHM Jansen, Kirsten Ruigrok-Ritstier, Lambert CJ Dorssers, Iris L van Staveren, Maxime P Look, Marion E Meijer-van Gelder, Anieta M Sieuwerts, Jozien Helleman, Stefan Sleijfer, Jan GM Klijn, et al. Downregulation of siah2, an ubiquitin e3 ligase, is associated with resistance to endocrine therapy in breast cancer. *Breast cancer research and treatment*, **116**[2]:263–271, 2009. 136

[150] Kevin Jarrett, Koray Kavukcuoglu, Yann LeCun, et al. What is the best multi-stage architecture for object recognition? In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2146–2153. IEEE, 2009. 37

[151] Simon A. Joosse, Juliane Hannemann, Julia Spötter, Andreas Bauche, Antje Andreas, Volkmar Müller, and Klaus Pantel. Changes in keratin expression during metastatic progression of breast cancer: Impact on the detection of circulating tumor cells. *Clinical Cancer Research*, **18**[4]:993–1003, 2012. 151

[152] Andrew R Joyce and Bernhard Ø Palsson. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*, **7**:198–210, 2006 Mar 2006. 17

[153] Giuseppe Jurman, Stefano Merler, Annalisa Barla, Silvano Paoli, Antonio Galea, and Cesare Furlanello. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics*, **24**[2]:258–264, January 2008. 44

[154] Md Monirul Kabir, Md Shahjahan, and Kazuyuki Murase. A new local search based hybrid genetic algorithm for feature selection. *Neurocomputing*, **74**[17]:2914–2928, 2011. 33

[155] Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, **12**[1]:95–116, 2007. 45

[156] Jessica Kao, Keyan Salari, Melanie Bocanegra, Yoon-La Choi, Luc Girard, Jeet Gandhi, Kevin A Kwei, Tina Hernandez-Boussard, Pei Wang, Adi F Gazdar, et al. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PloS one*, **4**[7]:e6146, 2009. 147

[157] Mariz Kasoha, Rainer M Bohle, Anita Seibold, Christoph Gerlinger, Ingolf Juhasz-Boess, and Erich-Franz Solomayer. Dickkopf-1 (dkk1) protein expression in breast cancer with special reference to bone metastases. *Clinical & experimental metastasis*, **35**[8]:763–775, 2018. 154

[158] Grit Kasper, Armin A Weiser, Andreas Rump, Katrin Sparbier, Edgar Dahl, Arndt Hartmann, Peter Wild, Uta Schwidetzky, Esmeralda Castaños-Vélez, and Kerstin Lehmann. Expression levels of the putative zinc transporter liv-1 are associated with a better outcome of breast cancer patients. *International journal of cancer*, **117**[6]:961–973, 2005. 138

[159] Hiroyuki Katayama, Clayton Boldt, Jon J Ladd, Melissa M Johnson, Timothy Chao, Michela Capello, Jinfeng Suo, Jianning Mao, JoAnn E Manson, Ross Prentice, et al. An autoimmune response signature associated with the development of triple-negative breast cancer reflects disease pathogenesis. *Cancer research*, **75**[16]:3246–3254, 2015. 151

[160] Benita S Katzenellenbogen, Monica M Montano, Kirk Ekena, Mary E Herman, and Eileen M McInerney. Antiestrogens: mecha-

nisms of action and resistance in breast cancer. *Breast cancer research and treatment*, **44**[1]:23–38, 1997. 16

[161] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 2016. 34

[162] Brian M Kennedy and Randall E Harris. Cyclooxygenase and lipoxygenase gene expression in the inflammogenesis of breast cancer. *Inflammopharmacology*, **26**[4]:909–923, 2018. 149

[163] James Kennedy. Particle swarm optimization. In *Encyclopedia of machine learning*, pages 760–766. Springer, 2011. 33

[164] Dan Koboldt. The future of cancer genomics. *Clinical OMICs*, **1**[7]:8–10, 2014. 50

[165] Pang Wei Koh, Emma Pierson, and Anshul Kundaje. Denoising genome-wide histone chip-seq with convolutional neural networks. *Bioinformatics*, **33**[14]:i225–i233, 2017. 34

[166] Isaac S Kohane, Atul J Butte, and Alvin Kho. *Microarrays for an integrative genomics*. MIT press, 2002. 49

[167] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, **97**[1-2]:273–324, 1997. 31

[168] Teuvo Kohonen. Exploration of very large databases by self-organizing maps. In *Proceedings of International Conference on Neural Networks (ICNN'97)*, **1**, pages PL1–PL6. IEEE, 1997. 21

[169] Nadezda V Kovalevskaya, Charlotte Whicher, Timothy D Richardson, Craig Smith, Jana Grajciarova, Xocas Cardama, José Moreira, Adrian Alexa, Amanda A McMurray, and Fiona GG Nielsen. Dnadigest and repositive: connecting the world of genomic data. *PLoS biology*, **14**[3]:e1002418, 2016. 50

[170] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 35

[171] Yanan Kuang, Bilal Siddiqui, Jiani Hu, Matthew Pun, MacIntosh Cornwell, Gilles Buchwalter, Melissa E Hughes, Nikhil Wagle, Paul Kirschmeier, Pasi A Jänne, et al. Unraveling the clinicopathological features driving the emergence of esr1 mutations in metastatic breast cancer. *NPJ breast cancer*, **4**[1]:22, 2018. 133

[172] Ludmila I Kuncheva and Lakhmi C Jain. Nearest neighbor classifier: Simultaneous editing and feature selection. *Pattern recognition letters*, **20**[11-13]:1149–1156, 1999. 33

[173] Miron Bartosz Kursa. Robustness of random forest-based gene selection methods. *BMC bioinformatics*, **15**[1]:8, 2014. 32

[174] Anne-Vibeke Laenkholm, Ann Knoop, Bent Ejlertsen, Tine Rudbeck, Maj-Britt Jensen, Sven Müller, Anne Elisabeth Lykkesfeldt, Birgitte Bruun Rasmussen, and Kirsten Vang Nielsen. Esr1 gene status correlates with estrogen receptor protein levels measured by ligand binding assay and immunohistochemistry. *Molecular oncology*, **6**[4]:428–436, 2012. 133

[175] Thomas Navin Lal, Olivier Chapelle, Jason Weston, and André Elisseeff. Embedded methods. In *Feature extraction*, pages 137–165. Springer, 2006. 30

[176] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, pages 473–480. ACM, 2007. 93

[177] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, **521**[7553]:436, 2015. 34, 100

[178] Jonathan T Lei, Jieya Shao, Jin Zhang, Michael Iglesia, Doug W Chan, Jin Cao, Meenakshi Anurag, Purba Singh, Xiaping He, Yoshimasa Kosaka, et al. Functional annotation of esr1 gene fusions in estrogen receptor-positive breast cancer. *Cell reports*, **24**[6]:1434–1444, 2018. 133

[179] CHRISTOPHE Lemetre. *Artificial neural network techniques to investigate potential interactions between biomarkers*. PhD thesis, Nottingham Trent University, 2010. 153

[180] K M Levine, K Ding, N Priedigkeit, M J Sikora, N Tasdemir, L Zhu, G C Tseng, R C Jankowitz, D J Dabbs, P F McAuliffe, A V Lee, and S Oesterreich. Abstract P5-04-21: FGFR4 is a novel druggable target for recurrent ER-positive breast cancers. *Cancer Research*, **79**[4 Supplement]:P5–04–21—-P5–04–21, 2019. 172

[181] Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**[9]:1175–1182, 2008. 20

[182] Cong Li, Fang Zhang, Meihong Lin, and Jingwen Liu. Induction of s100a9 gene expression by cytokine oncostatin m in breast cancer cells through the stat3 signaling cascade. *Breast cancer research and treatment*, **87**[2]:123–134, 2004. 157

[183] Lihua Li, Li Chen, D Goldgof, F George, Z Chen, A Rao, J Cragun, R Sutphen, and Johnathan M Lancaster. Integration of clinical information and gene expression profiles for prediction of chemo-response for ovarian cancer. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 4818–4821. IEEE, 2006. 20

[184] Xiurong Li, John K Cowell, and Khalid Sossey-Alaoui. CLCA2 tumour suppressor gene in 1p31 is epigenetically regulated in breast cancer. *Oncogene*, **23**:1474, feb 2004. 170

[185] Yang Li, Xu-Qing Tang, Zhonghu Bai, and Xiaofeng Dai. Exploring the intrinsic differences among breast tumor subtypes defined using

immunohistochemistry markers based on the decision tree. *Scientific reports*, **6**:35773, 2016. 140, 146

[186] Elgene Lim, François Vaillant, Di Wu, Natasha C Forrest, Bhupinder Pal, Adam H Hart, Marie-Liesse Asselin-Labat, David E Gyorki, Teresa Ward, Audrey Partanen, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in brca1 mutation carriers. *Nature medicine*, **15**[8]:907, 2009. 146, 147

[187] Yuh-Charn Lin, Yi-Ching Lee, Ling-Hui Li, Chien-Jui Cheng, and Ruey-Bing Yang. Tumor suppressor scube2 inhibits breast-cancer cell migration and invasion through the reversal of epithelial–mesenchymal transition. *J Cell Sci*, **127**[1]:85–100, 2014. 141

[188] Charles X Ling, Jin Huang, and Harry Zhang. Auc: a better measure than accuracy in comparing learning algorithms. In *Conference of the canadian society for computational studies of intelligence*, pages 329–341. Springer, 2003. 64

[189] Charles X Ling, Jin Huang, Harry Zhang, et al. Auc: a statistically consistent and more discriminating measure than accuracy. In *IJCAI*, **3**, pages 519–524, 2003. 64

[190] Caigang Liu, Lisha Sun, Jie Yang, Tong Liu, Yongliang Yang, Se-Min Kim, Xunyan Ou, Yining Wang, Li Sun, Mone Zaidi, et al. Fsip1 regulates autophagy in breast cancer. *Proceedings of the National Academy of Sciences*, **115**[51]:13075–13080, 2018. 142

[191] DaoTong Liu, Z Peng, JingYan Han, Fan-Zhong Lin, Xian-Min Bu, and Qing-Xia Xu. Clinical and prognostic significance of sox11 in breast cancer. *Asian Pacific journal of cancer prevention: APJCP*, **15**[13]:5483–5486, 2014. 150

[192] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, **17**[4]:491–502, 2005. 31

[193] Huiqing Liu, Jinyan Li, and Limsoon Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome informatics*, **13**:51–60, 2002. 19

[194] Peipei Liu, Wenhui Li, Yuanyuan Hu, and Youhong Jiang. Absence of AIF1L contributes to cell migration and a poor prognosis of breast cancer. *OncoTargets and therapy*, **11**:5485–5498, sep 2018. 169

[195] Siqi Liu, Sidong Liu, Weidong Cai, Sonia Pujol, Ron Kikinis, and Dagan Feng. Early diagnosis of alzheimer's disease with deep learning. In *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, pages 1015–1018. IEEE, 2014. 34

[196] Tong Liu, Hao Zhang, Li Sun, Danyu Zhao, Peng Liu, Meisi Yan, Neeha Zaidi, Sudeh Izadmehr, Animesh Gupta, Wahid Abu-Amer, Minna Luo, Jie Yang, Xunyan Ou, Yining Wang, Xuefeng Bai, Yan Wang, Maria I. New, Mone Zaidi, Tony Yuen, and Caigang Liu. Fsip1 binds her2 directly to regulate breast cancer growth and invasiveness. *Proceedings of the National Academy of Sciences*, **114**[29]:7683–7688, 2017. 142

[197] Felipe Llinares López. *Significant Pattern Mining for Biomarker Discovery*. PhD thesis, ETH Zurich, 2018. 19

[198] Ángeles López-López, Ángeles López-Gonzálvez, Tomás Clive Barker-Tejeda, and Coral Barbas. A review of validated biomarkers obtained through metabolomics. *Expert review of molecular diagnostics*, **18**[6]:557–575, 2018. 45

[199] Aoife J Lowery, Nicola Miller, Amanda Devaney, Roisin E McNeill, Pamela A Davoren, Christophe Lemetre, Vladimir Benes, Sabine Schmidt, Jonathon Blake, Graham Ball, et al. Microrna signatures predict oestrogen receptor, progesterone receptor and her2/neu receptor status in breast cancer. *Breast cancer research*, **11**[3]:R27, 2009. 16

[200] Edward M. Rubin, Susan Lucas, Paul Richardson, Daniel Rokhsar, and Len Pennacchio. Finishing the euchromatic sequence of the human genome. **431**, 09 2004. 2

[201] Shuangge Ma and Jian Huang. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, **9**[5]:392–403, 2008. 30

[202] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, **45**[D1]:D896–D901, 2016. 20

[203] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, **1**, pages 281–297. Oakland, CA, USA, 1967. 21

[204] James D Malley, Karen G Malley, and Sinisa Pajevic. *Statistical learning for biomedical data*. Cambridge University Press, 2011. 20

[205] Emily B Mardian, Ryan M Bradley, and Robin E Duncan. The hrasls (pla/at) subfamily of enzymes. *Journal of biomedical science*, **22**[1]:99, 2015. 157

[206] John WM Martens, Inko Nimmrich, Thomas Koenig, Maxime P Look, Nadia Harbeck, Fabian Model, Antje Kluth, Joan Bolt-de Vries, Anieta M Sieuwerts, Henk Portengen, et al. Association of dna methylation of phosphoserine aminotransferase with response to endocrine therapy in patients with recurrent breast cancer. *Cancer research*, **65**[10]:4101–4117, 2005. 149

[207] Lesley-Ann Martin, Ricardo Ribas, Nikiana Simigdala, Eugene Schuster, Sunil Pancholi, Tencho Tenev, Pascal Gellert, Laki Buluwela, Alison Harrod, Allan Thornhill, et al. Discovery of naturally occurring esr1 mutations in breast cancer cell lines modelling endocrine resistance. *Nature communications*, **8**[1]:1865, 2017. 133

[208] Jeanette N McClintick and Howard J Edenberg. Effects of filtering by present call on analysis of microarray experiments. *BMC bioinformatics*, **7**[1]:49, 2006. 49

[209] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, **6**:26094, 2016. 34

[210] Smita Misra and Gautam Chaudhuri. Abstract 3159: Zar2 transcriptionally represses the atpase atp6v0a4 to negatively regulate invasiveness of breast cancer cells. *Cancer Research*, **74**:3159–3159, 10 2014. 167

[211] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997. 22

[212] Martin Fodslette Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, **6**[4]:525–533, 1993. 98, 99

[213] Clément Morgat, Gaétan MacGrogan, Véronique Brouste, Valérie Vélasco, Nicolas Sevenet, Hervé Bonnefoi, Philippe Fernandez, Marc Debled, and Elif Hindie. Expression of gastrin-releasing peptide receptor in breast cancer and its association with pathologic, biologic, and clinical parameters: a study of 1,432 primary tumors. *Journal of Nuclear Medicine*, **58**[9]:1401–1407, 2017. 164

[214] Clément Morgat, Romain Schollhammer, Gaétan Macgrogan, Nicole Barthe, Valérie Vélasco, Delphine Vimont, Anne-Laure Cazeau, Philippe Fernandez, and Elif Hindié. Comparison of the binding of the gastrin-releasing peptide receptor (grp-r) antagonist 68ga-rm2 and 18f-fdg in breast cancer samples. *PloS one*, **14**[1]:e0210905, 2019. 164

[215] Barry K Moser and Gary R Stevens. Homogeneity of variance in the two-sample means test. *The American Statistician*, **46**[1]:19–21, 1992. 72

[216] I Moy, V Todorović, AD Dubash, JS Coon, James B Parker, M Buranapramest, CC Huang, Hong Zhao, Kathleen Janee

Green, and Serdar E Bulun. Estrogen-dependent sushi domain containing 3 regulates cytoskeleton organization and migration in breast cancer cells. *Oncogene*, **34**[3]:323, 2015. 162

[217] Sayan Mukherjee. Classifying microarray data using support vector machines. In *A practical approach to microarray data analysis*, pages 166–185. Springer, 2003. 24

[218] Naoki Nanashima, Kayo Horie, Toshiyuki Yamada, Takeshi Shimizu, and Shigeki Tsuchida. Hair keratin krt81 is expressed in normal and breast cancer cells and contributes to their invasiveness. *Oncology reports*, **37**[5]:2964–2970, 2017. 154

[219] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, **490**[7418]:61, 2012. 51

[220] Wing WY Ng, Guangjun Zeng, Jiangjun Zhang, Daniel S Yeung, and Witold Pedrycz. Dual autoencoders features for imbalance classification problem. *Pattern Recognition*, **60**:875–889, 2016. 37

[221] Jie Niu, Xiao-Meng Li, Xiao Wang, Chao Liang, Yi-Dan Zhang, Hai-Ying Li, Fan-Ye Liu, Hua Sun, Song-Qiang Xie, and Dong Fang. Dkk1 inhibits breast cancer cell migration and invasion through suppression of $\beta$-catenin/mmp7 signaling pathway. *Cancer cell international*, **19**[1]:168, 2019. 154

[222] Joanna Obacz, Veronika Brychtova, Jan Podhorec, Pavel Fabian, Petr Dobes, Borivoj Vojtesek, and Roman Hrstka. anterior gradient protein 3 is associated with less aggressive tumors and better outcome of breast cancer patients. *OncoTargets and therapy*, **8**:1523, 2015. 134

[223] Il-Seok Oh, Jin-Seon Lee, and Byung-Ro Moon. Hybrid genetic algorithms for feature selection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, [11]:1424–1437, 2004. 33

[224] Lydia E Pace and Nancy L Keating. A systematic assessment of benefits and risks to guide breast cancer screening decisions. *Jama*, **311**[13]:1327–1335, 2014. 15

[225] Marco Padilla-Rodriguez, Sara S Parker, Deanna G Adams, Thomas Westerling, Julieann I Puleo, Adam W Watson, Samantha M Hill, Muhammad Noon, Raphael Gaudin, Jesse Aaron, et al. The actin cytoskeletal architecture of estrogen receptor positive breast cancer cells suppresses invasion. *Nature communications*, **9**[1]:1–16, 2018. 141

[226] Soonmyung Paik, Gong Tang, Steven Shak, Chungyeul Kim, Joffre Baker, Wanseop Kim, Maureen Cronin, Frederick L Baehner, Drew Watson, John Bryant, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol*, **24**[23]:3726–3734, 2006. 141

[227] CS Park, TK Kim, HG Kim, YJ Kim, MH Jeoung, WR Lee, NK Go, K Heo, and S Lee. Therapeutic targeting of tetraspanin8 in epithelial ovarian cancer invasion and metastasis. *Oncogene*, **35**[34]:4540–4548, 2016. 174

[228] Yudi Pawitan, Stefan Michiels, Serge Koscielny, Arief Gusnanto, and Alexander Ploner. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **21**[13]:3017–3024, 2005. 19

[229] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, **27**[8]:1226–1238, 2005. 31

[230] 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, et al. The somatic mutation profiles of 2,433 breast

cancers refine their genomic and transcriptomic landscapes. *Nature communications*, **7**:11479, 2016. 62

[231] Charles M Perou, Therese Sørlie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, et al. Molecular portraits of human breast tumours. *nature*, **406**[6797]:747, 2000. 16

[232] Amelia A Peters, Peter T Simpson, Johnathon J Bassett, Jane M Lee, Leonard Da Silva, Lynne E Reid, Sarah Song, Marie-Odile Parat, Sunil R Lakhani, Paraic A Kenny, et al. Calcium channel trpv6 as a potential therapeutic target in estrogen receptor–negative breast cancer. *Molecular cancer therapeutics*, **11**[10]:2158–2168, 2012. 157

[233] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 30–41. Springer, 2016. 34

[234] Zheng Ping, Yuchao Xia, Tiansheng Shen, Vishwas Parekh, Gene P Siegal, Isam-Eldin Eltoum, Jianbo He, Dongquan Chen, Minghua Deng, Ruibin Xi, et al. A microscopic landscape of the invasive breast cancer genome. *Scientific reports*, **6**:27545, 2016. 54

[235] Richard Possemato, Kevin M Marks, Yoav D Shaul, Michael E Pacold, Dohoon Kim, Kivanç Birsoy, Shalini Sethumadhavan, Hin-Koon Woo, Hyun G Jang, Abhishek K Jha, et al. Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. *Nature*, **476**[7360]:346, 2011. 148

[236] Ivan O. Potapenko, Torben Lders, Hege G. Russnes, slaug Helland, Therese Srlie, Vessela N. Kristensen, Silje Nord, OleC. Lingjrde, Anne-Lise Brresen-Dale, and Vilde D. Haakensen. Glycan-related gene expression signatures in breast cancer subtypes; relation to survival. *Molecular Oncology*, **9**[4]:861 – 876, 2015. 149

[237] Christopher Poultney, Sumit Chopra, Yann L Cun, et al. Efficient learning of sparse representations with an energy-based model. In *Advances in neural information processing systems*, pages 1137–1144, 2007. 36, 37

[238] Michael James David Powell. Restart procedures for the conjugate gradient method. *Mathematical programming*, **12**[1]:241–254, 1977. 97

[239] Laszlo Radvanyi, Devender Singh-Sandhu, Scott Gallichan, Corey Lovitt, Artur Pedyczak, Gustavo Mallo, Kurt Gish, Kevin Kwok, Wedad Hanna, Judith Zubovits, et al. The gene associated with trichorhinophalangeal syndrome in humans is overexpressed in breast cancer. *Proceedings of the National Academy of Sciences*, **102**[31]:11005–11010, 2005. 134

[240] James M Rae, Michael D Johnson, Joshua O Scheys, Kevin E Cordero, Jose M Larios, and Marc E Lippman. Greb1 is a critical regulator of hormone dependent breast cancer growth. *Breast cancer research and treatment*, **92**[2]:141–149, 2005. 164

[241] Michael L Raymer, William F. Punch, Erik D Goodman, Leslie A Kuhn, and Anil K Jain. Dimensionality reduction using genetic algorithms. *IEEE transactions on evolutionary computation*, **4**[2]:164–171, 2000. 33

[242] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**[3]:368–375, 2003. 19

[243] Tomás Reinert, Rodrigo Gonçalves, and José Bines. Implications of esr1 mutations in hormone receptor-positive breast cancer. *Current treatment options in oncology*, **19**[5]:24, 2018. 133

[244] Tomas Reinert, Everardo D Saad, Carlos H Barrios, and José Bines. Clinical implications of esr1 mutations in hormone receptor-positive advanced breast cancer. *Frontiers in oncology*, **7**:26, 2017. 133

[245] Cielito C Reyes-Gibby, Jian Wang, Sai-Ching J Yeung, Patrick Chaftari, K Yu Robert, Ehab Y Hanna, and Sanjay Shete. Genome-wide association study identifies genes associated with neuropathy in patients with head and neck cancer. *Scientific reports*, **8**[1]:8789, 2018. 141

[246] Markus Ringnér. What is principal component analysis? *Nature biotechnology*, **26**[3]:303, 2008. 21, 29

[247] Lee Roth, Swati Srivastava, Moshit Lindzen, Aldema Sas-Chen, Michal Sheffer, Mattia Lauriola, Yehoshua Enuka, Ashish Noronha, Maicol Mancini, Sara Lavi, et al. Silac identifies lad1 as a filamin-binding regulator of actin dynamics in response to egf and a marker of aggressive breast tumors. *Sci. Signal.*, **11**[515]:eaan0949, 2018. 168

[248] David E Rumelhart and James L McClelland. Parallel distributed processing: explorations in the microstructure of cognition. volume 1. foundations. 1986. 34

[249] Graeme D Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, **17**[4]:688–690, 2006. 70

[250] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, **23**[19]:2507–2517, 2007. 30, 31

[251] Steven L Salzberg. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, **16**[3]:235–240, 1994. 28

[252] Yasushi Sasaki, Ryota Koyama, Reo Maruyama, Takehiro Hirano, Miyuki Tamura, Jun Sugisaka, Hiromu Suzuki, Masashi Idogawa, Yasuhisa Shinomura, and Takashi Tokino. CLCA2, a target of the p53 family, negatively regulates cancer cell migration and invasion. *Cancer biology & therapy*, **13**[14]:1512–1521, dec 2012. 170

[253] Aarti Sathyanarayana, Shafiq Joty, Luis Fernandez-Luque, Ferda Ofli, Jaideep Srivastava, Ahmed Elmagarmid, Teresa Arora, and Shahrad Taheri. Correction of: sleep quality prediction from wearable data using deep learning. *JMIR mHealth and uHealth*, **4**[4], 2016. 35

[254] C. D. Savci-Heijink, H. Halfwerk, J. Koster, H. M. Horlings, and M. J. van de Vijver. A specific gene expression signature for visceral organ metastasis in breast cancer. *BMC Cancer*, **19**[1]:333, Apr 2019. 167

[255] Pieter Segaert, Marta B Lopes, Sandra Casimiro, Susana Vinga, and Peter J Rousseeuw. Robust identification of target genes and outliers in triple-negative breast cancer data. *Statistical methods in medical research*, page 0962280218794722, 2018. 146, 147, 149

[256] Erin K. Shanle, Zibo Zhao, John Hawse, Kari Wisinski, Sunduz Keles, Ming Yuan, and Wei Xu. Research Resource: Global Identification of Estrogen Receptor $\beta$ Target Genes in Triple Negative Breast Cancer Cells. *Molecular Endocrinology*, **27**[10]:1762–1775, oct 2013. 164

[257] Jonathan H Shepherd, Ivan P Uray, Abhijit Mazumdar, Anna Tsimelzon, Michelle Savage, Susan G Hilsenbeck, and Powel H Brown. The sox11 transcription factor is a critical regulator of basal-like breast cancer growth, invasion, and basal-like gene expression. *Oncotarget*, **7**[11]:13106, 2016. 150

[258] Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Shawn C Baker, Patrick J Collins, Francoise De Longueville, Ernest S Kawasaki, Kathleen Y Lee, et al. The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, **24**[9]:1151, 2006. 45

[259] Andrew K Shiau, Danielle Barstad, Paula M Loria, Lin Cheng, Peter J Kushner, David A Agard, and Geoffrey L Greene. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, **95**[7]:927–937, 1998. 16

[260] Hoo-Chang Shin, Matthew R Orton, David J Collins, Simon J Doran, and Martin O Leach. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. *IEEE transactions on pattern analysis and machine intelligence*, **35**[8]:1930–1943, 2013. 37

[261] Wojciech Siedlecki and Jack Sklansky. A note on genetic algorithms for large-scale feature selection. In *Handbook of Pattern Recognition and Computer Vision*, pages 88–107. World Scientific, 1993. 33, 68

[262] J Silhava and Pavel Smrz. Additional predictive value of microarray data compared to clinical variables. In *4th IAPR International Conference on Pattern Recognition in Bioinformatics*, 2009. 20

[263] R John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, **73**[3]:751–754, 1986. 19

[264] Marcel Smid, Yixin Wang, Jan GM Klijn, Anieta M Sieuwerts, Yi Zhang, David Atkins, John WM Martens, and John A Foekens. Genes associated with breast cancer metastatic to bone. *Journal of Clinical Oncology*, **24**[15]:2261–2267, 2006. 138

[265] He Song, Ming Dong, Jianping Zhou, Weiwei Sheng, Xin Li, and Wei Gao. Expression and prognostic significance of trpv6 in the development and progression of pancreatic cancer. *Oncology reports*, **39**[3]:1432–1440, 2018. 157

[266] Qinbao Song, Jingjie Ni, and Guangtao Wang. A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE transactions on knowledge and data engineering*, **25**[1]:1–14, 2013. 30

[267] Therese Sørlie, Charles M Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, **98**[19]:10869–10874, 2001. 16

[268] Therese Sørlie, Robert Tibshirani, Joel Parker, Trevor Hastie, James Stephen Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the national academy of sciences*, **100**[14]:8418–8423, 2003. 16

[269] John M Stewart. Trpv6 as a target for cancer therapy. *Journal of Cancer*, **11**[2]:374, 2020. 157

[270] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**[16]:9440–9445, 2003. 19

[271] Jie Sun, Xiaojuan Zhang, Yanchun Han, Juan Zhen, Yuan Meng, and Min Song. Overexpression of seven in absentia homolog 2 protein in human breast cancer tissues is associated with the promotion of tumor cell malignant behavior in in vitro. *Oncology reports*, **36**[3]:1301–1312, 2016. 136

[272] Yijun Sun, Sinisa Todorovic, and Steve Goodison. Local-learning-based feature selection for high-dimensional data analysis. *IEEE transactions on pattern analysis and machine intelligence*, **32**[9]:1610–1626, 2010. 30

[273] Yuliang Sun, Scooter Willis, Xiaoqian Lin, Justin Achua, Casey Williams, and Brian Leyland-Jones. Is fgd3 a potentially prognostic marker for breast cancer, 2017. 161

[274] Django Sussman, Leia M Smith, Martha E Anderson, Steve Duniho, Joshua H Hunter, Heather Kostner, Jamie B Miyamoto, Albina Nesterova, Lori Westendorf, Heather A Van Epps, et al. Sgn–liv1a: A novel antibody–drug conjugate targeting liv-1 for the treatment of metastatic breast cancer. *Molecular cancer therapeutics*, **13**[12]:2991–3000, 2014. 138

[275] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 35

[276] Anna Louise Swan, Ali Mobasheri, David Allaway, Susan Liddell, and Jaume Bacardit. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *Omics: a journal of integrative biology*, **17**[12]:595–610, 2013. 18

[277] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 35

[278] Motoki Takaku, Sara A Grimm, John D Roberts, Kaliopi Chrysovergis, Brian D Bennett, Page Myers, Lalith Perera, Charles J Tucker, Charles M Perou, and Paul A Wade. Gata3 zinc finger 2 mutations reprogram the breast cancer transcriptional network. *Nature communications*, **9**[1]:1–14, 2018. 144

[279] Motoki Takaku, Sara A Grimm, and Paul A Wade. Gata3 in breast cancer: tumor suppressor or oncogene? *Gene Expression The Journal of Liver Research*, **16**[4]:163–168, 2015. 144

[280] Satoshi Takakura, Takashi Kohno, Ryokuhei Manda, Aikou Okamoto, Tadao Tanaka, and Jun Yokota. Genetic alterations and expression of the protein phosphatase 1 genes in human cancers. *International journal of oncology*, **18**[4]:817–824, 2001. 158, 172

[281] Willem Talloen, Djork-Arné Clevert, Sepp Hochreiter, Dhammika Amaratunga, Luc Bijnens, Stefan Kass, and Hinrich WH Göhlmann. I/ni-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics*, **23**[21]:2897–2902, 2007. 49

[282] Feng Tan, Xuezheng Fu, Yanqing Zhang, and Anu G Bourgeois. A genetic algorithm-based method for feature subset selection. *Soft Computing*, **12**[2]:111–120, 2008. 33

[283] Jie Tan, John H Hammond, Deborah A Hogan, and Casey S Greene. Adage-based integration of publicly available pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. *MSystems*, **1**[1]:e00025–15, 2016. 38

[284] Jie Tan, Matthew Ung, Chao Cheng, and Casey S Greene. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. In *Pacific Symposium on Biocomputing Co-Chairs*, pages 132–143. World Scientific, 2014. 38

[285] Shuya Tang, Yilong Hao, Yao Yuan, Rui Liu, and Qianming Chen. Role of fibroblast growth factor receptor 4 in cancer. *Cancer Science*, **109**[10]:3024, 2018. 171

[286] Sandra Tavares, André Filipe Vieira, Anna Verena Taubenberger, Margarida Araújo, Nuno Pimpao Martins, Catarina Brás-Pereira, António Polónia, Maik Herbig, Clara Barreto, Oliver Otto, et al. Actin stress fiber organization promotes cell stiffening and proliferation of pre-invasive breast cancer cells. *Nature communications*, **8**[1]:1–18, 2017. 141

[287] Jeremy MG Taylor, Donna P Ankerst, and Rebecca R Andridge. Validation of biomarker-based risk prediction models. *Clinical Cancer Research*, **14**[19]:5977–5983, 2008. 11, 45

[288] Kathryn M Taylor, Helen E Morgan, Kathryn Smart, Normawati M Zahari, Sara Pumford, Ian O Ellis, John FR Robertson, and Robert I Nicholson. The emerging role of the liv-1 subfamily of zinc transporters in breast cancer. In *Molecular medicine*, **13**, pages 396–406. Springer, 2007. 137

[289] Dan Theodorescu and Harald Mischak. Mass spectrometry based proteomics in urine biomarker discovery. *World journal of urology*, **25**[5]:435–443, 2007. 18

[290] C Tomasetto, C Regnier, C Moog-Lutz, MG Mattei, MP Chenard, R Lidereau, P Basset, and MC Rio. Identification of four novel human genes amplified and overexpressed in breast carcinoma and localized to the q11-q21. 3 region of chromosome 17. *Genomics*, **28**[3]:367–376, 1995. 154

[291] Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh. Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm). *Journal of biomedical informatics*, **54**:96–105, 2015. 34

[292] David Tritchler, Elena Parkhomenko, and Joseph Beyene. Filtering genes for cluster and network analysis. *BMC bioinformatics*, **10**[1]:193, 2009. 49

[293] Luis Tume, Karen Paco, Roberto Ubidia-Incio, and Jeel Moya. Cd133 in breast cancer cells and in breast cancer stem cells as another target for immunotherapy. *Gaceta Mexicana de Oncología*, **15**[1]:22–30, 2016. 147

[294] Natsue Uehiro, Fumiaki Sato, Fengling Pu, Sunao Tanaka, Masahiro Kawashima, Kosuke Kawaguchi, Masahiro Sugimoto, Shigehira Saji, and Masakazu Toi. Circulating cell-free dna-based epigenetic assay can detect early breast cancer. *Breast Cancer Research*, **18**[1]:129, 2016. 149

[295] Anita Umesh, Jenny Park, James Shima, Joseph Delaney, Robert Wisotzkey, Erin Kelly, Elizabeth Beatrice Chiu, Jyoti Madhusoodanan, Mamatha Shekar, and Ilya Kupershmidt. Identification of agr3 as a potential biomarker though public genomic data analysis of triple-negative (tn) versus triple-positive (tp) breast cancer (bc)., 2012. 134

[296] Ewa Urbanczyk-Wochniak, Alexander Luedemann, Joachim Kopka, Joachim Selbig, Ute Roessner-Tunali, Lothar Willmitzer, and Alisdair R Fernie. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO reports*, **4**[10]:989–993, 2003. 18

[297] Fatemeh Vafaee, Connie Diakos, Michaela B Kirschner, Glen Reid, Michael Z Michael, Lisa G Horvath, Hamid Alinejad-Rokny, Zhangkai Jason Cheng, Zdenka Kuncic, and Stephen Clarke. A data-driven, knowledge-based approach to biomarker discovery: application to circulating microrna markers of colorectal cancer prognosis. *NPJ systems biology and applications*, **4**[1]:20, 2018. 92

[298] Kimberly D van der Willik, Mieke M Timmermans, Carolien HM van Deurzen, Maxime P Look, Esther A Reijm, Wendy JHP van Zundert, Renée Foekens, Anita MAC Trapman-Jansen, Michael A den Bakker, Pieter J Westenend, et al. Siah2 protein expression in breast cancer is inversely related with er status and outcome to tamoxifen therapy. *American journal of cancer research*, **6**[2]:270, 2016. 136

[299] Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin Van Der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, **415**[6871]:530, 2002. 21, 141

[300] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer Russo Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul D. Thomas, Jinghui Zhang, George L. Gabor Miklos, Catherine Nelson, Samuel Broder, Andrew G. Clark, Joe Nadeau, Victor A. McKusick, Norton Zinder, Arnold J. Levine, Richard J. Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin

Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E. Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J. Heiman, Maureen E. Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen A. Ketchum, Zhongwu Lai, Yiding Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V. Merkulov, Natalia Milshina, Helen M. Moore, Ashwinikumar K Naik, Vaibhav A. Narayan, Beena Neelam, Deborah Nusskern, Douglas B. Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Zhen Yuan Wang, Aihui Wang, Xin Wang, Jian Wang, Ming-Hui Wei, Ron Wides, Chunlin Xiao, and et al. Yan, Chunhua. The sequence of the human genome. *Science*, **291**[5507]:1304–1351, 2001. 2

[301] Jorge R Vergara and Pablo A Estévez. A review of feature selection methods based on mutual information. *Neural computing and applications*, **24**[1]:175–186, 2014. 30

[302] Jean-Louis Vincent, Jordi Rello, John Marshall, Eliezer Silva, Antonio Anzueto, Claude D Martin, Rui Moreno, Jeffrey Lipman, Charles Gomersall, Yasser Sakr, et al. International study of the prevalence and outcomes of infection in intensive care units. *Jama*, **302**[21]:2323–2329, 2009. 37

[303] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, **11**[Dec]:3371–3408, 2010. 35

[304] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of

gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, **101**[1]:5–22, 2017. 20

[305] Larissa Wakefield, James Robinson, Hilary Long, J Claire Ibbitt, Susanna Cooke, Helen C Hurst, and Edith Sim. Arylamine n-acetyltransferase 1 expression in breast cancer cell lines: A potential marker in estrogen receptor-positive tumors. *Genes, Chromosomes and Cancer*, **47**[2]:118–126, 2008. 138

[306] Randall Wald, Taghi M Khoshgoftaar, and Amri Napolitano. Stability of filter-and wrapper-based feature subset selection. In *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on*, pages 374–380. IEEE, 2013. 45

[307] Dujuan Wang, Guohong Liu, Balu Wu, Li Chen, Lihua Zeng, and Yunbao Pan. Clinical significance of elevated s100a8 expression in breast cancer patients. *Frontiers in oncology*, **8**:496, 2018. 156

[308] Fu-Wen Wang, Xiang Ao, and Shao-Mei Fu. Expression of sox11 and her2 and their association with recurrent breast cancer. *Translational Cancer Research*, **8**[1]:248–254, 2019. 150

[309] Shuang Wang, Dou Quan, Xuefeng Liang, Mengdan Ning, Yanhe Guo, and Licheng Jiao. A deep learning framework for remote sensing image registration. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018. 35

[310] Yu Wang, Igor V Tetko, Mark A Hall, Eibe Frank, Axel Facius, Klaus FX Mayer, and Hans W Mewes. Gene selection from microarray data for cancer classificationa machine learning approach. *Computational biology and chemistry*, **29**[1]:37–46, 2005. 20

[311] PH Watson, SK Chia, Charles C Wykoff, C Han, RD Leek, WS Sly, KC Gatter, P Ratcliffe, and AL Harris. Carbonic anhydrase xii is a marker of good prognosis in invasive breast carcinoma. *British journal of cancer*, **88**[7]:1065, 2003. 140

[312] Geoffrey I Webb. Discovering significant rules. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 434–443. ACM, 2006. 19

[313] Geoffrey I Webb. Discovering significant patterns. *Machine learning*, **68**[1]:1–33, 2007. 19

[314] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, **45**[10]:1113, 2013. 2

[315] Bernard L Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, **34**[1/2]:28–35, 1947. 30

[316] H Gilbert Welch, Philip C Prorok, A James OMalley, and Barnett S Kramer. Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness. *New England Journal of Medicine*, **375**[15]:1438–1447, 2016. 15

[317] John B Welsh, Lisa M Sapinoso, Suzanne G Kern, David A Brown, Tao Liu, Asne R Bauskin, Robyn L Ward, Nicholas J Hawkins, David I Quinn, Pamela J Russell, et al. Large-scale delineation of secreted protein biomarkers overexpressed in cancer tissue and serum. *Proceedings of the National Academy of Sciences*, **100**[6]:3410–3415, 2003. 134

[318] Scooter Willis, Yuliang Sun, Mark Abramovitz, Teng Fei, Brandon Young, Xiaoqian Lin, Min Ni, Justin Achua, Meredith M Regan, Kathryn P Gray, et al. High expression of fgd3, a putative regulator of cell morphology and motility, is prognostic of favorable outcome in multiple cancers. *JCO Precision Oncology*, **1**:1–13, 2017. 161

[319] Stephan M Winkler, Michael Affenzeller, Witold Jacak, and Herbert Stekel. Identification of cancer diagnosis estimation models

using evolutionary algorithms: a case study for breast cancer, melanoma, and cancer in the respiratory system. In *Proceedings of the 13th annual conference companion on Genetic and evolutionary computation*, pages 503–510. ACM, 2011. 33

[320] Ian H Witten and Eibe Frank. Data mining practical learning tools and techniques with java implementations, 2000. 41

[321] Pu Xia. Cd133 mrna may be a suitable prognostic marker for human breast cancer. *Stem cell investigation*, **4**, 2017. 147

[322] Bin Xiao, Jianfeng Hang, Ting Lei, Yongyin He, Zhenzhan Kuang, Li Wang, Lidan Chen, Jia He, Weiyun Zhang, Yang Liao, et al. Identification of key genes relevant to the prognosis of er-positive and er-negative breast cancer based on a prognostic prediction system. *Molecular biology reports*, pages 1–9, 2019. 138

[323] M Xu, S Chen, W Yang, X Cheng, Y Ye, J Mao, X Wu, L Huang, and J Ji. FGFR4 Links Glucose Metabolism and Chemotherapy Resistance in Breast Cancer. *Cellular Physiology and Biochemistry*, **47**[1]:151–160, 2018. 172

[324] Bing Xue, Mengjie Zhang, Will N Browne, and Xin Yao. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*, **20**[4]:606–626, 2016. 33

[325] Mutsuko Yamamoto-Ibusuki, Yutaka Yamamoto, Saori Fujiwara, Aiko Sueta, Satoko Yamamoto, Mitsuhiro Hayashi, Mai Tomiguchi, Takashi Takeshita, and Hirotaka Iwase. C6orf97-esr1 breast cancer susceptibility locus: influence on progression and survival in breast cancer patients. *European Journal of Human Genetics*, **23**[7]:949, 2015. 136

[326] Meisi Yan, Jinsong Wang, Yanlv Ren, Lin Li, Weidan He, Ying Zhang, Tong Liu, and Zhigao Li. Over-expression of fsip1 promotes breast cancer progression and confers resistance to docetaxel via mrp1 stabilization. *Cell death & disease*, **10**[3]:1–13, 2019. 143

[327] Jihoon Yang and Vasant Honavar. Feature subset selection using a genetic algorithm. In *Feature extraction, construction and selection*, pages 117–136. Springer, 1998. 33

[328] Kaidi Yang, Jian Gao, and Mao Luo. Identification of key pathways and hub genes in basal-like breast cancer using bioinformatics analysis. *OncoTargets and therapy*, **12**:1319, 2019. 153

[329] Pengyi Yang, Yee Hwa Yang, Bing B Zhou, and Albert Y Zomaya. A review of ensemble methods in bioinformatics. *Current Bioinformatics*, **5**[4]:296–308, 2010. 32

[330] Fang Yao, Chi Zhang, Wei Du, Chao Liu, and Ying Xu. Identification of gene-expression signatures and protein markers for breast cancer grading and staging. *PloS one*, **10**[9]:e0138213, 2015. 161

[331] Youngjin Yoo, Tom Brosch, Anthony Traboulsee, David KB Li, and Roger Tam. Deep learning of image features from unlabeled data for multiple sclerosis lesion segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 117–124. Springer, 2014. 34

[332] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 856–863, 2003. 31

[333] Qingyan Zhang, Danhui Huang, Zhenfei Zhang, Yuzhen Feng, Meiting Fu, Min Wei, Zhou Jueyu, Yuanjin Huang, Shuguang Liu, and Rong Shi. High expression of tmem40 contributes to progressive features of tongue squamous cell carcinoma. *Oncology Reports*, **41**, 10 2018. 151

[334] Sen Zhang and Yu Zhang. Seeking for correlative genes and signaling pathways with bone metastasis from breast cancer by integrated analysis. *Frontiers in Oncology*, **9**:138, 2019. 147

[335] Zhen-Fei Zhang, Han-Rong Zhang, Qing-Yan Zhang, Shu-Yu Lai, Yu-Zhen Feng, Yi Zhou, Si-Rong Zheng, Rong Shi, and

Zhou Jueyu. High expression of tmem40 is associated with the malignant behavior and tumorigenesis in bladder cancer. *Journal of Translational Medicine*, **16**:9, 12 2018. 151

[336] Shuang Zhao, Shuang-Shuang Chen, Yuan Gu, En-Ze Jiang, and Zheng-Hong Yu. Expression and clinical significance of sushi domain-containing protein 3 (susd3) and insulin-like growth factor-i receptor (igf-ir) in breast cancer. *Asian Pac J Cancer Prev*, **16**[18]:8633–8636, 2015. 162

[337] Xiangdong Zhao, Faliang Xu, Nestor P Dominguez, Yuanping Xiong, Zhongxun Xiong, Hong Peng, Chloe Shay, and Yong Teng. Fgfr4 provides the conduit to facilitate fgf19 signaling in breast cancer progression. *Molecular carcinogenesis*, **57**[11]:1616–1625, 2018. 172

[338] Wei Zheng, Jirong Long, Yu-Tang Gao, Chun Li, Ying Zheng, Yong-Bin Xiang, Wanqing Wen, Shawn Levy, Sandra L Deming, Jonathan L Haines, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25. 1. *Nature genetics*, **41**[3]:324, 2009. 136

[339] Jing-Min Zhong, Jing Li, An-Ding Kang, San-Qian Huang, Wen-Bin Liu, Yun Zhang, Zhi-Hong Liu, and Liang Zeng. Protein s100-a8: A potential metastasis-associated protein for breast cancer determined via itraq quantitative proteomic and clinicopathological analysis. *Oncology letters*, **15**[4]:5285–5293, 2018. 155

[340] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, **12**[10]:931, 2015. 34

[341] Jindan Zhu, Amit Pande, Prasant Mohapatra, and Jay J Han. Using deep learning for energy expenditure estimation with wearable sensors. In *E-health Networking, Application & Services (HealthCom), 2015 17th International Conference on*, pages 501–506. IEEE, 2015. 35

[342] Rongxuan Zhu, Olivier Gires, Liqun Zhu, Jun Liu, Junjian Li, Hao Yang, Gaoda Ju, Jing Huang, Weiyu Ge, Yi Chen, et al.

Tspan8 promotes cancer cell stemness via activation of sonic hedgehog signaling. *Nature communications*, **10**[1]:1–14, 2019. 174

[343] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang. Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sensing Lett.*, **12**[11]:2321–2325, 2015. 35

[344] Marketa Zvelebil, Erik Oliemuller, Qiong Gao, Olivia Wansbury, Alan Mackay, Howard Kendrick, Matthew J Smalley, Jorge S Reis-Filho, and Beatrice A Howard. Embryonic mammary signature subsets are activated in brca1-/-and basal-like breast cancers. *Breast Cancer Research*, **15**[2]:R25, 2013. 149