Classification of Epithelial Ovarian Cancer Subtypes.

Hannah Simpson-Clancy

This thesis is submitted to the School of Computing and Engineering, University of Huddersfield, in partial fulfilment of the requirements for the degree of MSc by Research.
January 2024.

Abstract

Superior prognosis is associated with earlier staged diagnosis of ovarian cancer. Patient prognosis is further affected by histology of epithelial ovarian cancer. Accurate methods to distinguish between these histological subtypes is crucial for faster diagnosis, effective treatment choices and both improved prognosis and treatment response in patients. Subtypes include clear cell, endometrioid, mucinous, high grade and low grade serous. This study trained and validated linear, radial basis function and polynomial support vector machines, K-nearest neighbours and random forest models to classify patient samples in four separate cases: (1) clear cell vs endometrioid, (2) clear cell vs non-clear cell, (3) LGSOC vs non-LGSOC, (4) mucinous vs non-mucinous. Models for each case were trained using gene signatures of lengths 11, 15, 18 and 25, respectively. Gene signatures were formed by a selection of differentially expressed probes representing genes unique to a case's specific histology identified by non-parametric hypothesis tests (BH adjusted, $p<0.05$) and fold change analysis ($FC<0.5$ or $FC>2$).

The polynomial support vector machine model for classifying case (1) performed best achieving a test F_1 score of 100%. The best performing model for classifying case (2) was a random forest model achieving testing F_1 score of 90.9%. The best performing model for classifying case (3) was the radial basis function support vector machine model exhibiting a train and test F_1 score of 100%. A linear support vector machine model performed best for classifying case (4), displaying a train and test F_1 score of 100%. Based on the proposed gene signatures in this study, it is indicated that mucinous and non-mucinous samples were completely linearly separable. Moreover, gene signatures suggested for LGSOC and mucinous classifications present possible diagnostic capabilities for their corresponding EOC subtype. Clear cell and endometrioid EOC can also be differentiated between by using the suggested gene signature within this study.

Further study on the selected genes in each gene signature is suggested to determine their role in ovarian cancer and specific histological subtypes of epithelial ovarian cancer.

Keywords: epithelial ovarian cancer; classification; statistics.

Dedications

First and foremost, a thank you to my supervisors Dr Ann Smith and Professor William Lee for their encouragement, support and time throughout my work and submission. Secondly, thank you to my family for their ongoing support. A very special thank you to my daughter, my greatest source of inspiration and determination.

Contents

1 Introduction **6**

1.1 Epithelial Ovarian Cancer (EOC) 6
1.1.1 Screening 7
1.1.2 Symptoms 7
1.1.3 Current Treatments 8
1.1.4 PARP Inhibitor Treatment Trials 11
1.2 EOC Histological Subtypes 13
1.2.1 Clinical Features 13
1.2.2 Genetic Features (Known Mutations) 16
1.3 Classification 21
1.3.1 Overview of classification techniques 21
1.4 Classification in Disease 22
1.4.1 Cervical Cancer 22
1.4.2 Diabetes 23
1.4.3 Heart Disease 23
1.4.4 Covid-19 23
1.4.5 Alzheimer's Disease 24
1.4.6 Breast cancer 24
1.5 Classification in Ovarian Cancer 25
1.5.1 Ovarian Cancer Presence. 25
1.5.2 Stages 26
1.5.3 Subtypes 27
1.5.4 Treatment Response 29
1.5.5 Survival 31
1.5.6 Recurrence 32
1.6 Current Study 32
1.6.1 Aims 32
1.6.2 Thesis Overview 33

2 Methodology **35**
2.1 Training Data . . . 35
2.2 Testing Data . . . 37
2.3 Data Preparation . . . 39
2.4 Hypothesis testing . . . 39
2.4.1 Fold Change Analysis . . . 39
2.4.2 Wilcoxon Rank-Sum Test . . . 40
2.4.3 Kruskal Wallis Test . . . 41
2.4.4 Dunn Test . . . 42
2.4.5 Benjamini-Hochberg P-value Correction . . . 43
2.4.6 Selection of gene signatures . . . 44
2.5 Principal Component Analysis . . . 45
2.6 Classifiers . . . 46
2.6.1 Choice of Classifier . . . 46
2.6.2 K-Nearest Neighbours (KNN) . . . 47
2.6.3 Linear Support Vector Machines (Linear SVM) . . . 48
2.6.4 Non-Linear Support Vector Machines (Non-linear SVM) 50
2.6.5 Random forest . . . 51
2.6.6 Stratified K-fold cross validation . . . 52
2.7 Evaluation Metrics . . . 53
2.7.1 Confusion Matrix . . . 53
2.7.2 Unbalanced Evaluation Metrics . . . 54
2.7.3 Precision-Recall Curve . . . 57
2.7.4 Balanced Evaluation Metrics . . . 58
2.7.5 Receiver Operating Curves . . . 58
2.8 R statistical software . . . 60

3 Results **61**
3.1 Determining Statistically Significant Genes. . . . 61
3.2 Classifying Patients with Clear Cell EOC . . . 64
3.2.1 Gene Signature . . . 64
3.2.2 Principal Component Analysis . . . 64
3.2.3 Classification Model Training and Testing . . . 66
3.3 Classifying Patients with Mucinous EOC . . . 71
3.3.1 Gene Signature . . . 71
3.3.2 Principal Component Analysis . . . 72
3.3.3 Classification Model Training and Testing . . . 74
3.4 Classifying Patients with LGSOC . . . 78
3.4.1 Gene Signature . . . 78
3.4.2 Principal Component Analysis . . . 79
3.4.3 Classification Model Training and Testing . . . 81

3.5 Classifying Endometrioid vs Clear Cell Patients 85
3.5.1 Gene Signature 85
3.5.2 Principal Component Analysis 86
3.5.3 Classification Model Training and Testing 87

4 Discussion **93**
4.1 Overview of Ovarian Cancer 93
4.2 Prior Classification Studies of EOC Subtypes 94
4.3 Current Study Purpose 94
4.4 Case 1: Clear Cell versus Non-Clear Cell EOC 95
4.4.1 Classification Results Overview 95
4.4.2 Clear Cell Gene Signature 96
4.5 Case 2: Mucinous versus Non-mucinous EOC 99
4.5.1 Classification Results Overview 99
4.5.2 Mucinous Gene Signature 101
4.6 Case 3: LGSOC versus non-LGSOC 105
4.6.1 Classification Results Overview 105
4.6.2 LGSOC Gene Signature 106
4.7 Case 4: Endometrioid versus Clear Cell EOC 109
4.7.1 Classification Results Overview 109
4.7.2 Clear Cell versus Endometrioid Gene Signature 110
4.8 Limitations of Study 112
4.9 Future Research 113

5 Conclusion **115**

6 References **117**

A Tables **187**

B Copyright Statement **189**

List of Tables

1.1 Type I vs II OC. 7

2.1 EOC Histological Subtypes of Training Data 36
2.2 EOC Histological Subtypes of Test Data 37
2.3 Example of Dunn test criteria for gene signature selection. . . 45
2.4 Confusion Matrix for Binary Classification. 53

3.1 Differentially Expressed and Dys-regulated Probes. 61
3.2 Retained Differentially Expressed Probes for Gene Signatures. 63
3.3 Gene Signature: Clear Cell versus Non-Clear Cell EOC. 64
3.4 Clear Cell versus Non-Clear Cell Training Data Metrics. . . . 68
3.5 Clear Cell versus Non-Clear Cell Test Data Metrics. 68
3.6 Gene Signature: Mucinous versus Non-Mucinous EOC. 71
3.7 Mucinous versus Non-Mucinous Training Data Metrics. 75
3.8 Mucinous versus Non-Mucinous Test Data Metrics. 75
3.9 Gene Signature: LGSOC versus Non-LGSOC. 78
3.10 LGSOC versus Non-LGSOC Training Data Metrics. 82
3.11 LGSOC versus Non-LGSOC Test Data Metrics. 82
3.12 Gene Signature: Clear Cell versus Endometrioid EOC. 85
3.13 Clear Cell versus Endometrioid Training Data Metrics. 89
3.14 Clear Cell versus Endometrioid Test Data Metrics. 89

A.1 Grading System for Ovarian Cancer 187
A.2 FIGO staging of Ovarian Cancer. 188

Chapter 1

Introduction

An estimated 7,500 newly diagnosed ovarian cancer cases arise annually in the UK; projections indicate an increase to 9,400 by 2038-2040 (Cancer Research UK, n.d.b). Approximately 94.5% of stage I ovarian cancer patients will survive for five years post diagnosis. When comparing this to the 16% five-year survival probability for patients diagnosed with stage IV ovarian cancer, it is clear that an earlier diagnosis of ovarian cancer is crucial to improve patient survival and prognosis (Cancer Research UK, 2022). Of the ovarian cancer cases diagnosed in England in 2020, 1,858 were early stage (I-II) and 3,015 were late stage (III-IV) (Cancer Research UK, 2022).

1.1 Epithelial Ovarian Cancer (EOC)

The literature describes ovarian cancer (OC) manifestation in three forms: epithelial, germ-cell and sex cord stromal (Matulonis et al., 2016). This study will focus on epithelial ovarian cancer (EOC) with no further discussion of germ-cell or sex cord stromal OC.

EOC tumours can be described as benign, borderline or malignant (Chen et al., 2003). Malignant tumours are categorised by five histologies: clear cell, endometrioid, mucinous, serous and Brenner (Chen et al., 2003). Serous EOC has been divided into two categories based on grade at diagnosis; grade I is defined as low-grade serous OC (LGSOC) and grade II-III defined as high-grade serous OC (HGSOC) (Bodurka et al., 2012). See appendix A:table A.1 for details on OC grading systems. Studies report further division of EOC histotypes, separating EOC histological subtypes into type I or type II groups allowing the incorporation of derived location (Shih & Kurman, 2004; Kurman & Shih, 2016).

Table 1.1: Type I vs II OC. Adapted from Kurman, R.J., & Shih, I.-M. (2016). The dualistic model of ovarian carcinogenesis: Revisited, revised and expanded. *The American Journal of Pathology, 186*, 733-747.

Type I				Type II
Endometriosis	Fallopian Tube	Germ Cell	Transitional Cell	Fallopian Tube
Endometrioid Carcinoma	LGSOC	Mucinous Carcinoma	Mucinous Carcinoma	HGSOC
Clear Cell Carcinoma			Brenner Tumours	Carcinosarcoma
Seromucinous Carcinoma				Undifferentiated Carcinoma

The histological subtypes of EOC focused on in this study are clear cell, endometrioid, mucinous, HGSOC and LGSOC.

1.1.1 Screening

Studies have provided evidence that screening for disease has the ability to reduce mortality rates: colorectal, cervical and breast cancer are three cancers in which screening has been deemed as successful (Hewitson et al., 2008; Peto et al., 2004; Berry et al., 2005). However, screening cannot be recommended for the public in the case of OC. The use of CA-125 blood testing and trans-vaginal ultrasounds for patients deemed high or non-high risk has indicated improvements in early detection of OC but no significant improvement in mortality rates in OC patients (Menon et al., 2021; Rosenthal et al., 2013, 2017). This lack of screening ability reduces the chances of accurate and efficient early OC diagnosis.

1.1.2 Symptoms

Ovarian cancer has been described as asymptomatic in its earlier stages hindering early detection of the disease. However, studies have taken place to determine the symptoms that may suggest the presence of OC. Goff et al. (2004) observed that malignant OC patients experienced a significantly greater number of symptoms than non-malignant and healthy patients. The symptoms Goff et al. (2004) proposed to be associated with malignant OC were increased abdominal size, bloating, urinary frequency and pelvic pain. A symptoms index was later developed for OC with pelvic/abdominal pain, urinary urgency/frequency, increased abdominal size/bloating and difficulty eating/feeling full being significantly associated with OC when the symptoms occurred for more than 12 days per month within a 12 month period (Goff et al., 2007). Similarly, Lurie et al. (2009) suggested abdominal pain to be predictive of OC but for early stage tumours. Further symptoms this study

found to be predictive of early stage OC were distended and hard abdomen, vaginal bleeding and palpable abdominal mass (Lurie et al., 2009).

A case-control study conducted by Kim et al. (2009) employed a symptoms index identical to that previously proposed by (Goff et al., 2007). Kim et al. (2009) found a greater proportion of OC patients had a positive symptoms index compared with benign and control patients. In addition, the index was positive for a greater percentage of late stage patients when compared with early stage (72.9% versus 44.8%). Kim et al. (2009) concluded that the symptom index could independently predict OC. Another study utilised the symptom index introduced by Goff et al. (2007) with additional symptoms such as back pain, indigestion, nausea, vomiting, weight loss, constipation, diarrhoea, menstrual irregularities, bleeding after menopause, pain during intercourse, fatigue, leg swelling and difficulty breathing (Ore et al., 2017). From this study, Ore et al. (2017) found that 88.8% of participants reported at least one symptom during the study; only 3.9% of these patients met the frequency and duration stated by Goff et al. (2007). Participants' meeting frequency and duration criteria experienced significantly greater severity of symptoms in comparison to participants not meeting the criteria (Ore et al., 2017).

Currently, National Institute for Health and Care Excellence (NICE) guidelines recommend that testing for OC should be considered if symptoms including bloating, appetite loss, abdomen or pelvic pain and increased urinary frequency are reported (National Institute for Health and Care Excellence., 2011). Although these symptoms instigate investigation, they cannot be relied upon to indicate OC, particularly in its earliest stages.

1.1.3 Current Treatments

Current treatments for OC are not specific to EOC histological subtypes.

Primary Surgery

Surgery in OC is utilised to make a diagnosis and remove tumour tissue to improve the susceptibility of remaining tissue to chemotherapy (Coleman et al., 2013).

The outcome of primary surgery for OC patients is described as (1) complete resection (no residual disease), (2) optimal debulking (residual disease with diameter $<$ 1cm), (3) suboptimal debulking (residual disease with diameter $>$ 1cm) (du Bois et al., 2009).

Complete resection in stages III-IV is associated with improved prognosis, longer overall and progression-free survival and lower recurrence risk

when compared to optimal and sub-optimal debulking (Ataseven et al., 2016; Polterauer et al., 2012; Luyckx et al., 2012; Chang et al., 2012; Winter et al., 2007; Wimberger et al., 2010; du Bois et al., 2009). However, advanced stage OC has been associated with lower complete resection rates (Polterauer et al., 2012).

Adjuvant chemotherapy

Adjuvant chemotherapy is performed after primary surgery; its aim is to reduce the recurrence risks in patients by eliminating any residual tumour (Lawrie et al., 2015).

Previous observations indicate that both recurrence-free and overall survival of patients is greater in those who receive immediate adjuvant chemotherapy in comparison to those who do not; particularly those with high risk, stage I OC (Collinson et al., 2014). Association of longer intervals between primary debulking surgery and administration of adjuvant chemotherapy with poor overall survival in completely resected patients has been reported (Timmermans et al., 2018).

Neoadjuvant Chemotherapy

Neoadjuvant chemotherapy is performed in advanced OC cases as primary treatment in patients who cannot have primary debulking surgery or whose tumours cannot be debulked (Kessous et al., 2017).

Reports have determined that although neoadjuvant chemotherapy succeeded by interval debulking surgery does not significantly improve overall survival of advanced stage OC patients when compared with primary debulking surgery, it may increase the chance of complete resection in debulking surgery (Rauh-Hain et al., 2012; Bian et al., 2016; Vergote et al., 2010; Kehoe et al., 2015; Zeng et al., 2016).

Standard Chemotherapy Treatments

Standard chemotherapy for OC is platinum-based carboplatin and paclitaxel in three weekly doses for six cycles (Karam et al., 2017).

Studies have demonstrated significant improvement of progression-free and overall survival in patients who received dose-dense weekly treatments of paclitaxel and three weekly doses of carboplatin (Katsumata et al., 2009, 2013). In contrast, a clinical trial ICON8 trialled EOC patients with three varying treatments

1. three weekly treatments of carboplatin and paclitaxel for six cycles

2. weekly treatments of paclitaxel and three weekly treatments of carboplatin

3. weekly treatments of paclitaxel and carboplatin

and reported no significant improvement in patient progression-free survival when receiving dose-dense treatments compared with standard treatments (Clamp et al., 2019).

Bevacizumab

The introduction of bevacizumab as cancer treatment is based on prevention of angiogenesis by inhibition of the VEGF protein (Hicklin & Ellis, 2005). VEGF promotes angiogenesis - the formation of blood vessels that aid development and survival of tumours (Hanahan & Weinberg, 2011).

Bevacizumab has been introduced in various trials to work alongside chemotherapy. Varying combinations of treatments have been trialled including

1. Standard chemotherapy versus standard chemotherapy with concurrent bevacizumab (Pujade-Lauraine et al., 2014; Perren et al., 2011).

2. Gemcitabine and carboplatin chemotherapy with placebo versus gemcitabine and carboplatin with bevacizumab (Aghajanian et al., 2012).

3. Standard chemotherapy with concurrent bevacizumab versus standard chemotherapy with concurrent and maintenance bevacizumab (Tewari et al., 2019).

4. Standard chemotherapy with placebo versus standard chemotherapy with concurrent bevacizumab and placebo versus standard chemotherapy with maintenance bevacizumab (Burger et al., 2011).

5. Standard chemotherapy with and without bevacizumab versus dense-dose chemotherapy with and without bevacizumab (Chan et al., 2016).

It is reported that bevacizumab treated groups have significantly improved progression-free survival when compared with control groups (Pujade-Lauraine et al., 2014; Aghajanian et al., 2012; Perren et al., 2011; Burger et al., 2011). Additionally, significant improvement in patient objective response rate (Pujade-Lauraine et al., 2014; Aghajanian et al., 2012) and duration of response (Aghajanian et al., 2012) was associated with bevacizumab treatment. The objective response rate is defined as the proportion of people with partial or complete response to treatment in a specified time period

(National Cancer Institute, n.d.). The duration of response is defined as the time to disease progression or death in patients with a partial or complete response to treatment (Delgado & Guddati, 2021). However, improved overall survival was not found in bevacizumab treatment compared with the control groups (Pujade-Lauraine et al., 2014; Burger et al., 2011; Tewari et al., 2019), with the exception of patients with high disease progression risk (Perren et al., 2011). Chan et al. (2016) reported from results of a clinical trial comparing weekly and dose-dense paclitaxel chemotherapy treatment with and without bevacizumab that progression-free survival did not significantly differ between groups.

Bevacizumab treatment was associated with increased hypertension frequency (Aghajanian et al., 2012; Burger et al., 2011), proteinuria (Aghajanian et al., 2012) and other toxic effects (Perren et al., 2011).

1.1.4 PARP Inhibitor Treatment Trials

PARP protein inhibition destroys tumour cells by preventing tumour cells repairing; trials have considered a variety of PARP inhibitors as OC treatment (Konecny & Kristeleit, 2016), a few of which will be discussed below. These trials have considered specific histological subtypes of EOC. Particularly the most prevalent subtype: serous.

Olaparib

Primary and recurrent high-grade serous and endometrioid OC (HGSOC and HGEOC, respectively), with BRCA mutations and platinum-sensitivity were common tumour characteristics of patients admitted to Olaparib maintenance trials (Moore et al., 2018; Ray-Coquard et al., 2019; Pujade-Lauraine et al., 2017a,b; Ledermann et al., 2014, 2016; Liu et al., 2014b). Platinum-sensitivity is defined as no disease progression at least 6 months after the last dose of platinum-based chemotherapy (Pujade-Lauraine et al., 2017a,b; Ledermann et al., 2014, 2016; Liu et al., 2014b).

When trialling Olaparib doses of twice daily 300mg against a placebo drug, improvements in progression-free survival were noted in newly diagnosed, advanced stage HGSOC and HGEOC with BRCA mutations and platinum-sensitivity regardless of the presence of bevacizumab treatment (Moore et al., 2018; Ray-Coquard et al., 2019). However, Olaparib maintenance treatment when compared with placebo treatments did not display significant differences in overall survival for recurrent, BRCA-mutated, platinum-sensitive, serous EOC patients (Ledermann et al., 2014, 2016). Whereas, for BRCA-mutated, platinum-sensitive, relapsed, HGSOC or HGEOC

patients exhibited improved progression-free survival when treated with Olaparib maintenance treatment compared with placebo treatment (Pujade-Lauraine et al., 2017a,b).

Clinical trials also considered a dose of 400mg Olaparib once or twice daily. In advanced OC patients the objective response rate was greater in BRCA-mutated, platinum-sensitive than BRCA-mutated, platinum-resistant patients (Domchek et al., 2016). Platinum-resistance is defined as disease progression occurrence within 6 months of the last platinum-based chemotherapy (Domchek et al., 2016). However, the duration of response was similar for both patient groups (Domchek et al., 2016). Gelmon et al. (2011) compared objective response rates between BRCA positive and negative patients; this rate was greater in positive patients who were not HGSOC histology.

Audeh et al. (2010) compared twice daily 400mg Olaparib doses with twice daily 100mg doses focusing on advanced, BRCA-mutated OC patients. Results indicate the objective response rate and progression-free survival was greater in 400mg Olaparib doses regardless of BRCA status (Audeh et al., 2010).

Trials to compare combination treatments including Olaparib as a maintenance have also been performed. Oza et al. (2015) compared standard platinum chemotherapy (carboplatin and paclitaxel) with standard chemotherapy plus Olaparib. Whereas, Liu et al. (2014b) compared combined cediranib and Olaparib with Olaparib. Significant improvement of progression-free survival in BRCA-mutated, platinum-sensitive, serous OC patients receiving standard chemotherapy combined with Olaparib was observed (Oza et al., 2015). The combination of Olaparib with another chemotherapy drug, cediranib, was found to improve progression-free survival in patients with recurrent, platinum-sensitive, HGSOC or HGEOC when comparing this combined treatment with Olaparib as a single drug treatment (Liu et al., 2014b).

In all studies discussed here, the adverse events with most frequent occurrence were nausea and fatigue (Audeh et al., 2010; Gelmon et al., 2011; Domchek et al., 2016; Ledermann et al., 2016; Pujade-Lauraine et al., 2017a; Ledermann et al., 2014; Oza et al., 2015; Liu et al., 2014b; Moore et al., 2018; Ray-Coquard et al., 2019).

Rucaparib

Trials of rucaparib as oral and IV treatments for germline BRCA-mutated OC maintenance concluded that oral rucaparib treatment had a greater objective response rate than IV treatments with an optimal dose of 600mg twice daily (Drew et al., 2016; Kristeleit et al., 2017).

This recommended dose was trialled as maintenance treatment for high-

grade, platinum-sensitive, primary and recurrent patients with BRCA1/2 mutations (Swisher et al., 2017; Coleman et al., 2017; Kristeleit et al., 2017; Oza et al., 2017). An objective response rate of greater than 50% was identified in high-grade OC patients (Oza et al., 2017; Kristeleit et al., 2017) and significantly improved progression-free survival was observed in high-grade primary and recurrent, platinum-sensitive patients (Coleman et al., 2017; Swisher et al., 2017).

Adverse events with the most frequent occurrence with grade 3 or greater severity were anaemia, fatigue and increased aspartate transaminase and alanine transaminase levels (Coleman et al., 2017; Swisher et al., 2017; Kristeleit et al., 2017). An adverse event with grade 3 severity is one that is severe and could lead to hospitalisation but is not life-threatening (National Cancer Institute, 2017).

Niraparib

Clinical trials consider the use of niraparib as treatment for a number of OC types. These include platinum-sensitive, newly diagnosed, advanced HGSOC or HGEOC (González-Martín et al., 2019) and recurrent OC patients (Oza et al., 2018; del Campo et al., 2019; Mirza et al., 2016).

Patients diagnosed with recurrent, platinum-sensitive OC had significantly improved progression-free survival when receiving niraparib treatment versus placebo treatment (González-Martín et al., 2019; Mirza et al., 2016; del Campo et al., 2019). A patient's BRCA mutation status was reported as having no effect on the significance of this result (Mirza et al., 2016; del Campo et al., 2019).

Grade 3 or greater toxicities most frequently observed in the described trials were thrombocytopenia, anaemia, neutropenia, hypertension and fatigue (González-Martín et al., 2019; del Campo et al., 2019; Oza et al., 2018); it was concluded that quality of life could be maintained during a niraparib treatment (Oza et al., 2018).

1.2 EOC Histological Subtypes

1.2.1 Clinical Features

Here, a comparison of clinical features presented by the EOC subtypes will be discussed.

Stage at diagnosis

Greater than 75% of type II tumours are advanced stage (IV) at diagnosis with serous being diagnosed more frequently in later development stages (III/IV) (Kurman & Shih, 2016; Torre et al., 2018). In comparison, diagnosis in earlier development stages (I-II) is more frequent for type I tumours (Torre et al., 2018; Alcázar et al., 2013). For example, when compared with serous OC, clear cell EOC was observed to have a significantly greater chance of having a stage I diagnosis (Sugiyama et al., 2000). When considering histological subtypes of EOC independently, HGSOC was more frequently diagnosed in later stages of tumour development (Köbel et al., 2010). More frequent early stage diagnosis is found in clear cell (Köbel et al., 2010; Shu et al., 2015; Zhu et al., 2021; Winterhoff et al., 2016), endometrioid (Winterhoff et al., 2016) and mucinous (Brown & Frumovitz, 2014).

Age at diagnosis

Studies have previously suggested that age at diagnosis is associated with histology of EOC. It has been argued that HGSOC is diagnosed at a significantly greater age than clear cell, endometrioid, mucinous and LGSOC; significant differences discovered between type I and type II EOC patients provide agreements with this statement (Mackay et al., 2010; Torre et al., 2018; Alcázar et al., 2013; Peres et al., 2019).

Survival Rates and Prognosis

Another topic of research for EOC histotypes is survival rates and prognosis. Prior to the introduction of LGSOC and HGSOC subtypes, studies found that for advanced stage (III-IV) tumours, mucinous (Mackay et al., 2010; Zaino et al., 2011; Bamias et al., 2010; Alexandre et al., 2010; Winter et al., 2007) and clear cell (Mackay et al., 2010; Sugiyama et al., 2000; Winter et al., 2007) patient overall survival was significantly lower than in serous. Another study discovered that for stage I tumours, serous and clear cell had significantly worse prognosis than endometrioid (Gilks et al., 2008). This same study also determined that in stage II tumours, serous and mucinous had worse prognosis than endometrioid. Also for stage III tumours, clear cell had worse prognosis than serous (Gilks et al., 2008). When introducing the separation of serous EOC into low grade and high grade, Bodurka et al. (2012) established that LGSOC had a lower risk of death than HGSOC and significantly longer progression-free survival. Post 2012, overall survival has been compared between all five histotypes both with and without defined development stage. Similar to previous studies, mucinous and clear cell

histotypes were found to have worse prognosis at an advanced stage of development when compared to other histotypes (Lan & Yang, 2019; Zhou et al., 2018b; Mizuno et al., 2015; Simons et al., 2017, 2015; Karabuk et al., 2013). Whereas, Peres et al. (2019) established that for localised disease ten years post-diagnosis, HGSOC and mucinous had the worst prognosis. Both Peres et al. (2019) and Lan & Yang (2019) found that regardless of stage, LGSOC and endometrioid had the best prognosis with LGSOC having the best overall prognosis for each time interval. A study observed that early stage clear cell EOC had significantly greater three-year overall and progression-free survival than patients with advanced clear cell EOC (Zhu et al., 2021).

The differences in survival rates and prognosis across the EOC subtypes highlights the importance of a fast, accurate diagnosis. This is vital for efficient treatment decisions and administration that could improve the survival rates and prognosis of OC patients. Particularly with survival rates and prognosis being affected by a tumour's stage of development.

Chemotherapy Response

A consistent statement throughout analysis of chemotherapy response in EOC histotypes is that mucinous EOC has a limited response to platinum-based chemotherapy; mucinous response is significantly lower than the response in serous EOC (Alexandre et al., 2010; Pisano et al., 2005; Hess et al., 2004; Bamias et al., 2010; Pectasides et al., 2005; Simons et al., 2015; Karabuk et al., 2013; Shimada et al., 2009). Similarly, it has been indicated that clear cell EOC has high levels of platinum-based chemo-resistance with these levels being significantly greater than in serous OC (Sugiyama et al., 2000; Itamochi et al., 2002, 2008). Shu et al. (2015) suggested that advanced clear cell tumours have higher platinum-based chemo-resistance than early stage clear cell tumours. LGSOC has also been found to hold platinum-based chemo-resistant features. For neoadjuvant chemotherapy, only 4% of patients had complete response implying resistance (Schmeler et al., 2008).

For comparison of advanced LGSOC and HGSOC response to first-line platinum-chemotherapy, Grabowski et al. (2016) discovered that HGSOC response rate was significantly greater than that of LGSOC with prior sub-optimal debulking surgery.

Endometrioid EOC is discussed as having a frequent response to first-line platinum-based chemotherapy but high rates of relapse (Tomasi Cont, 2015). Oseledchyk et al. (2017) concluded that grade 3, stage IC endometrioid EOC tumours had significant positive changes in five-year overall survival when receiving adjuvant chemotherapy.

Contradicting results are found when researching adjuvant chemotherapy

use in clear cell, stage I tumours. Where studies have found that adjuvant chemotherapy in stage IA and IB clear cell tumours have significantly improved overall survival (Nasioudis et al., 2018), the opposite result has also been observed (Bogani et al., 2020; Mizuno et al., 2012; Oseledchyk et al., 2017).

Varying susceptibility to chemotherapy for EOC subtypes emphasises the importance of drug developments specific to individual subtypes. Additionally, it emphasises the need to accurately diagnose a patient with the correct EOC subtype in order to provide treatments that their tumour will most successfully respond to.

Debulking Surgery Response

Grabowski et al. (2016) found that in the case of both advanced HGSOC and LGSOC, complete resection in primary debulking surgery indicated a significantly better progression-free and overall survival than those not achieving complete resection. This conclusion was observed for LGSOC in a number of studies (Fader et al., 2013; Previs et al., 2014; Gershenson et al., 2015).

Previous studies suggested that for advanced clear cell EOC patients to have improved prognosis, complete resection is needed during debulking surgery (Takano et al., 2006). Mackay et al. (2010) identified that late stage mucinous and clear cell EOC had a significantly greater chance of being completely resected than serous EOC. Whereas, Alexandre et al. (2010) reported that mucinous had a lower frequency of optimally debulked stage II-III patients compared with serous.

1.2.2 Genetic Features (Known Mutations)

A number of genetic features have previously been studied to determine whether they could be indicative of specific EOC histological subtype presence. A few of these genetic features will be discussed below.

KRAS mutations

KRAS mutations are found to be most frequent in mucinous EOC tumours; a number of studies have found that between 60-80% of mucinous EOC tumours present KRAS mutations (Mackenzie et al., 2015; Mueller et al., 2018; Cheasley et al., 2019; Lee et al., 2016; Chang et al., 2016; Gorringe et al., 2020). The proportion of LGSOC tumours presenting KRAS mutations is lower than that in mucinous OC tumours, ranging from 23-35% (ElNaggar et al., 2022; Cheasley et al., 2021; Gershenson et al., 2022; Singer et al.,

2003). It was also suggested that KRAS mutations are frequent in recurrent LGSOC (Tsang et al., 2013). KRAS mutations are reported to be present in 9-20% of clear cell EOC tumours (Itamochi et al., 2017; Friedlander et al., 2016; Murakami et al., 2017; Kim et al., 2018; Shibuya et al., 2018) and 15-42% of endometrioid EOC tumours (Hollis et al., 2020; Cybulska et al., 2019; Pierson et al., 2020). KRAS mutations are low in HGSOC; multiple studies found that less than 10% of HGSOC tumours had KRAS mutation presence (Yang et al., 2018b; Singer et al., 2002, 2003; Cybulska et al., 2019).

BRAF mutations

BRAF gene mutations have oncogenic properties which lead to uncontrolled cell growth and division along with preventing apoptosis (Croce et al., 2019). BRAF mutations have been suggested to be in 2-38% of LGSOC; advanced stage LGSOC have the lowest frequency of BRAF mutations indicating an association between high mutation frequency and early stage LGSOC (Grisham et al., 2013; Cheasley et al., 2021; Gershenson et al., 2022; ElNaggar et al., 2022; Jones et al., 2012; Singer et al., 2003). A low frequency of BRAF mutations are seen in clear cell EOC tumours; mutations were found in no tumours or 2% of tumours in the studies by Singer et al. (2003) and Itamochi et al. (2017), respectively. Singer et al. (2003) also found that 24% of endometrioid tumours had BRAF mutations. Mucinous EOC tumours also show low presence of BRAF mutations; studies have found that between 5-12% of mucinous tumours show BRAF mutations (Mackenzie et al., 2015; Cheasley et al., 2019; Gorringe et al., 2020).

PTEN mutations

PTEN is a tumour suppressing gene, whose mutation leads to uncontrolled cell growth and division, promoting tumour growth (Song et al., 2012). PTEN mutations are rarely found in HGSOC; studies found these mutations in less than 10% of HGSOC tumours (Yang et al., 2018b; Singer et al., 2002, 2003; Cybulska et al., 2019). Similarly, clear cell EOC tumours are found to have a low frequency of PTEN mutations with approximately 2-5% of tumours presenting these mutations (Itamochi et al., 2017; Murakami et al., 2017). Mucinous EOC tumours also have low presentation of PTEN tumours, ranging from 2 to 9% of tumours (Mackenzie et al., 2015; Gorringe et al., 2020). Endometrioid EOC tumours present PTEN mutations in greater proportion; Hollis et al. (2020) found 29% of tumours had PTEN mutations and Parra-Herran et al. (2017) found that 44% of tumours had PTEN loss. McConechy et al. (2014) states that PTEN mutations are of

lower frequency in low-grade endometrioid OC.

PIK3CA mutations

PIK3CA mutations are found in less than 10% of HGSOC tumours (Yang et al., 2018b; Singer et al., 2002; Cybulska et al., 2019). These mutations are also low in mucinous EOC tumours with studies finding 8-14% of tumours presenting this mutation (Cheasley et al., 2019; Gorringe et al., 2020; Mackenzie et al., 2015). Endometrioid and clear cell EOC tumours have greater frequency of PIK3CA mutations. Studies show PIK3CA mutations are presented in 35-51% of clear cell tumours (Itamochi et al., 2017; Kim et al., 2018; Okamoto et al., 2014; Oda et al., 2018; Iida et al., 2021; Friedlander et al., 2016; Bolton et al., 2022; Murakami et al., 2017; Shibuya et al., 2018) and 31-43% of endometrioid tumours (Hollis et al., 2020; Cybulska et al., 2019; Pierson et al., 2020). Campbell et al. (2004) reported that 45% of endometrioid and clear cell tumours in their study displayed a mutation or amplification in this gene.

ARID1A mutations

ARID1A mutations are presented in 8-12% of both mucinous and LGSOC tumours (Cheasley et al., 2021, 2019; Gorringe et al., 2020). ARID1A mutations are reported in greater proportion of endometrioid tumours (17-36%) (Wiegand et al., 2010; Hollis et al., 2020; Cybulska et al., 2019; Parra-Herran et al., 2017). The histotype of EOC with the greatest presence of ARID1A mutations is clear cell with a reported 40-67% of tumours showing mutations (Wiegand et al., 2010; Jones et al., 2010; Maeda et al., 2010; Kim et al., 2018; Oda et al., 2018; Itamochi et al., 2017; Iida et al., 2021; Murakami et al., 2017; Shibuya et al., 2018).

CDKN2A mutations

Frequent deletions of CDKN2A are reported in LGSOC (Cheasley et al., 2021). A small proportion (6%) of endometrioid tumours are reported to have CDKN2A mutations (Cybulska et al., 2019). Mutations in mucinous tumours are more frequent ranging from 19-76% (Mackenzie et al., 2015; Mueller et al., 2018; Cheasley et al., 2019).

TP53 mutations

Many studies agree that TP53 mutations present in more than 90% of HGSOC tumours; multiple studies found this was the case for 96% of HGSOC

cases (Ahmed et al., 2010; Cancer Genome Atlas Research Network et al., 2011; Yang et al., 2018b; Cole et al., 2016; Cybulska et al., 2019). In comparison, LGSOC has low frequency of TP53 mutations with some studies finding no TP53 mutations (Wong et al., 2010). TP53 mutations have been detected in less than 20% of clear cell tumours (Friedlander et al., 2016; Bolton et al., 2022). Mucinous tumours show presence of TP53 mutations in greater frequency than clear cell tumours but not as frequently as HGSOC tumours. Studies have found that 49-75% of mucinous tumours showed presence of TP53 mutations (Mackenzie et al., 2015; Mueller et al., 2018; Cheasley et al., 2019; Kang et al., 2021a; Gorringe et al., 2020).

CCNE1 Amplification

CCNE1 is over-expressed in 22-30% of HGSOC (Kuhn et al., 2016; Etemadmoghadam et al., 2013; Stronach et al., 2018; Margolis et al., 2021; Noske et al., 2015).

ERBB2/HER2 Mutations/Amplification

Studies have found that ERBB2/HER2 is amplified in 11% of clear cell tumours and mutated in less than 10% (Itamochi et al., 2017). However, ERBB2/HER2 amplifications have been found in greater frequency in mucinous tumours; studies found them in 19-33% of mucinous tumours (Anglesio et al., 2013; Cheasley et al., 2019; Goundiam et al., 2015; Chang et al., 2016; Gorringe et al., 2020).

PPP2R1A Mutations

This gene is found to be mutated in both clear cell (4-19%) and endometrioid tumours (12%) (McConechy et al., 2011; Murakami et al., 2017; Shibuya et al., 2018).

CTNNB1 Mutations

CTNNB1 mutations are most frequently found in endometrioid tumours. Studies have found that 25-43% of endometrioid tumours present these mutations (Hollis et al., 2020; Cybulska et al., 2019). McConechy et al. (2014) found CTNNB1 mutations in 53% of low grade endometrioid OC tumours in their study. Another study by Palacios & Gamallo (1998) found that CTNNB1 mutations were observed in 38% of endometrioid tumour samples, presenting as early stage, grade I/II.

WT1 Staining

WT1 staining is positive in both high- and low-grade serous tumours (Köbel et al., 2008; Ali et al., 2013; Köbel et al., 2016). Whereas, WT1 staining is negative in clear cell, endometrioid and mucinous tumours (Köbel et al., 2008; Köbel et al., 2016; McCluggage, 2011). In the cases of Köbel et al. (2008) and Ali et al. (2013), positive staining is defined as $> 5\%$ of tumour cells presenting nuclear staining. Köbel et al. (2016) defines is at $> 1\%$ of tumour cells. Negative staining has a general definition of absence of any nuclear staining in tumour cells (Köbel et al., 2008; Ali et al., 2013; Köbel et al., 2016).

Estrogen Receptor (ER) and Progesterone Receptor (PR) Staining

Positive staining for PR, (defined as $> 1\%$ of tumour cells presenting nuclear staining), has been observed in greater frequency for endometrioid and LGSOC tumours and found lowest in mucinous and clear cell tumours (Sieh et al., 2013). Studies suggest the greatest PR expression can be found in endometrioid tumours (Lee et al., 2005; Geisler et al., 2008). One study found that ER is more highly expressed in HGSOC and endometrioid than other subtypes (Lee et al., 2005). However, multiple studies have identified that ER expression is greater in LGSOC than HGSOC (Escobar et al., 2013), with as many as 96% of LGSOC patients in one study exhibiting ER expression (Fader et al., 2017; Llaurado Fernandez et al., 2020). Furthermore, greater expression of PR has been observed in LGSOC than HGSOC (Escobar et al., 2013). Köbel et al. (2008) identified low levels of PR (3%) and ER (10%) in clear cell EOC tumours. Overall, it is reported that ER staining is greater in all histotypes than PR staining (Sieh et al., 2013).

BRCA1 and BRCA2 Mutations

A study observed that 17.1% of HGSOC tumours exhibited germline BRCA1/2 mutations (Alsop et al., 2012a,b). This same study found that approximately 16.6% of all serous EOC tumours presented with BRCA mutations and this was in greater proportion compared to other EOC subtypes Alsop et al. (2012b). BRCA1 mutations have previously been associated with an immunoreactive molecular subtype of HGSOC (George et al., 2013).

Another study observed that 85% of invasive serous patients carried a mutation (Risch et al., 2006). A study prior to this also found that invasive serous cancer patients accounted for 93% of patients with BRCA1/2 mutations and were twice as likely to carry BRCA1 versus BRCA2 mutations (Risch et al., 2001). This study also identified four women with endometrioid

tumours carrying BRCA1/2 mutations (Risch et al., 2001). Further evidence of high BRCA1/2 mutations in HGSOC was provided by Pal et al. (2005); the study found that of the BRCA1/2-mutated patients, 63% of them had been diagnosed with invasive serous EOC.

Norquist et al. (2016) observed that LGSOC had fewer BRCA1 and BRCA2 mutations than HGSOC. Additionally, clear cell EOC tumours had a significantly lower overall frequency of BRCA1/2 mutations than HGSOC at 8.6% and 19.6%, respectively (Norquist et al., 2016).

More recently, frequencies of 25.7% and 0% for BRCA mutations were observed in HGSOC and HGEOC, respectively (Manchana et al., 2019).

It is commonly observed that mucinous EOC tumours do not exhibit BRCA1/2 mutations (Norquist et al., 2016; Pal et al., 2005; Risch et al., 2001, 2006).

1.3 Classification

Classification is the process of grouping objects into predetermined categories. It is often used for determining similarities between observations but is also useful for the detection of anomalies. For example, in the financial sector, detection of anomalies is critical for the prevention of fraud so classification is used to predict fraudulent activity (Dornadula & Geetha, 2019; Thennakoon et al., 2019; Mittal & Tyagi, 2019; Martin et al., 2022). The cyber-security sector also uses classification models to detect anomalous emails to prevent cyber-security attacks (Sahingoz et al., 2019; Alhogail & Alsabih, 2021; Rawal et al., 2017). Classification is also used in many businesses for the prediction of product demand (Huber & Stuckenschmidt, 2020; Antunes et al., 2018; Abera & Khedkar, 2020).

1.3.1 Overview of classification techniques

Supervised versus unsupervised

Classification techniques can be described as supervised or unsupervised. Supervised classification is based on the prior knowledge of class labels. The chosen classifier trains a model on data where cases have a predefined class label and then uses this information to predict the class label of new data (Saravanan & Sujatha, 2018). Whereas, unsupervised classification does not consider predefined class labels; the data is grouped based on similarities or dissimilarities between provided variables (Saravanan & Sujatha, 2018).

Examples of supervised classification methods include support vector machines (Vapnik, 1999), naïve Bayes (Aggarwal, 2014c), logistic regression (Ag-

garwal, 2014d), decision trees (Aggarwal, 2014a), random forest (Breiman, 2001), K-nearest neighbours (Cover & Hart, 1967) and linear discriminant analysis (Aggarwal, 2014b). Unsupervised classification methods include k-means clustering (Hastie et al., 2009a), hierarchical clustering (Hastie et al., 2009b) and principal component analysis (Everitt & Dunn, 1991a).

Single versus ensemble classifiers

Most examples of supervised classifiers given above are single classifiers - one model providing one result. In the case of an ensemble classifier, a number of single classifier models are trained and their results are then combined to provide a single prediction (Hastie et al., 2009c). For example, the random forest classifier produces a number of decision trees and then uses majority voting to predict the class label of a new case (Breiman, 2001).

1.4 Classification in Disease

Classification is used for a variety of diseases in the context of diagnosis, prognosis prediction, treatment response prediction, recurrence prediction and disease progression prediction. The following section will provide insights into the use of classification for a few specific diseases.

1.4.1 Cervical Cancer

An abundance of studies have discussed the application of classification and machine learning models in the diagnosis and detection of cervical cancer. For example, the use of pathological imaging combined with classification techniques such as random forest models and convolutional neural networks are employed to predict metastasis risk and recurrence in cervical cancer patients (Ye et al., 2022b). Convolutional neural networks have also been presented to distinguish between normal and abnormal cervixes through both pap smear and liquid based cytology (Zhang et al., 2017b). A study found that random forest models for detecting cervical cancer through pap smear imaging were superior to K-nearest neighbours and decision trees (Diniz et al., 2021). Another study found that the implementation of features that describe the texture of patient sample images into artificial neural networks and support vector machines improved detection of cervical cancer (Arya et al., 2018). Support vector machines have also been used for prediction of patient outcome in cervical cancer based on pre-treatment MRI images (Torheim et al., 2014).

1.4.2 Diabetes

A number of studies have utilised classification models in diabetes research. Classification models such as decision trees, support vector machines, naïve Bayes, random forest, K-nearest neighbours, logistic regression, linear discriminant analysis, gradient boosting and k-means clustering are some of the models used in order to detect diabetes presence in its earliest stage (Sisodia & Sisodia, 2018; Sneha & Gangil, 2019; Abdulhadi & Al-Mousa, 2021) and distinguish between diabetic and non-diabetic patients (Mujumdar & Vaidehi, 2019; Birjais et al., 2019). Classification techniques have also been used to predict complications due to diabetes (Ahlqvist et al., 2018). Complications of diabetes that studies have considered for prediction include retinopathy (S K & P, 2017; Gadekallu et al., 2023; Tsao et al., 2018; Dagliati et al., 2018; Gulshan et al., 2016), nephropathy (Dagliati et al., 2018; Rodriguez-Romero et al., 2019; Huang et al., 2015), neuropathy (Kazemi et al., 2016) and cardiovascular disease (Hossain et al., 2021).

Other ways classification and machine learning models have been utilised in diabetes include predicting the risk of diabetes development due to lifestyle and family history (Tigga & Garg, 2020) and prediction of exercise response in pre-diabetic patients (Liu et al., 2020b).

1.4.3 Heart Disease

Studies have employed classification and machine learning techniques such as naïve Bayes, random forest, support vector machines, logistic regression, K-nearest neighbours, artificial neural networks, decision trees and random forests to indicate the presence of heart disease in patients (Gárate-Escamila et al., 2020; Otoom et al., 2015; Vembandasamy et al., 2015; Dwivedi, 2018; Shah et al., 2020; Ali et al., 2021).

1.4.4 Covid-19

Covid-19 presence has previously been predicted by use of neural networks based on X-ray imaging (Ozturk et al., 2020). Gradient boosting models have predicted Covid-19 test results (Zoabi et al., 2021) and K-nearest neighbours, naïve Bayes, random forest and support vector machines were compared when determining Covid-19 presence through blood testing (Brinati et al., 2020; Tschoellitsch et al., 2020). XGRboost and logistic regression models have also allowed for prediction of Covid-19 presence (Dong et al., 2022).

Prognosis of patients with Covid-19 was also predicted through use of various classification models. For example, support vector machines (Booth

et al., 2021), logistic regression (Hu et al., 2020) and gradient boosting decision trees (Li et al., 2023) have been used for mortality prediction.

1.4.5 Alzheimer's Disease

A frequently researched application of classification is Alzheimer's disease. A common area using classification models is determining the stage of a patients disease. This includes distinguishing between Alzheimer's, mild cognitive impairment and healthy patients (Long et al., 2017; Sørensen et al., 2016; Shankar et al., 2022; Acharya et al., 2019; Altaf et al., 2018). Classification models tested on the prediction of disease stage include support vector machines (Long et al., 2017; Sørensen et al., 2016; Shankar et al., 2022; Altaf et al., 2018), naïve Bayes (Shankar et al., 2022), decision trees (Shankar et al., 2022; Altaf et al., 2018) and K-nearest neighbours (Shankar et al., 2022; Altaf et al., 2018). Support vector machines have further been selected to distinguish between Alzheimer's and behavioural variant frontotemporal dementia (Möller et al., 2016).

1.4.6 Breast cancer

Studies have used classification models such as naïve Bayes, K-nearest neighbours, decision trees, artificial neural networks, random forests, support vector machines and logistic regression to aid diagnosis of patients with malignant or benign breast cancer (Amrane et al., 2018; Shan et al., 2016; Islam et al., 2020).

Classification models have also allowed for division of breast cancer patients into subtypes. A random forest classifier found that immunogenomic profiling separated triple negative breast cancer patients into three subgroups that related to prognosis of said patients (He et al., 2018). Triple negative breast cancer is a subtype of breast cancer characterized by limited expression of ER, PR and HER2 (Gluz et al., 2009). Furthermore, survival of varying breast cancer subtypes has previously been predicted through use of random forest classifiers (Montazeri et al., 2016).

Classification of patients with varying progression risks in breast cancer has also been studied; the ability to group patients by high or low risk was considered by Ferroni et al. (2019).

Treatment response prediction can be observed by use of classification models. Support vector machines and random forest models are two examples described in the literature that are used to determine response to treatment in breast cancer (Dorman et al., 2016; Meti et al., 2021).

1.5 Classification in Ovarian Cancer

This section will provide an overview of the many ways in which classification has been used in OC research.

1.5.1 Ovarian Cancer Presence.

EOC vs Healthy

A number of studies have considered the use of classification for identifying the presence of ovarian cancer. For initial diagnosis of OC, the ability to distinguish healthy patients from OC patients is critical. With serous EOC being the prevalent histological subtype, an aim in literature has been to differentiate patients with serous EOC from healthy patients resulting in varying accuracies (Sans et al., 2019; Boylan et al., 2017; Han et al., 2018). A validation accuracy of 95% and a sensitivity of 100% was achieved using a biomarker signature of small metabolites to differentiate between serous and healthy samples (Sans et al., 2019). This same study was able to achieve a validation accuracy of 92.3% when differentiating between HGSOC and healthy samples (Sans et al., 2019). Han et al. (2018) provided a 4-marker signature that achieved a ROC-AUC of 0.9896 when comparing HGSOC samples with healthy samples. Barnabas et al. (2019) suggested a 9-marker signature that achieved a sensitivity of 70% and specificity of 76.2% when distinguishing between HGSOC and healthy samples. Boylan et al. (2017) separated serous EOC samples into early and late stage to provide a comparison with healthy samples. For a specificity of 95%, Boylan et al. (2017) found that the biomarker CA125 achieved a sensitivity of 100% for comparing late stage serous EOC with healthy samples. Whereas, for early stage serous EOC compared with healthy samples, a 12 protein signature was able to achieve a sensitivity of 95.2% at a specificity of 95% when using naïve Bayes classification. Use of principal component analysis and hierarchical clustering could also separate serous patients from healthy patients regardless of diagnosis stage (Boylan et al., 2017). Yu et al. (2020) was able to distinguish between serous cancerous EOC samples and non-cancerous samples through the use of a convolutional neural network based on imaging with an AUC greater than 0.95.

Suryawanshi et al. (2013) aimed to distinguish between endometrioisis-related EOC (clear cell and endometrioid (Kurman & Shih, 2016)) and healthy samples achieving a sensitivity and specificity of 86% and 85%, respectively. This study used linear discriminant analysis based on a signature of plasma miRNA (Suryawanshi et al., 2013).

EOC vs benign tumours

Studies have also considered the use of classification techniques for determining whether a patient has a benign or malignant tumour. For example, Vanderstichele et al. (2017) used chromosomal instability in cell free DNA to differentiate between HGSOC samples and benign samples achieving a ROC-AUC of 0.94. When considering a combination of borderline and invasive tumours versus benign tumours, this same study achieved a ROC-AUC of 0.89 (Vanderstichele et al., 2017). Han et al. (2018) achieved a ROC-AUC of 0.9608 when comparing benign tumours with HGSOC tumours. Kawakami et al. (2019) tested multiple classification techniques to predict the malignancy or benignity of a tumour; both gradient boosting machine and random forest presented the greatest classification accuracy of 93.7%. Ke et al. (2015) provided a marker signature that achieved a ROC-AUC of 0.91 for EOC compared with benign samples and 0.8385 for early stage EOC compared with benign samples. This was an improvement on the use of the single standard biomarker CA125. Prior to the study by Ke et al. (2015), Acharya et al. (2014) identified 11 features that could classify benign and OC tumours with 100% accuracy through use of K-nearest neighbour and probabilistic neural network classification methods. The use of a 2-marker signature by Lu et al. (2020b) could predict malignant versus benign tumours with accuracies of 92.1% and 97.4% using decision tree and logistic regression models, respectively. Further studies suggest marker signatures allowing distinction between benign and OC tumours including Enroth et al. (2019) who achieved a model with 99% sensitivity and 100% specificity when classifying benign tumours versus stage III-IV OC tumours. Ultrasound imaging of ovaries has also allowed the classification of malignant tumours versus benign tumours. Aramendía-Vidaurreta et al. (2016) discovered that particular features of ultrasound images combined with patient age gave a classification accuracy of 98.78% with both sensitivity and specificity greater than 98%.

1.5.2 Stages

Both early stage versus late stage and stage I versus stages II-IV classification was performed by Ke et al. (2015) yielding ROC-AUC values of 0.8801 and 0.9557, respectively, using 53 metabolite markers. A random forest classifier was able to predict the clinical stage of EOC patients with 69% accuracy when stages were grouped as early (I-II) and late (III-IV) (Kawakami et al., 2019).

1.5.3 Subtypes

Molecular Subtypes

Predicting the molecular subtypes of individual EOC histotypes is one way in which classification models have been utilised. For example, classification models have distinguished between epithelial-like and mesenchymal-like clear cell EOC; both subgroups were associated with differing development stages and time to disease progression or death (progression-free survival) (Tan et al., 2019). Four molecular subtypes of HGSOC and endometrioid OC have been defined as immunoreactive, differentiated, proliferative and mesenchymal (Cancer Genome Atlas Research Network et al., 2011; Tothill et al., 2008). Leong et al. (2015) later provided a 48 gene signature that was able to distinguish between molecular subtypes of HGSOC with 100% and 80% accuracy for fresh frozen and formalin-fixed, paraffin-embedded samples, respectively.

Histological Subtypes

The clustering of clear cell EOC samples based on a number of clinical features was able to identify an advanced stage, poorer outcome TP53 mutated cluster with abnormal p53 expression and a second cluster containing samples that suggested cells with abnormal chromosome numbers and mutations in ARID1A/PIK3CA (Cunningham et al., 2022).

Multiple studies discuss the classification of HGSOC versus endometrioid EOC. Assem et al. (2018) presented the differentiation of grade 3 endometrioid and HGSOC samples with an accuracy of 69%. However, the use of known WT1 and p53 status was able to improve this accuracy to 96% (Assem et al., 2018). A later study by Dieters-Castator et al. (2019) classified endometrioid EOC and HGSOC samples with an accuracy of 99.2% using logistic regression.

Multiple studies examining the classification of EOC histotypes focused on imaging features rather than biological features with accuracies ranging from $78.6 - 95\%$ (BenTaieb et al., 2016; Farahani et al., 2022; An et al., 2021). Farahani et al. (2022) determined their best convolution neural network model had a test accuracy of 80.97%; all clear cell samples were correctly classified in this case. However, approximately half of the test samples corresponding to endometrioid EOC were misclassified (Farahani et al., 2022). BenTaieb et al. (2016) was able to perform support vector machine multiclass classification using tumour images on validation data with an accuracy of 95%. When considering binary classification, clear cell EOC and HGSOC could be distinguished with 100% accuracy unlike endometrioid EOC and

LGSOC (BenTaieb et al., 2016). Overall, the mean classification accuracies comparing HGSOC, LGSOC, mucinous, clear cell and endometrioid EOC with all other subtypes were approximately 88.75%, 73.75%, 66.5%, 82% and 63.5%, respectively (BenTaieb et al., 2016).

Both Köbel et al. (2016) and Klein et al. (2019) used biological features to classify patients by EOC histotype. A signature of eight immunohistochemical biomarkers classified patients with an accuracy of 93% with endometrioid EOC having the greatest classification error (Köbel et al., 2016). Klein et al. (2019) performed classification on HGSOC, LGSOC, clear cell and serous borderline tumour samples considering features from MALDI imaging finding an accuracy of 85% for the convolutional neural network model. Schwartz et al. (2002) has previously proposed a gene signature consisting of 158 genes that was able to classify all non-clear cell samples and all but one clear cell sample correctly.

Binary classification models have also been used to distinguish between individual EOC histotypes. For example, Woodbeck et al. (2019) discovered that the combination of markers PR and VIM could predict whether a patient had mucinous or endometrioid EOC with an accuracy of 96%. As previously discussed, tumour imaging provided 100% accuracy in differentiating between clear cell and HGSOC (BenTaieb et al., 2016).

Using 32 blood peripheral biomarkers, Kawakami et al. (2019) was able to predict the histological subtype of an ovarian cancer tumour through random forest models with accuracies of 75.8%, 67.7%, 55.6%, and 96.0% and ROC-AUC of 0.785, 0.650, 0.597, and 0.728 for histotypes HGSOC, clear cell, endometrioid and mucinous, respectively.

A 3 miRNA signature was proposed to distinguish between endometriosis related EOC (clear cell and endometrioid histology) and serous EOC achieving a sensitivity and specificity of 86% and 79%, respectively (Suryawanshi et al., 2013).

Another method for classification of histological subtypes has been to classify HGSOC versus non-HGSOC patients. This has also been described as classifying type I versus type II EOC patients. The use of traditional MRI imaging and logistic regression by Qian et al. (2020) achieved a sensitivity and specificity of 0.88 and 0.97, respectively. Zhang et al. (2019) didn't achieve a sensitivity or specificity as high when classifying type I versus type II tumours using MRI radiomics. The sensitivity and specificity achieved in this study were 0.7647 and 0.8649, respectively (Zhang et al., 2019). Wang et al. (2022a) made use of CT imaging features and logistic regression to distinguish between HGSOC and non-HGSOC subtypes. The testing sensitivity and specificity of this model were 0.799 and 0.784, respectively (Wang et al., 2022a). An et al. (2021) provided random forest classification models to

distinguish HGSOC samples from non-HGSOC samples (LGSOC, clear cell, mucinous and endometrioid), using morphological features or texture combined with clinical features; the former combination achieved an accuracy of 78.6% and the latter an accuracy of 83.3%.

1.5.4 Treatment Response

Chemotherapy Response

The prediction of chemotherapy response among patients is vital for providing patients with the most effective treatment. Through use of biomarkers, response to chemotherapy has been predicted in patient samples. Sun et al. (2016) could predict HGSOC chemotherapy resistance with 83.9% accuracy using a radial basis kernel support vector machine. Similarly, Huang et al. (2018) was able to predict chemotherapy response in ovarian cancer with an accuracy exceeding 80% by support vector machine. Following this study, Lu et al. (2019) used a support vector machine model with a gene signature of length 10 to predict an OC cell line's response to chemotherapy (low, medium or high). In this study, validation of the model suggested high accuracy in classification due to the samples in high response group having a significantly longer recurrence-free and overall survival than the low response group (Lu et al., 2019).

Through the use of imaging, Yu et al. (2020) could distinguish between patients who had longer platinum-free intervals and those who had shorter when focusing on serous EOC patients. Platinum-free intervals provide information on the length of time before a patient had platinum-resistant recurrence, therefore, distinguishing between these two groups implies that the group with shorter platinum-free intervals is platinum-resistant (Yu et al., 2020).

Schilling et al. (2023) used a selection of classification models including naïve Bayes, logistic regression, support vector machines, random forest and XGBoost to predict platinum chemotherapy response of patients. For prediction of chemotherapy-resistance versus sensitivity, the greatest classification F1 score of 0.91 was achieved by the random forest model, correctly classifying 36 out of 38 patients using 172 genes (Schilling et al., 2023).

First-line platinum-based chemotherapy response in serous EOC patients using the expression of genes AGGF1 and MAP4 has also been studied with use of linear support vector machines and artificial neural networks (Zhao et al., 2019). The artificial neural net achieved greater ROC-AUC values of 0.8056 and 0.7245 for two separate data sets (Zhao et al., 2019).

Yi et al. (2021) combined pre-chemotherapy treatment CT imaging and

single-nucleotide polymorphisms of SULF1 to predict response to platinum-based chemotherapy in OC using support vector machines and random forest models. These combined features produced a validation AUC of 0.967 (Yi et al., 2021).

Buttarelli et al. (2022) used a random forest model to determine the response of BRCA wild-type, HGSOC patients to first-line platinum-based chemotherapy through use of a 10 gene signature. This model provided an accuracy of approximately 93% when classifying patients as chemotherapy resistant or sensitive, with a precision of approximately 94% (Buttarelli et al., 2022).

Hwangbo et al. (2021) used six variables associated with platinum-sensitivity, to predict sensitivity in HGSOC patients by training and validating logistic regression, random forest, support vector machine and deep neural network models. The study found that logistic regression was most successful with an AUC, sensitivity and specificity of 0.741, 0.778 and 0.622, respectively (Hwangbo et al., 2021).

Surgery Response

Artificial neural networks using features such as histology, grade, stage and CA125 levels in pre-operative patients' have previously been used to predict OC patients' response to surgery. Enshaei et al. (2015) classified optimal debulking versus suboptimal debulking in patients with an accuracy of 77% and ROC-AUC of 0.73. XGboost models have also predicted optimal debulking in advanced stage EOC patients with an AUC of 0.866 and an F1 score of 0.89 (Laios et al., 2022). Logistic regression using features such as disease score, stage, ascites and the interaction between stage and age predicted optimal debulking in EOC patients with an AUC of 0.83 (Horowitz et al., 2018). The use of CT imaging and clinical variables for patients with advanced stage EOC was able to predict gross residual disease with a AUC of 0.762 (Kumar et al., 2019). Lu et al. (2023) predicted residual disease in post-operative HGSOC patients with AUCs of 0.936 and 0.9 through use of pre-operative MRIs.

An ordinal classification model was produced and tested by Kawakami et al. (2019) to predict whether a patient would achieve complete resection or non-complete resection reaching an accuracy and ROC-AUC of 64.9% and 0.697, respectively. Furthermore, Kawakami et al. (2019) developed a model to distinguish patients who would achieve suboptimal resection from patients with non-suboptimal resection and produced an accuracy and ROC-AUC of 62.9% and 0.667, respectively.

Riester et al. (2014) predicted whether patients were at a high or low risk

of suboptimal resection post debulking surgery. Use of three proteins significantly associated with debulking status of late stage OC patients achieved a classification accuracy and ROC-AUC of 92.8% and 0.89, respectively (Riester et al., 2014).

1.5.5 Survival

Survival of patients with OC is commonly predicted by use of survival analysis. However, Paik et al. (2019) was able to produce a gradient boosting model to predict patient overall survival for the second year with a ROC-AUC of 0.843. The conventional Cox-proportional hazards model for this same data had a ROC-AUC of 0.597 (Paik et al., 2019). An artificial neural network proposed by Enshaei et al. (2015) was able to predict the overall survival of OC patients with an accuracy of 93% but a ROC-AUC of 0.74.

Zeng et al. (2021) developed a random forest model using a combination of 100 genes displaying high frequencies of somatic mutations in EOC tumours and histopathological image features. This combined model provided an AUC value of 0.834 for 5-year overall survival prediction in HGSOC patients, improving upon models using just histopathological image features alone (Zeng et al., 2021). A random forest model, developed by Arezzo et al. (2022), provided the greatest classification accuracy when compared with K-nearest neighbours and logistic regression for predicting 12 month progression-free survival in OC patients. Ultrasound imaging features along with age, menopause, CA125 levels, histotype and FIGO stage were used to train this model and led to an accuracy and AUC of 93.7% and 0.92, respectively (Arezzo et al., 2022). Furthermore, a random forest model provided an accuracy of 88.72% when predicting patient survival (Sorayaie Azar et al., 2022). Another random forest model was developed by use of genes USP19 and RPL23; Kang et al. (2021b) proposed that these genes could serve as prognostic biomarkers for HGSOC providing a prediction of prognosis with sensitivity and specificity of 0.67 and 0.92, respectively.

Schilling et al. (2023) used machine learning models to predict patient outcome, labelling a bad outcome as death within two years of treatment and a good outcome as survival for five or more years. This study found that logistic regression achieved the greatest classification F1 score of 0.88 using 149 genes. This model correctly classified 25 of 29 patients (Schilling et al., 2023).

1.5.6 Recurrence

Support vector machines have often been employed to predict the recurrence of EOC. Through the use of 19 mi-RNAs, the AUC of both training and test data were > 0.9 allowing for accurate recurrence prediction (Dong & Xu, 2019). Furthermore, 39 genes were considered to predict recurrence in EOC patients providing accuracies of 93.3% and 96.6% on validation data (Zhou et al., 2018a). Zhang et al. (2018) considered metabolic signatures pre- and post-surgery as features used to train support vector machine models. Individual use of the pre- and post-surgery signatures led to an AUC of 0.815 and 0.909, respectively. This AUC value increased to 0.964 when both signatures were combined to train the support vector machine model (Zhang et al., 2018). Another signature of six mi-RNAs combined with clinical features had an accuracy of 91.86% when predicting OC recurrence using an XGBoost model (Sujamol et al., 2021). Chen et al. (2021a) developed a support vector machine model combining CT radiomic features with FIGO stage and residual disease. This led to a HGSOC early recurrence prediction AUC of 0.749 and 0.769 in the training and test data, respectively. This combined model performed better than models using radiomic features or clinical features alone (Chen et al., 2021a). Another support vector machine model was developed combining clinical and MRI radiomic features leading to an AUC of 0.85 for predicting recurrence-free survival in advanced HGSOC (Li et al., 2021a).

1.6 Current Study

1.6.1 Aims

The ability to accurately detect and diagnose EOC in its earliest stages is vital to improvements in survival rates and prognosis of EOC patients. However, an overall diagnosis of EOC is not enough. Research into distinguishing between the five histological subtypes of EOC is essential. As discussed in section 1.2, the five subtypes display a number of differences. An accurate individual diagnosis of EOC subtype is crucial for factors such as treatment decisions. Varying responses to standard treatment emphasises the need for accurate subtype diagnosis. Without this, inadequate treatments will be administered to patients which will reduce treatment success. This in turn reduces the survival rates and negatively impacts prognosis of patients. Furthermore, survival rates and prognosis already differ between EOC subtypes so the ability to provide efficient and accurate subtype diagnosis should also improve speed of treatment decision and administration. As the prognosis

and survival rates differ not only between subtypes but also for tumour development stages, a method that identifies a patient's EOC subtype quickly is crucial.

Discussed in section 1.2.2, differences in the genetic make-up of these five histological subtypes have been identified. Hence, the use of gene signatures could be a successful method in determining the histological subtype of a patient's EOC tumour. Currently, even with known genetic differences, no set of genes can be used to accurately diagnose the subtype of a patient's EOC tumour.

Following on from this discussion, the overall purpose of this research considers two objectives: detection and diagnosis. The first objective considers whether genes can detect the presence or non-presence of EOC in a sample. The second objective then considers whether subsets of these genes can determine the histological subtype diagnosis of a sample that indicates EOC presence. This study aims to contribute individual gene signatures with promising diagnostic abilities with regards to specific EOC histological subtypes.

Binary classification models will be used to evaluate a gene signature's diagnostic abilities in this study. Gene expression will be the only feature considered when preparing classification models. This is due to missing data values for features such as age, stage and grade in the training cohort. Due to the rarity of particular EOC subtypes, a main limitation of this research is the data size available for analysis. In addition, a lack of prior studies applying a similar binary classification approach to the current study in this context means there are limited results available for direct comparison of classification predictions.

It is assumed that all tumour samples used in this study are primary tumours with correct information regarding their actual histological diagnosis.

Through reviewing the literature, identification and application of individual gene signatures for the diagnosis of a specific EOC subtype using binary classification techniques is sparse.

1.6.2 Thesis Overview

The thesis will be organised as follows. In chapter 1, the context and motivation behind this research has been introduced. Furthermore, the main aims of this study have been identified along with a brief description of this research's limitations, assumptions and contributions.

Chapter 2 will provide a description of the data used within this study. The choice of statistical and classification methods will be explored as well

as evaluation methods selected for determining the success of classification models.

Chapter 3 provides the results of statistical analysis methods described in chapter 2 for selecting gene signatures. This chapter displays the gene signatures selected for classification along with their ability to classify patients correctly by their EOC subtype for four classification cases. The cases are (1) clear cell versus non clear cell EOC, (2) mucinous versus non-mucinous EOC, (3) LGSOC versus non-LGSOC and (4) clear cell versus endometrioid EOC.

Chapter 4 provides an overview of the motivation behind this study. For each of the four EOC subtype classification cases considered, a brief description of the classification results will be provided. That is followed by a discussion of the corresponding gene signature and the individual genes previous associations with cancer and disease. Limitations of this study are also described along with future research that should follow.

Chapter 5 provides conclusions on the success of each classification case considered. Furthermore, it describes the possible clinical contribution of this study.

Chapter 2

Methodology

2.1 Training Data

The Gene Expression Omnibus (GEO) database was searched for ovarian cancer gene expression data. Data sets containing gene expression data for clear cell, endometrioid, mucinous, low grade serous or high grade serous EOC as well as normal ovaries were selected. The data sets were used for training of classification models in order to provide representation for each histological subtype. All of the data sets had the common platform GPL570 meaning they contained the same probe IDs. The data sets selected for classification training models were GSE26193, GSE65986, GSE39204, GSE44104, GSE52037, GSE54388 and GSE27651 (Mateescu et al., 2011; Uehara et al., 2015; Abiko et al., 2013; Wu et al., 2014; Hill et al., 2014; Yeung et al., 2017; King et al., 2011). The samples included in this training data could be separated into six categories defined in table 2.1. The samples were regarded as cases in the analysis.

Table 2.1: Table providing information on the samples' histological types in each of the training data sets.

Data	Clear Cell	Endometrioid	Mucinous	LGSOC	HGSOC	Normal Ovaries
GSE26193	6	8	8	4	75	0
GSE65986	25	14	0	0	0	0
GSE39204	16	13	2	0	0	0
GSE44104	12	11	9	0	0	0
GSE52037	0	0	0	0	0	10
GSE54388	0	0	0	0	0	6
GSE27651	0	0	0	13	0	7
TOTAL	59	46	19	17	75	23

Table 2.1 provides information on the sample size and sample histological diagnosis of each data set used for classification training data. A total of 239 samples were included in this training data set. The data consisted of $54,675$ unique probes that were regarded as variables in this analysis. Limited data was available for clinical features such as patient age, diagnosed stage of disease and diagnosed grade of disease. However, the data sets provided enough information on serous grade diagnosis for it to be separated into HGSOC and LGSOC. The data consisted of continuous variables as the variables were represented by gene expression values for each sample. The training data did not follow a normal distribution indicated by the quantile-quantile plots in figure 2.1.

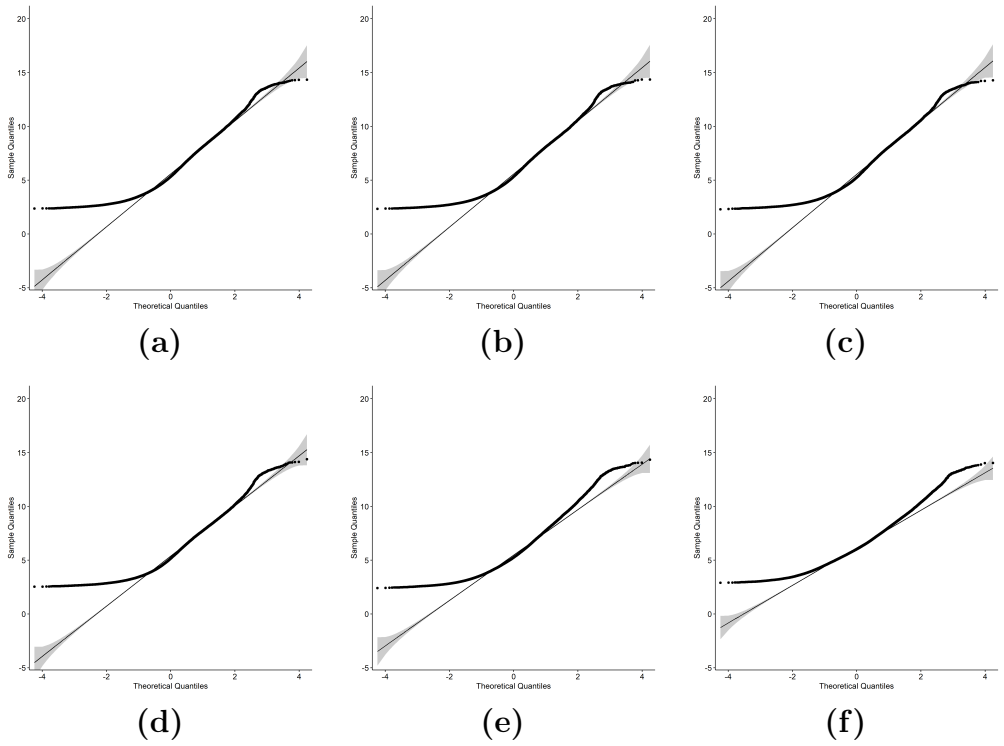


Figure 2.1: **(a)** QQ-plot for clear cell EOC training samples. **(b)** QQ-plot for endometrioid EOC training samples. **(c)** QQ-plot for mucinous EOC training samples. **(d)** QQ-plot for HGSOC training samples. **(e)** QQ-plot for LGSOC training samples. **(f)** QQ-plot for control training samples.

2.2 Testing Data

The data set GSE20565 (Meyniel et al., 2010) was employed for use as the test data to validate the classification models in this study. This data set was not used in the training of classification model. Therefore there was no need for the splitting of the training data to provide a test set. The samples (cases) included in this test data could be separated into five categories defined in table 2.2.

Table 2.2: Table providing information of the sample types in the test data set.

Data	Clear Cell	Endometrioid	Mucinous	LGSOC	HGSOC
GSE20565	6	6	7	3	67

Table 2.2 presents the test data set for classification validation. The test data set contained 89 samples; this data set did not contain any normal ovary samples as the overall purpose of the classification models was differentiation between EOC histotype samples. Similarly to the training data, information regarding clinical features was not complete so was not considered. However, enough information regarding serous grade diagnosis was available to separate serous samples into HGSOC and LGSOC. As the test data sets were taken from the same platform as the training data sets, the same $54,675$ probes were available for analysis as variables. These variables were continuous in nature as they represented gene expression values for each of the samples. The data did not follow a normal distribution as shown by the quantile-quantile plots in figure 2.2.

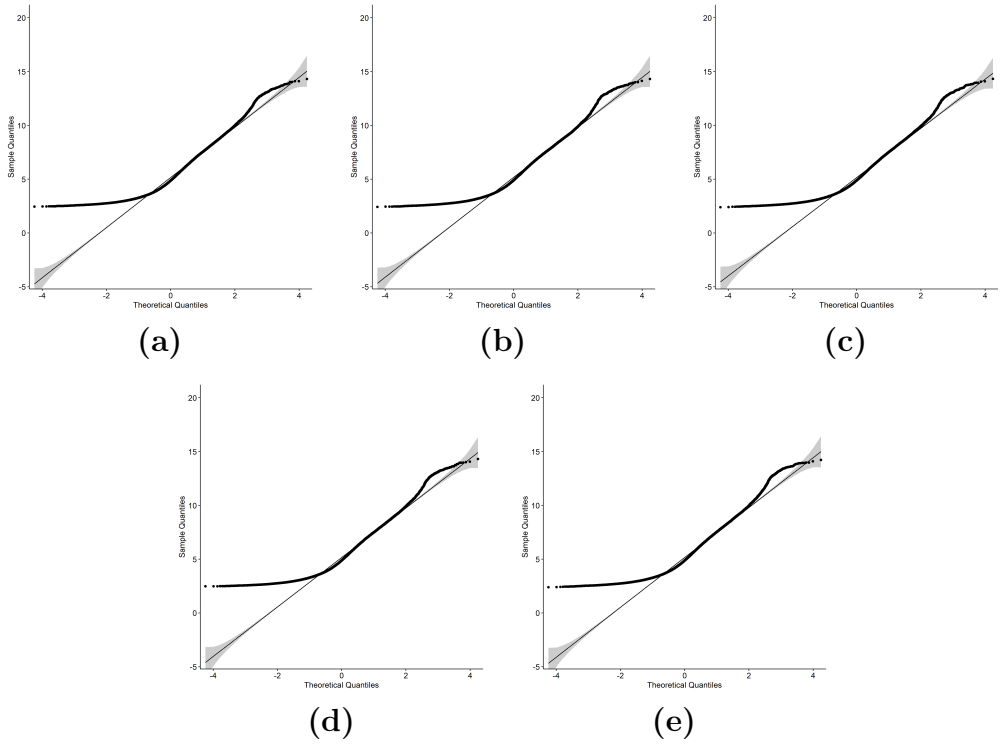


Figure 2.2: **(a)** QQ-plot for clear cell EOC test samples. **(b)** QQ-plot for endometrioid EOC test samples. **(c)** QQ-plot for mucinous EOC test samples. **(d)** QQ-plot for HGSOC test samples. **(e)** QQ-plot for LGSOC test samples.

2.3 Data Preparation

Following best practice as established by Gov & Arga (2017), Lisowska et al. (2016) and Bao et al. (2020), robust multi-array average (RMA) was performed using the function 'rma' in the 'affy' package from R (Bolstad et al., 2003; Irizarry et al., 2003a,b) on the raw data. Background corrections, log_2 transformations, quantile normalization and probe normalization was performed through this processing method (Bolstad et al., 2003; Irizarry et al., 2003a,b). The data was then returned as expression values for each gene through the use of the function 'exprs' in the package 'Biobase' (Huber et al., 2015). The training data was then combined into one large data set for analysis of differentially expressed genes. The test data was kept separately from training data.

Of the $54,675$ available unique probe IDs corresponding to probes that measure expression levels for genes, a number were found to have no corresponding gene symbol. Therefore, they were removed prior to analysis; $45,118$ probe IDs remained. The removal of these probes ensured that corresponding gene names could later be matched to selected probe IDs. All probe IDs, gene symbols and gene names were taken from the GPL570 platform.

The non-normality of the data distribution limited the statistical and classification techniques that could be performed. This was due to parametric assumptions in a number of common methods. Hence, the following methods were chosen due to being non-parametric in assumption preventing a violation in model assumptions.

2.4 Hypothesis testing

2.4.1 Fold Change Analysis

Fold change analysis, described in Quackenbush (2002) and (Deng et al., 2009), is the comparison of gene expression level between a control and case group. Let the mean expression level of a control group be denoted $\bar{x}_{control}$ and the mean expression level of a case group be denoted $\bar{x}_{case}$. Then for gene expression data, the fold change value of a gene is calculated as (Jeffery et al., 2006; Deng et al., 2009; Lee, 2014)

$$FC = \frac{\bar{x}_{case}}{\bar{x}_{control}} \tag{2.1}$$

However, as the expression data has been log_2 transformed, this ratio will calculate the log_2 fold change value.

$$log_2FC = log_2\left(\frac{(\bar{x}_{case})}{(\bar{x}_{control})}\right) \tag{2.2}$$

$$= log_2(\bar{x}_{case}) - log_2(\bar{x}_{control}) \tag{2.3}$$

Therefore, the fold change value of a gene between a case and control group is calculated as

$$FC = 2^{log_2(FC)} \tag{2.4}$$

Fold change analysis is then considered a hypothesis test. The fold change value is compared to a threshold C and considered dys-regulated for cases (Lee, 2014)

$$FC > C \quad or \quad FC < \frac{1}{C} \tag{2.5}$$

Here, the former indicates up-regulation and the latter indicates down-regulation of a gene in case samples. For this study, following Arend et al. (2018) and Gov (2020) we chose $C = 2$.

2.4.2 Wilcoxon Rank-Sum Test

The Wilcoxon rank-sum test is not dependent on the normality assumption of the data distribution. It does, however, depend on the assumption of independent randomly sampled data samples (Hollander et al., 2013). These assumptions are held for the data in this study.

The Wilcoxon rank-sum test is performed to determine whether the distribution of two groups are equal or not; Hollander et al. (2013) provides a description of the theory behind the Wilcoxon rank-sum test which will be discussed below.

Consider $K = 2$ groups for which $j = 1, 2$ with each group containing $i = 1, ..., n_j$ observations denoted X_{ij}. Here, let the total number of observations be $N = \sum_{j=1}^{K} n_j$. The population distributions of groups $j = 1$ and $j = 2$ are denoted as F_1 and F_2, respectively.

The Wilcoxon rank-sum test defines a null and alternative hypothesis, H_0 and H_A, respectively.

$$H_0 : F_1 = F_2 \tag{2.6}$$

$$H_A : F_1 \neq F_2 \tag{2.7}$$

The N observations are placed in ascending order and ranked. The test statistic, denoted W, is the sum of group $j = 1$ observation ranks, denoted

r_{i1}.

$$W = \sum_{i=1}^{n_1} r_{i1} \tag{2.8}$$

If the null hypothesis is true, we can calculate the expected value and variance of W, $E[W]$ and $Var[W]$, respectively.

$$E[W] = \frac{n_1(n_1 + n_2 + 1)}{2} \tag{2.9}$$

$$Var[W] = \frac{n_1 n_2(n_1 + n_2 + 1)}{12} \tag{2.10}$$

Standardizing W, we find that

$$W^* = \frac{W - E[W]}{\sqrt{Var[W]}} \tag{2.11}$$

It is then possible to find the standard normal statistic and compare this with the standardized test statistic W^* for a two sided test.

$$Reject \quad H_0 : |W^*| \geq z_{\alpha/2} \tag{2.12}$$

$$Not \quad reject \quad H_0 : -z_{\alpha/2} < W^* < z_{\alpha/2} \tag{2.13}$$

As this is a two-sided hypothesis test, a standard $\alpha = 0.05$ is selected.

2.4.3 Kruskal Wallis Test

The Kruskal Wallis test does not depend on the normality of the data distribution. The assumptions that are considered include the samples being randomly sampled and independent (Hollander et al., 2013). These assumptions are held for the data in this study.

In the same way as the Wilcoxon rank-sum test, the Kruskal Wallis test is performed to determine whether population distributions of groups are equal. In this case, there are $K > 2$ groups compared; Hollander et al. (2013) provides the methodology behind the Kruskal Wallis test which is described below.

As before, let there be $j = 1, ..., K$ groups and each group holds $i = 1, ..., n_j$ observations that are denoted X_{ij}. The total number of observations is $N = \sum_{j=1}^{K} n_j$. Denote the population distributions of these $i = 1, ..., K$ groups to be F_i. Then we consider the following null and alternative hypothesis.

$$H_0 : F_1 = F_2 = ... = F_K \tag{2.14}$$

$$H_A : not \quad all \quad F_1, ..., F_K \quad are \quad equal \tag{2.15}$$

The N observations are placed in ascending order and ranked. The rank of observation X_{ij} is denoted r_{ij}. Let

$$R_j = \sum_{i=1}^{n_j} r_{ij} \tag{2.16}$$

where R_j is the sum of observation ranks for group j. Then let $\bar{R}_j$ be the mean rank of the observations in group j.

$$\bar{R}_j = \frac{R_j}{n_j} \tag{2.17}$$

The Kruskal Wallis test statistic, denoted H, is calculated by the following equation.

$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{K} \frac{R_j^2}{n_j} \right) - 3(N+1) \tag{2.18}$$

When H_0 is true, H can be approximated by the χ^2_{K-1} distribution. Therefore it is found that

$$Reject\, H_0 : H \geq \chi^2_{K-1,\alpha} \tag{2.19}$$

$$Not\, reject\, H_0 : H < \chi^2_{K-1,\alpha} \tag{2.20}$$

Here, a standard $\alpha = 0.05$ is used.

2.4.4 Dunn Test

The Dunn test, introduced by Dunn (1964), is used as a post-hoc test for the Kruskal Wallis test to determine which of the group populations differ if H_0 is rejected.

Following from the Kruskal Wallis test in section 2.4.3, there are K groups and we will consider that at least one pair of these K groups have significantly different population distributions. The following description of the Dunn test is described by Dunn (1964) and Dinno (2015). The Dunn test performs l pairwise comparisons between groups.

The test statistic for this test is calculated as

$$z_m = \frac{y_m}{\sigma_m} \tag{2.21}$$

for $m = 1, ..., l$ pairwise comparisons.

For equation 2.21, y_m represents the difference in sum of mean ranks for groups j and $j^{'}$

$$y_m = \bar{R}_j - \bar{R}_{j'} \tag{2.22}$$

and σ_m is the standard deviation of y_m

$$\sigma_m = \sqrt{\left(\frac{N(N+1)}{12} - \frac{\sum_{s=1}^{t} \tau_s^3 - \tau_s}{12(N-1)}\right)\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)} \tag{2.23}$$

where t is the number of tied ranks and τ_s is the number of observations tied at the sth tied value. If no ties exist, σ_m becomes

$$\sigma_m = \sqrt{\left(\frac{N(N+1)}{12}\right)\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)} \tag{2.24}$$

The test statistic for each comparison m is approximated by the standard normal distribution such that we have a z statistic calculated as $z_{1-\alpha/2l}$. Therefore, the null hypothesis H_0 is rejected for the following case.

$$|z_m| \geq z_{1-\alpha/2l} \tag{2.25}$$

Again, here we take $\alpha = 0.05$.

2.4.5 Benjamini-Hochberg P-value Correction

As the Wilcoxon rank-sum test, Kruskal Wallis and Dunn test are calculated for each individual probe ID considered in this study, the Benjamini-Hochberg correction for multiple testing was used to control the false discovery rate. The false discovery rate is defined as the proportion of null hypotheses incorrectly rejected (Benjamini & Hochberg, 1995).

Let there be $g = 1, ..., G$ genes and a hypothesis test has been performed on each gene; each gene has been presented a p-value denoted p_g and are ranked in ascending order such that $p_1 \leq p_2 \leq ... \leq p_G$ (Benjamini & Hochberg, 1995). Denote i as the number rank a p-value is assigned. Benjamini & Hochberg (1995) describes the method of calculating adjusted p-values as

$$q_g = \frac{p_g G}{i} \tag{2.26}$$

where q_g is the adjusted p-value for gene g, p_g is the original p-value for the gene g, G is the total number of genes tested and i is the rank of a gene g in ascending order. Hypotheses, denoted H_g, are rejected when

$$q_g < q^* \tag{2.27}$$

where q^* controls the false discovery rate. In this study, $q^* = 0.05$ following Arend et al. (2018) and Buttarelli et al. (2022).

2.4.6 Selection of gene signatures

Objective 1: Detection of EOC

For this study, differentially expressed and dys-regulated probes were identified through the use of fold change analysis and the Wilcoxon rank-sum tests. Fold change analysis and the Wilcoxon rank-sum test was performed simultaneously to compare each of the individual EOC subtypes with healthy ovaries. The probes identified as differentially expressed and dys-regulated here were indicative of presence or non-presence of EOC.

Objective 2: Diagnosis of Individual EOC subtypes

The next step in analysis was determining which of the probes were only differentially expressed and dys-regulated between **one** EOC subtype and normal ovaries. This made a probe unique in its EOC detection; it only indicated the presence or non-presence of **one** EOC subtype.

The probes found to be uniquely indicative of EOC presence or non-presence for one subtype were further filtered. This filtering was completed by use of the Kruskal Wallis test and the post-hoc Dunn test. The Kruskal Wallis test was performed to identify whether the expression of selected probes was also found to differ between any pair of the five EOC subtypes.

We shall let the number of EOC subtypes be K, where $j = 1, ..., K$. The jth subtype will be denoted S_j.

If for a given probe, the Kruskal Wallis test rejected the null hypothesis, the Dunn test was then performed on this probe to determine which of the $j = 1, ..., K$ subtype pairs presented differential expression. Probes found to be differentially expressed between a specific subtype S_j and all other subtypes $S_{j'}$, for $j \neq j^{'}$, were retained for consideration in a 'gene' signature for the corresponding subtype S_j. Probes found to present differential expression between any other subtype pairs were not retained. An example of criteria for a probe to be selected for gene signature consideration in clear cell EOC can be found in table 2.3, below.

Table 2.3: Example of criteria for further probe selection through the Dunn test for a specific EOC subtype (clear cell). A cross indicates non-statistical significance of p-value. A tick indicates statistical significance of p-value. A probe meeting this criteria was retained.

	Clear Cell	Endometrioid	Mucinous	LGSOC	HGSOC
Clear Cell	-	✓	✓	✓	✓
Endometrioid	✓	-	×	×	×
Mucinous	✓	×	-	×	×
LGSOC	✓	×	×	-	×
HGSOC	✓	×	×	×	-

The hypothesis was that the probes meeting the gene signature selection criteria should be able to identify the presence of one EOC subtype whilst simultaneously indicating non-presence of all four other EOC subtypes.

Multiple probe IDs can represent a single gene. If the same gene symbol was considered more than once for a 'gene' signature, the function 'jmap' in 'jetset' R (Li et al., 2011) was used to determine the optimal probe ID to represent the gene. This probe ID was retained; the other probe IDs representing the gene were removed. If there was no optimal probe ID available for that gene, the probe ID holding the smallest adjusted p-value from the Kruskal Wallis test was selected.

2.5 Principal Component Analysis

Let a multivariate data set be denoted as X_{np} where there are $i = 1, ..., n$ cases and $j = 1, ..., p$ variables. The intention of principal component analysis is to reduce the dimensions of the data by finding principal components $m < p$ that retain the maximum amount of variance possible from the original p variables (Everitt & Dunn, 1991b). Each principal component is a linear combination of the original p variables as described by Everitt & Dunn (1991a).

$$y_j = \mathbf{a}_j^{\mathrm{T}}\mathbf{x} \tag{2.28}$$

where y_j is the jth principal component, $\mathbf{a}_j^{\mathrm{T}}$ is a row vector of coefficients corresponding to the column vector of original variables $\mathbf{x}$ and $\mathbf{a}_j^{\mathrm{T}}$ denotes the transpose of $\mathbf{a}_j$. Principal components are restricted by two conditions in which we normalise $\mathbf{a}_j$ (equation 2.29) and ensure principal components

are uncorrelated (equation 2.30).

$$\mathbf{a}_j^{\mathrm{T}}\mathbf{a}_j = 1 \tag{2.29}$$

$$\mathbf{a}_j^{\mathrm{T}}\mathbf{a}_i = 0 \quad i < j \tag{2.30}$$

The first principal component has the greatest variance of all the principal components.

$$Var(y_i) > Var(y_j) \quad i < j$$

This variance is calculated as

$$Var(y_j) = \mathbf{a}_j^{\mathrm{T}}\mathbf{S}\mathbf{a}_j \tag{2.31}$$

where $\mathbf{S}$ is the covariance matrix of the p original variables. The maximisation of variance of the first principal component y_1 is subject the constraint given in equation 2.29. Maximisation of variance of all principal components $y_2, ..., y_p$ are subject to the constraints given in equations 2.29 and 2.30.

We can denote the eigenvalue of $\mathbf{S}$ as λ_j for the jth principal component. Thus, the following holds.

$$Var(y_j) = \lambda_j \tag{2.32}$$

See Chatfield & Collins (1980) for details on the use of Langrange multipliers for maximising variance.

2.6 Classifiers

2.6.1 Choice of Classifier

The first factor considered when choosing classification methods for this study was the non-normality of the data distribution. Non-parametric classifiers were chosen to avoid the violation of assumptions based the distribution of the data. Secondly, due to the data providing class labels and the aim being to determine whether genes presented diagnostic capabilities, supervised classification methods were chosen. Furthermore, as this study only considers binary classification, the ability of classifiers to perform multi-class classification was not necessary.

Support vector machines hold the advantage that they do not place any assumption on the underlying distribution of the data. Furthermore, they are well equipped for binary classification. With gene signatures of varied sizes chosen for this analysis, another advantage of support vector machines is their ability to deal with high-dimensional data. SVMs can perform both linear and non-linear classification through use of kernels. Additionally, no matter

the kernel chosen the optimal separation between classes will be selected (Boateng et al., 2020). However, SVMs also face disadvantages. Boateng et al. (2020) describes disadvantages such as their inability to deal with large training data sample sizes. SVM classifiers can be sensitive to imbalanced data; this sensitivity can be reduced by the introduction of class weights prior to model training. Moreover, the choice of kernel for non-linear classification is not intuitive and calls for trialling of kernels to find the most appropriate one.

The K-nearest neighbours technique also has the advantage of being a non-parametric method, making no assumptions on the underlying distribution of the data. It also has the capabilities to classify non-linear data. KNN classifiers are easy to understand due to the classification being based only on distance measures. However, a large sample size in training the data can cause issues in computational time for KNN as the distance measure has to be calculated between every training sample and the test sample (Boateng et al., 2020).

The third type of classification technique chosen for this analysis was random forest. This model does not hold assumptions on the distribution of the data. Random forest classifiers are implemented due to their ability to reduce overfitting in comparison to single decision trees (Boateng et al., 2020). However, random forest models can be sensitive to changes in the training data (Boateng et al., 2020).

In each case, the optimal parameters needed to train the models can be determined through validation techniques (discussed in section 2.6.6), to improve the performance of the classifiers. This can be computationally expensive dependent on how many parameters are to be optimised and how large that training data set is.

The classifiers mentioned above will be discussed in more detail below.

Suppose there are $i = 1, ..., p$ features and $j = 1, ..., n$ samples. This study will consider features to be genes and samples to be patient samples of each EOC subtype and normal ovaries. An observation of sample j for feature i is denoted x_{ij}. This study considers an observation to be the expression of gene i in patient sample j. For each sample $\mathbf{x}_j$, there is a corresponding binary outcome denoted y_j. Each sample $\mathbf{x}_j$ can be represented as a point in p-dimensional space.

2.6.2 K-Nearest Neighbours (KNN)

K-nearest neighbours, described by Cover & Hart (1967), is a method that classifies a test sample $\mathbf{x}^{'}$ such that the class labels of the k closest training samples $\mathbf{x}_j$ influence the overall class label of the test sample. The Euclidean

distance, denoted d, between the test sample $\mathbf{x}^{'}$ and each individual training sample $\mathbf{x}_j$ is calculated

$$d(\mathbf{x}^{'}, \mathbf{x}_j) = \sqrt{\sum_{i=1}^{p}(x_i^{'} - x_{ij})^2} \tag{2.33}$$

for $j = 1, ..., n$. The KNN model then determines the k training samples $\mathbf{x}_j$ with $min(d(\mathbf{x}^{'}, \mathbf{x}_j))$. The predicted class label $y^{'}$ of test sample $\mathbf{x}^{'}$ is then determined by majority vote with respect to the predicted classes y_j of the k training samples $\mathbf{x}_j$.

2.6.3 Linear Support Vector Machines (Linear SVM)

Linear support vector machines aim to construct a linear hyperplane in p-dimensional space that distinctly separates data points based on their class label (Awad & Khanna, 2015).

For this case, suppose we have n samples, $\mathbf{x}_j$, where $j = 1, ..., n$ that are represented in p-dimensional space ($i = 1, ..., p$ features). Each sample x_j corresponds to a class y_j and $y_j \in \{-1, 1\}$.

When using a linear support vector machine, the data can either be completely linearly separable or not completely linearly separable.

The optimal hyperplane for a completely linear separable case, described by James et al. (2013) and Awad & Khanna (2015), follows.

The objective of linear SVM for a completely separable case is to find a hyperplane

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{2.34}$$

$$subject \quad to \quad the \quad constraints$$

$$y_j(\mathbf{w}^T\mathbf{x}_j + b) \geq 0 \quad \forall j \tag{2.35}$$

where $\mathbf{w}$ is normal to the hyperplane and b is a bias term that means the hyperplane does not go through the origin for $b \neq 0$. The constraint indicates that all points $\mathbf{x}_j$ are classified correctly; $f(\mathbf{x}_j) > 0$ for $y_j = 1$ and conversely, $f(\mathbf{x}_j) < 0$ for $y_j = -1$.

The optimal hyperplane is described by the following equation

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{2.36}$$

Suppose we have the following equations to represent the boundaries of a margin m.

$$\mathbf{w}^T\mathbf{x} + b = 1 \tag{2.37}$$

$$\mathbf{w}^T\mathbf{x} + b = -1 \tag{2.38}$$

In addition, we have that a distance, denoted d, from the hyperplane to any $\mathbf{x}$ is defined as

$$d = \frac{|\mathbf{w}^T\mathbf{x} + b|}{||\mathbf{w}||_2} \tag{2.39}$$

where $||\mathbf{w}||_2 = \sqrt{\mathbf{w}^T\mathbf{w}}$ (Euclidean norm).

The linear SVM aims to find a margin that maximises the minimum distance between a point $\mathbf{x}$ and the hyperplane. As we have the margin being bounded by equations 2.37 and 2.38, we find that the margin width m is calculated by

$$\frac{|-1|}{||\mathbf{w}||_2} + \frac{|1|}{||\mathbf{w}||_2} = \frac{2}{||\mathbf{w}||_2} \tag{2.40}$$

This classifier aims to separate the two classes by as much distance as possible. Therefore, the margin width is maximised to maintain as large a distance between the points.

$$maximise_{\mathbf{w},b} \quad \frac{2}{||\mathbf{w}||_2} \tag{2.41}$$

$$subject \quad to \quad the \quad constraints \tag{2.42}$$

$$y_j(\mathbf{w}^T\mathbf{x}_j + b) \geq 0 \quad \forall j \tag{2.43}$$

which is reformulated as

$$minimise_{\mathbf{w},b} \quad \frac{||\mathbf{w}||_2^2}{2} \tag{2.44}$$

$$subject \quad to \quad the \quad constraints$$

$$y_j(\mathbf{w}^T\mathbf{x}_j + b) \geq 1 \quad \forall j \tag{2.45}$$

Hence, all $\mathbf{x}_j$ with corresponding class $y_j = 1$ can be linearly separated from all $\mathbf{x}_j$ with corresponding class $y_j = -1$ due to each class having a distance of at least $m/2$ from the hyperplane and $f(\mathbf{x}_j) > 1$ for $y_j = 1$ and conversely, $f(\mathbf{x}_j) < -1$ for $y_j = -1$.

In the case where data is not completely linearly separable, data points may violate the constraints in the optimization problem equations 2.44 and 2.45 and thus the constraints for a completely separable case must be adapted.

The minimisation problem for non-separable data, discussed by Awad &

Khanna (2015) and Boyle (2011), becomes

$$minimise_{\mathbf{w},b,\epsilon} \quad \frac{||\mathbf{w}||_2^2}{2} + C\sum_{j=1}^{n}\epsilon_j \tag{2.46}$$

subject to the constraints

$$y_j(\mathbf{w}^T\mathbf{x}_j + b) \geq 1 - \epsilon_j \quad \forall j \tag{2.47}$$

$$\epsilon_j \geq 0 \quad \forall j \tag{2.48}$$

Here, ϵ_j are slack variables; slack variables allow a point $\mathbf{x}_j$ to move inside of the margin boundaries set in the separable case.

For $\epsilon_j = 0$, the corresponding sample $\mathbf{x}_j$ does not violate the margin boundary. An $\epsilon_j > 0$ indicates a violation of the margin boundary and $\epsilon_j > 1$ indicates both a violation of the margin boundary and of the hyperplane. Hence, a sample $\mathbf{x}_j$ with a corresponding $\epsilon_j > 1$ is misclassified (James et al., 2013). Any sample $\mathbf{x}_j$ with corresponding $\epsilon_j > 0$ are defined as support vectors.

Awad & Khanna (2015) provides a description of the parameter C such that an increase in this value decreases the width of the margin and therefore minimises misclassification. Whereas, a decrease in C increases the width of the margin and allows for more violations; in this case more emphasis is placed on the maximising of the margin and less on the minimising of misclassifications.

2.6.4 Non-Linear Support Vector Machines (Non-linear SVM)

In the case that the data cannot be linearly separated, non-linear kernels are used to map the original data points into a higher dimensional space ($\Phi(\mathbf{x})$) so that the soft margin linear classifier can be performed.

Awad & Khanna (2015) defines the optimisation problem for a non-linear support vector machine as

$$minimise_{\mathbf{w},\epsilon} \quad \frac{||\mathbf{w}||_2^2}{2} + C\sum_{j=1}^{n}\epsilon_j \tag{2.49}$$

subject to the constraints

$$y_j(\mathbf{w}^T\varphi(\mathbf{x}_j) + b) \geq 1 - \epsilon_j \tag{2.50}$$

$$\epsilon_j \geq 0 \quad \forall j \tag{2.51}$$

where $\varphi(\mathbf{x}_j)$ is taken from $K(\mathbf{x}_j, \mathbf{x}') = \varphi(\mathbf{x}_j)^T\varphi(\mathbf{x}')$

Radial Basis Function Kernel (Radial SVM)

The radial basis function kernel is of the form (Cortes & Vapnik, 1995)

$$K(\mathbf{x}, \mathbf{x}^{'}) = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}^{'}||_2^2}{\sigma^2}\right) \tag{2.52}$$

The denominator can also be evaluated as $2\sigma^2$ or equivalently $\gamma = \frac{1}{2\sigma^2}$ (Prajapati & Patle, 2010; Ghosh et al., 2019; Vapnik, 1998). The value of σ is said to influence the level of non-linearity in the model (Mastropietro et al., 2023). A smaller value of σ increases the curvature of the decision boundary (Ben-Hur et al., 2008).

Polynomial Kernel (Polynomial SVM)

The polynomial kernel is of the form (Cortes & Vapnik, 1995)

$$K(\mathbf{x}, \mathbf{x}^{'}) = (\mathbf{x}^T\mathbf{x}^{'} + c)^d \tag{2.53}$$

for a polynomial of degree d. For a greater value of d, the curvature of the decision boundary increases (Ben-Hur et al., 2008).

2.6.5 Random forest

The random forest algorithm, introduced and described by Breiman (2001), creates B, $(b = 1, ..., B)$, decision trees by use of bagging. Bagging produces each decision tree using a subset of random samples from the original training set with replacement (Breiman, 2001).

For some decision tree T_b, bagging produces a new training set X_b containing $j = 1, ..., n^{'}$ samples based on $i = 1, ..., p$ features, denoted $\mathbf{x}_j$. The set X_b has corresponding class labels $\mathbf{y}_b$. The samples not selected in the bagging process for a decision tree T_b will be denoted $\mathbf{x}_b^{'}$. The samples $\mathbf{x}_b^{'}$ not included in the tree T_b can then be used to test the classification strength of the random forest model. Each tree T_b will predict the class label of each sample $\mathbf{x}_b^{'}$. Following this, the overall majority class prediction for each $\mathbf{x}_b^{'}$ based on all decision tree votes will be the final class prediction of this sample. The proportion of incorrectly classified $\mathbf{x}_b^{'}$ out of total number of $\mathbf{x}_b^{'}$ is defined as the out of bag error. A test sample $\mathbf{x}_t$ can then be classified by the majority vote subject to all of the decision trees in the random forest.

For determining the decision node splits during model training, random feature selection, discussed by Breiman (2001), is implemented. A random set of $m < p$ features are evaluated at each decision node with the feature

providing the optimal split for this node chosen. The use of this method prevents the correlation of the B decision trees. Commonly, the value of m is chosen as $m = \sqrt{p}$. All random forest models for this study were made from 500 trees.

2.6.6 Stratified K-fold cross validation

K-fold cross validation is a common method used to tune parameters in classification models.

Described in detail by Geisser (1975), the training data $\mathbf{x}_j$, for all $j = 1, ..., n$ samples, is split into K equal (or relatively equal) sized folds based on the total number of samples in the training data. A classifier is then trained for varying parameter values on $K - 1$ of these folds and validated on the Kth fold. This classifier is trained K times such that each fold is used once as a validation set; an evaluation metric $eval_k$, for $k = 1, ..., K$, is calculated for each of the k validation sets.

The overall evaluation metric for a parameter is then calculated as

$$\frac{1}{K}\sum_{k=1}^{K} eval_k \tag{2.54}$$

If the evaluation metric focuses on similarity between predicted and known sample classes, the optimal parameter value will be the one that maximises the overall evaluation metric. Whereas, if the evaluation metric focuses on differences between predicted and known sample classes then the optimal parameter value will minimise this overall evaluation metric.

K cross validation can produce a large variation in results, repeat k-fold cross validation is implemented (Kim, 2009). If k-fold cross validation is repeated M times, where $m = 1, ..., M$, the process of k-fold cross validation described above is repeated M times, where for each m the original data is randomly sampled into K folds differing to those used in other repeats. The overall evaluation metric for a given parameter is then the mean evaluation metric for the K folds across the M repeats.

$$\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} eval_{mk} \tag{2.55}$$

where $eval_{mk}$ is the evaluation metric for the validation fold k in repeat m.

Stratified cross validation is employed in order to maintain a representation of each class in each of the K folds and is recommended for use when data contains imbalanced class proportions (Gunasegaran & Cheah, 2017).

Due to the imbalance of sample sizes in the data used in this study, stratified 10-fold cross validation with 10 repeats was used to select the optimal parameters for each classification model (Berrar, 2019).

Furthermore, due to class imbalance, intuitively sensible weights of the inverse of a classes prevalence in the training data set divided by total number of classes were applied to a class in order to balance the importance of correctly classifying positive and negative classes (Fernández et al., 2013).

2.7 Evaluation Metrics

2.7.1 Confusion Matrix

A common representation of the results of a classification model is a confusion matrix. This matrix allows the comparison of the actual class labels in the original data and the predicted class labels based on the classification model (Fawcett, 2006).

	Actual Class	
Predicted	0 (Positive)	1 (Negative)
0 (Positive)	TP	FP
1 (Negative)	FN	TN

Table 2.4: Confusion Matrix for Binary Classification. TP=true positive, TN=true negative, FN=false negative and FP=false positive. Adapted from Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters, 27*, 861-874.

The confusion matrix represented by table 2.4 is based on binary classification; there are two classes a data sample can be labelled as. These two classes are commonly defined as positive and negative.

A *true negative* result implies that a sample is both predicted and known to be of negative class. Whereas, a *true positive* sample is both predicted and known to be of the positive class.

A *false positive* result, which is also referred to as a type I error, implies that a sample is known to be of the negative class but is classified as positive. A *false negative*, or a type II error, is known to be of the positive class but is predicted to be negative (Han et al., 2011).

From the confusion matrix, it is possible to calculate a variety of evaluation metrics to determine how well the corresponding classification model works.

The positive classes for this study will be defined as the EOC subtype associated with the specific gene signature used to train the classification models.

Metrics for data with unbalanced classes will first be described in section 2.7.2. This will be followed by metrics for balanced classes in section 2.7.4. As the data in this study contains unbalanced classes, the metrics chosen for use to evaluate classification models are found in section 2.7.2.

2.7.2 Unbalanced Evaluation Metrics

Sensitivity

The sensitivity of a model, also known as the recall, is defined as the proportion of known positive samples that are correctly predicted as positive (Altman & Bland, 1994b).

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.56}$$

This is commonly used as an evaluation metric for imbalanced data with a small number of positives (He & Garcia, 2009). This metric will be used in this study due to imbalances classes for both training and testing data.

Specificity

Specificity is defined as the proportion of known negative samples that are correctly predicted to be negative (Altman & Bland, 1994b).

$$Specificity = \frac{TN}{TN + FP} \tag{2.57}$$

Prevalence

Prevalence is defined as the proportion of known positive samples out of the total samples (Altman & Bland, 1994a).

$$Prevalence = \frac{TP + FN}{TP + FN + FP + TN} \tag{2.58}$$

Positive predictive value (PPV)

The PPV of a model, also known as the precision, determines the proportion of all positively predicted samples that are correctly predicted as positive.

PPV can be calculated by equation 2.59 (Altman & Bland, 1994a).

$$PPV = \frac{sensitivity \times prevalence}{(sensitivity \times prevalence) + (1 - specificity)(1 - prevalence)} \tag{2.59}$$

Using equations 2.56, 2.57 and 2.58, PPV can be calculated in terms of true positives and false positives as shown in equation 2.60.

$$PPV(Precision) = \frac{TP}{TP + FP} \tag{2.60}$$

This metric is commonly used for evaluation of classification models in imbalanced data with a small number of positives (He & Garcia, 2009). This metric will be used in this study due to the imbalance of classes for both training and test data.

$\mathbf{F}_\beta$ score

The F_β score is a way of combining both precision and recall into one metric. It has been defined by the equation (Chinchor & Sundheim, 1993)

$$F_\beta = \frac{(\beta^2 + 1) \times precision \times recall}{(\beta^2 \times precision) + recall} \tag{2.61}$$

If precision (PPV) and recall (sensitivity) are considered to be equally as important then $\beta = 1$ and the formula F_1 is described as the harmonic mean of precision and recall. This is calculated by equation 2.62 (Chinchor & Sundheim, 1993).

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{2.62}$$

Using equations 2.60 and 2.56, this can be re-formulated in terms of true positives, false positives and false negatives by equation 2.63.

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{2.63}$$

The F_1 score is a metric often used in the case of imbalanced data with smaller number of positives (He & Garcia, 2009). The closer this value is to 1, the better the classification performance achieved. This metric will be used to determine classifier performance as this study uses data containing class imbalances.

Jaccard Index

The Jaccard Index is a metric commonly used to determine the similarity between two sets. Developed in 1901 by Jaccard (Jaccard, 1901), the Jaccard index is generally represented by equation 2.64

$$J(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} \tag{2.64}$$

where y and $\hat{y}$ are two sets, $\cap$ represents the intersection of sets and $\cup$ represents the union of sets.

When considering the Jaccard index in the context of binary classification, the metric is calculated with respect to the positive class. This metric determines the ratio of the intersection and union of positive actual class labels and those predicted as positive. This is represented by equation 2.65 and can be defined by confusion matrix notation equation 2.66 (Shattuck et al., 2009; Berman et al., 2018)

$$J_p(y, \hat{y}) = \frac{y_p \cap \hat{y_p}}{y_p \cup \hat{y_p}} \tag{2.65}$$

$$J_p = \frac{TP}{FP + FN + TP} \tag{2.66}$$

where y_p are the actual positive class, $\hat{y_p}$ are the predicted positive class. The closer the Jaccard index is to 1, the more similar the actual class labels and predicted class labels are. This metric is also used to determine class performance due to class imbalances in the training and test data.

Fowlkes Mallows Index

Introduced by Fowlkes & Mallows (1983) as a measure of finding the similarity between clusters, the Fowlkes Mallows index has been adapted for use in classification. It is also described as the geometric mean of precision and recall.

$$FMI = \sqrt{\left(\frac{TP}{TP + FP}\right)\left(\frac{TP}{TP + FN}\right)} \tag{2.67}$$

It provides a measure for the similarity between predicted and actual class labels; a value closer to 1 indicates greater similarity between predicted and actual class labels. This study employs this metric to establish classification performance due to the imbalance in classes used.

2.7.3 Precision-Recall Curve

A precision-recall curve considers the proportion of known positive samples that are predicted correctly versus the proportion of predicted positive samples that predicted correctly (TPR versus PPV) (Davis & Goadrich, 2006). Figure 2.3 provides a visual representation of precision-recall curve space.

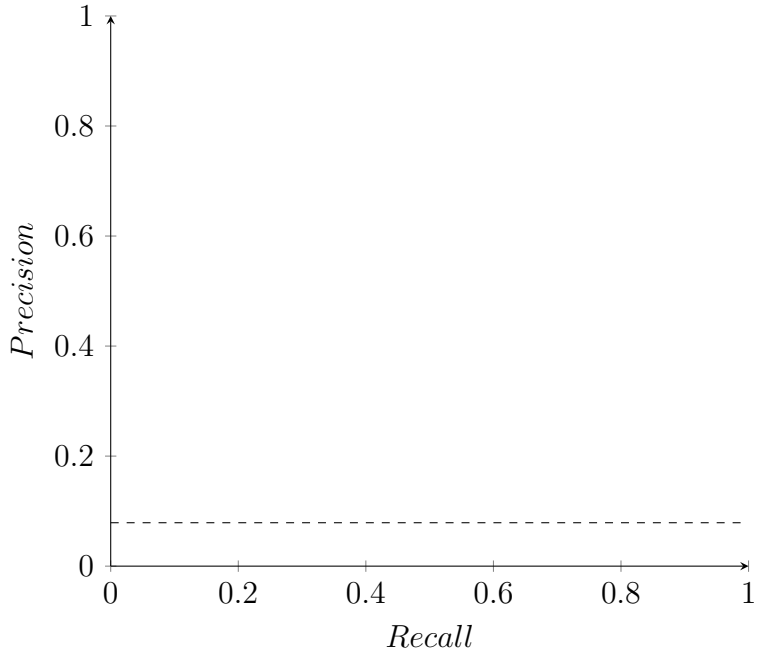


Figure 2.3: Precision-recall curve graph. Horizontal dotted line represents a random classifier's performance.

In the case of precision-recall curves, a horizontal line represents a random classifier model, this has been described as a baseline, and it is represented by the prevalence of positive known samples in the data (Saito & Rehmsmeier, 2015). A classifier exceeding this baseline performs better than a random classifier. For a precision-recall curve, a perfect classifier will reach the coordinate (1,1), where both precision and recall are 100%; implying that all predictions are correct (Sofaer et al., 2019).

In a similar way to the ROC curve (see section 2.7.5), the PR-curve compares the probability or score of a sample being predicted as a member of the positive class t with varying threshold value t^* such that for $t > t^*$, a sample is predicted as the positive class. For each threshold value t^*, a corresponding precision and recall value can be calculated for a classifier.

PR-curves are reported to be more reliable in evaluating classification

models that are based on imbalanced class proportions (He & Garcia, 2009; Saito & Rehmsmeier, 2015). As the data in this study in each classification case uses imbalanced class proportions, the precision-recall curve is used to determine whether classification model performance can be improved based on thresholds.

2.7.4 Balanced Evaluation Metrics

Accuracy

A metric often used when class proportions are balanced, the accuracy of a model determines the proportion of all samples that are correctly predicted to be the same class as their known class (He & Garcia, 2009).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.68}$$

Previously, it has been discussed that when used to evaluate classifiers with imbalanced class, accuracy becomes unreliable due to the lack of effect the smaller class has on the overall accuracy of a model (Sun et al., 2009; García & Herrera, 2008; Hossin et al., 2011; He & Garcia, 2009).

Negative predictive value (NPV)

The NPV of a model determines the proportion of predicted negative samples that are correctly predicted as negative. It is calculated by the formula 2.69 (Altman & Bland, 1994a).

$$NPV = \frac{specificity(1 - prevalence)}{prevalence(1 - sensitivity) + specificity(1 - prevalence)} \tag{2.69}$$

Using equations 2.57, 2.58 and 2.56, NPV can also be calculated in terms of true negatives and false negatives. This is shown by equation 2.70.

$$NPV = \frac{TN}{TN + FN} \tag{2.70}$$

2.7.5 Receiver Operating Curves

Receiver operating curves (ROC), described in Fawcett (2006), represent true positive and false positive rates, which can also be defined as *sensitivity* and $1 - specificity$, respectively, for varying thresholds indicating a samples chance of being predicted as part of the positive class.

Each sample is provided a probability or score t of being classified in the positive and negative class. To determine the class prediction of a sample, a threshold value t^* is chosen such that for $t > t^*$, a sample is classified as positive; a standard threshold for positive classification is $t > 0.5$.

A ROC-curve is produced by varying this threshold t^*, in order to determine a true positive rate (TPR) and corresponding false positive rate (FPR) for each t^*. Figure 2.4 is a visual representation of a ROC-curve space.

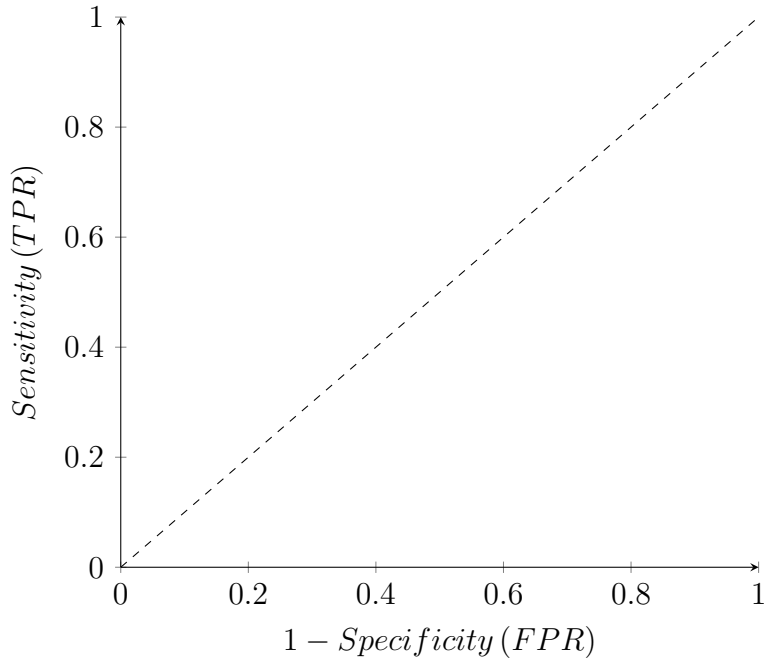


Figure 2.4: ROC curve graph. Dotted line of equation $y = x$ represents a random classifier's performance. Adapted from Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters, 27*, 861-874.

Fawcett (2006) provides insight into the various areas of a ROC-curve plot and their meaning in regards to the classification model. A classifier whose ROC curve follows the dotted line of equation $y = x$ shown in figure 2.4 implies that said classifier performs no better than a randomly chosen classifier. The coordinate (0,1) is representative of perfect classification with 100% TPR and 0% FPR. The coordinate (0,0) indicates that the classifier never supplies a positive prediction with both a TPR and FPR of 0%. Conversely, a coordinate of (1,1) indicates the classifier never predicting the negative class with both a TPR and FPR of 100%.

Fawcett (2006) also provides details on the use of the area under the ROC curve (ROC-AUC) and describes the AUC as the probability that ran-

domly selected positive class samples will rank higher than randomly selected negative class samples. For an AUC of 0.5, the chance of positive class samples ranking higher than negative class samples is 50% and shows no ability to discriminate between sample classes; any $AUC \leq 0.5$ indicates the bad classification performance. Whereas, $AUC > 0.5$ indicates classification performances such that there is an increased chance of positive class samples being ranked higher than negative samples. An $AUC = 1$ represents the perfect classifier in which 100% of randomly selected positive samples are ranked higher than randomly selected negative samples.

Similarly to accuracy, false negatives have little effect on the result of a ROC curve and its area under the curve which can lead to unreliable results (He & Garcia, 2009; Saito & Rehmsmeier, 2015).

2.8 R statistical software

All analysis within this study was performed in R (R Core Team, 2022). The predictive classification models trained and tested were produced using the 'caret' package (Kuhn et al., 2016). The radial basis function kernel presented in the 'caret' package uses 'kernlab' (Karatzoglou et al., 2004) in which the hyperparameter tuned is defined as σ. However, this σ is equated to γ unlike the relation between the two parameters described in section 2.6.4. To avoid confusion, when discussing the radial basis function kernel for models trained and tested in the results chapter (3), the hyperparameter will be denoted γ.

For the production of PR-curves for classification models, the function 'evalm' in package 'MLeval' was utilised (John, 2020).

Chapter 3

Results

3.1 Determining Statistically Significant Genes.

Due to the non-normality of the data analysed, non-parametric hypothesis tests were used to determine statistically significant probes. A total of 45,118 probe IDs were tested for their significance in difference of expression level between normal ovary samples and five subtypes of EOC (clear cell, endometrioid, mucinous, low grade serous and high grade serous) using both the Wilcoxon rank-sum test and fold change analysis simultaneously. Probes were considered to be statistically significant with a Benjamini-Hochberg adjusted p-value of less than 0.05 and a fold change value of either greater than 2 or less than 0.5.

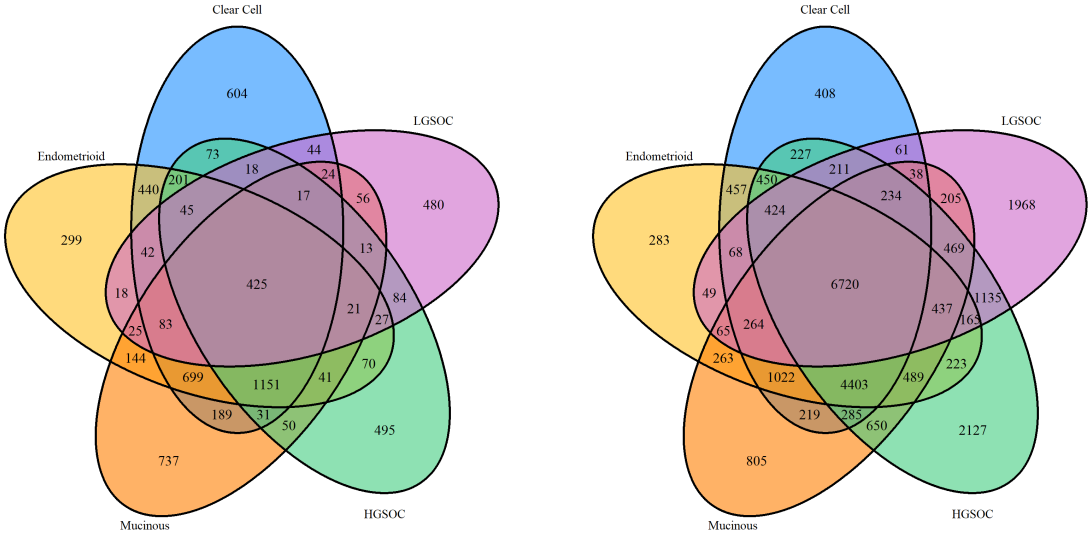
Table 3.1: Table presenting the number of up-regulated, down-regulated and total dys-regulated, differentially expressed probes when comparing a subtype of EOC with normal ovaries.

Comparison	Up-regulated	Down-regulated	Total Dys-regulated
Clear Cell vs Control	4086	15491	19577
Endometrioid vs Control	3731	15782	19513
Mucinous vs Control	3706	16568	20274
Low Grade Serous vs Control	2762	18649	21411
High Grade Serous vs Control	1422	12513	13935

Table 3.1 displays the number of probes found in each EOC subtype that were significantly differentially expressed and dys-regulated when compared with normal ovary samples. These probes were described as being indicative of cancer presence. In total, LGSOC had a greater number of dys-regulated, differentially expressed probe IDs when comparing their expression level with

normal ovaries. HGSOC had the least. However, clear cell EOC presented the greatest number of up-regulated, differentially expressed genes when compared with normal ovaries. It is clear that genes found to be down-regulated in ovarian cancer subtypes are much more common than those up-regulated.

The dys-regulated, differentially expressed probes were then compared across EOC subtype groups to find the probes that were uniquely differentially expressed between the normal ovary samples and an individual EOC subtype.



(a) Five-way Venn Diagram For Statistically Significant, Up-regulated Probe IDs.

(b) Five-way Venn Diagram For Statistically Significant, Down-regulated Probe IDs.

Figure 3.1: Venn Diagrams showing number of shared and unique dys-regulated probe IDs for EOC subtypes.

Figure 3.1 provides insight into the number of dys-regulated, differentially expressed probes both shared between EOC subtypes and unique to each EOC subtype for up-regulated (3.1a) and down-regulated (3.1b) probes. For both up- and down-regulated probes, endometrioid EOC had the smallest number of unique probes (299 and 283, respectively). The EOC subtype with the largest number of unique probes was mucinous when considering up-regulated genes (737). In contrast, HGSOC had the greatest number of unique down-regulated probes (2127). Figure 3.1a also shows that endometrioid and clear cell had the greatest number of shared probes (440) in comparison to endometrioid and LGSOC sharing the least (18). Similarly, based on figure 3.1b, endometrioid and LGSOC also shared the smallest

number of down-regulated probes (49). However the two subtypes sharing the greatest number of down-regulated probe IDs were HGSOC and LGSOC (1135).

The probes considered for further hypothesis testing were those found to be

1. Uniquely significantly up-regulated and differentially expressed for an EOC subtype.

2. Uniquely signficantly down-regulated and differentially expressed for an EOC subtype.

Therefore, a probe may be uniquely up-regulated in one EOC subtype but found to be down-regulated between normal ovaries and other subtypes and vice versa.

Application of the Dunn test to each dys-regulated probe unique to a subtype, S, then allowed the probes to be filtered further. Any probe considered to be statistically significant when comparing distributions of expression level between only subtype S and each of the remaining subtypes was retained for further analysis.

Table 3.2: Table presenting the number of probe IDs uniquely up-regulated or down-regulated and differentially expressed between a subtype S and normal ovaries. Furthermore, the table presents the number of probe IDs retained due to the Dunn test indicating differential expression between only subtype S and all other subtypes.

			Dunn test significance between S and all other subtypes	
EOC subtype S	Up-regulated Probes	Down-regulated Probes	Up-regulated Probes	Down-regulated Probes
Clear Cell	604	408	26	14
Endometrioid	299	283	9	3
Mucinous	737	805	24	46
Low Grade Serous	480	1968	95	43
High Grade Serous	495	2127	147	632

Table 3.2 indicates that although there were a large number of differentially expressed, dys-regulated probes between each subtype and normal ovaries, the number of these probes that are also differentially expressed between the given subtype and all other subtypes is small. Endometioid EOC had the smallest number of probes retained once filtering had taken place, whereas HGSOC had the most. These probes were considered for selection to produce a gene (probe) signature for a given subtype of EOC in order to differentiate said subtype from all others.

Uniqueness of probes were based on probe ID due to each probe ID being unique. Once probes were considered for selection in the production of

a signature for classification, corresponding gene symbols were introduced to ensure multiple probe IDs for the same gene were not selected for gene signatures.

3.2 Classifying Patients with Clear Cell EOC

3.2.1 Gene Signature

Fifteen of the forty genes (probes) found to be differentially expressed between clear cell and non-clear cell samples were used in the gene signature for classifying samples into the two groups. These genes are provided in table 3.3, below.

Table 3.3: Gene signature for classifying clear cell vs non-clear cell samples.

Probe ID	Gene Symbol	Gene Name
208868_s_at	GABARAPL1	GABA type A receptor associated protein like 1
228739_at	CYS1	cystin 1
207052_at	HAVCR1	hepatitis A virus cellular receptor 1
220324_at	LINC00472	long intergenic non-protein coding RNA 472
218704_at	RNF43	ring finger protein 43
225597_at	SLC45A4	solute carrier family 45 member 4
200697_at	HK1	hexokinase 1
220786_s_at	SLC38A4	solute carrier family 38 member 4
209606_at	CYTIP	cytohesin 1 interacting protein
241455_at	C6orf132	chromosome 6 open reading frame 132
225545_at	LOC101930123///EEF2K	eukaryotic elongated factor 2 kinase///eukaryotic elongation factor 2 kinase
217917_s_at	DYNLRB1	dynein light chain roadblock-type 1
217796_s_at	NPLOC4	NPL4 homolog, ubiquitin recognition factor
1563209_a_at	MACROD2	MACRO domain containing 2
216238_s_at	FGB	fibrinogen beta chain

3.2.2 Principal Component Analysis

Principal component analysis was used as a form of exploratory data analysis. Due to high dimensionality of the data, principal component analysis was used in order to visualise the training data samples in two-dimensions. These two dimensions were produced using linear combinations of the 15 genes selected for the gene signature.

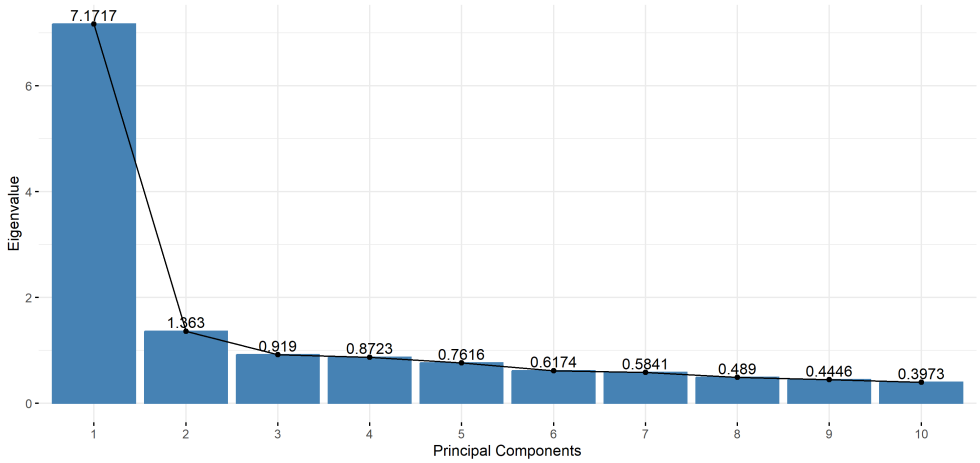


Figure 3.2: Screeplot of the first ten principal components for principal component analysis of clear cell EOC gene signature.

Figure 3.2 shows that the first two principal components have eigenvalues greater than 1 which suggests that they provide greater variance than any of the original 15 genes. Retaining 95% of the variance from the original 15 genes required 12 principal components. Therefore, the use of principal component analysis to reduce the dimensions of the data to two would not be suitable in a diagnostic context. For this reason, principal component analysis was not used to pre-process the data prior to classification model training.

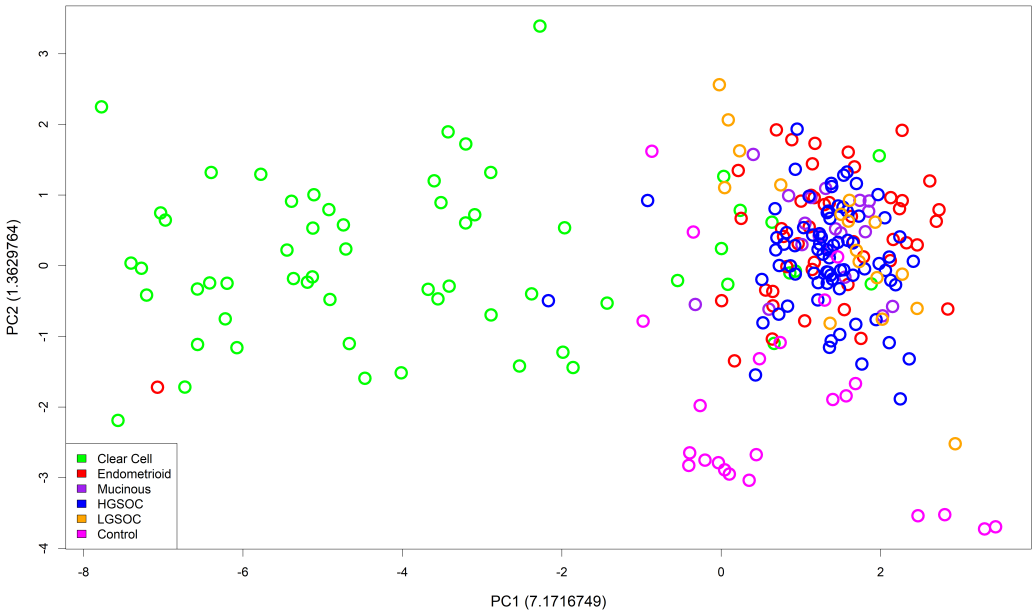


Figure 3.3: Two dimensional principal component plot of samples using clear cell gene signature.

Figure 3.3 provides a visualisation of both clear cell and non-clear cell EOC samples when projected within the first two principal components. It indicated that the linear combination of the 15 genes representing the first principal component was able to separate a large proportion of clear cell samples from non-clear cell samples. This supports the hypothesis that the gene signature suggested has potential in distinguishing between clear cell and non-clear cell EOC samples. Only two non-clear cell samples were found within the large cluster of clear cell samples; these were of endometrioid and HGSOC histology. All other non-clear cell samples were clustered on the right hand side of the principal component plot. Although the two groups of samples could not be perfectly separated, it was clear that the two groups could be segregated well based on principal components calculated from the suggested gene signature.

3.2.3 Classification Model Training and Testing

Clear cell EOC was designated as the positive class.

Linear SVM

The optimal cost parameter value for maximising the mean precision-recall curve AUC was $C = 0.01$. The optimal mean AUC based on this parameter was 0.74375. This model used a total of 87 support vectors; 21 of which were clear cell samples and 66 were non-clear cell samples. The trained data has a total of 10 false negative and 3 false positive classifications.

Radial SVM

The optimal parameters for radial SVM were $C = 0.25$ and $\gamma = 0.06$. These parameters had the greatest mean precision-recall AUC score of 0.75426. This model used 112 support vectors; 34 clear cell samples and 78 were non-clear cell samples. This model had 4 false positives and 9 false negatives.

Polynomial SVM

The optimal parameters for achieving the maximum mean precision-recall AUC of 0.74525 were $C = 0.5$, $scale = 0.01$ and $degree = 2$. The number of support vectors used in this model was 88; 22 were clear cell samples and 66 were non-clear cell samples. The model had 3 false positive and 10 false negative classifications.

KNN

Eleven nearest neighbours was selected as the optimal parameter value for k when using Euclidean distance. This gave the maximum mean precision-recall AUC of 0.32068. This training model had 2 false positives and 14 false negatives.

Random Forest

The optimal random forest model tested one random gene at each split. This model gave a maximum mean precision-recall AUC of 0.73761. This model has 3 false positives and 12 false negatives.

Table 3.4: Table providing comparisons of Jaccard index (JI), Fowlkes Mallows index (FMI), precision, recall, specificity and F_1 score for the clear cell training classification models.

Classifier	JI	FMI	Precision	Recall	Specificity	F_1 Score
Linear SVM	0.79032	0.88464	0.94231	0.83051	0.98333	0.88288
Radial SVM	**0.79365**	**0.88582**	0.92593	**0.84746**	0.97778	**0.88496**
Polynomial SVM	0.79032	0.88464	0.94231	0.83051	0.98333	0.88288
KNN	0.73770	0.85455	**0.95745**	0.76271	**0.98889**	0.84906
RF	0.75806	0.86534	0.94000	0.79661	0.98333	0.86239

Table 3.5: Table providing comparisons of Jaccard index (JI), Fowlkes Mallows index (FMI), precision, recall, specificity, F_1 score and PR-AUC for clear cell test predictions.

Classifier	JI	FMI	Precision	Recall	Specificity	F_1 Score	PR-AUC
Linear SVM	0.71429	0.83333	0.83333	0.83333	0.98795	0.83333	0.69
Radial SVM	0.62500	0.77152	0.71429	0.83333	0.97590	0.76923	0.72
Polynomial SVM	0.71429	0.83333	0.83333	0.83333	0.98795	0.83333	0.69
KNN	0.71429	0.83333	0.83333	0.83333	0.98795	0.83333	0.41
RF	**0.83333**	**0.91287**	**1.00000**	0.83333	**1.00000**	**0.90909**	0.74

Tables 3.4 and 3.5 provide the evaluation metrics for each trained and tested model for classifying clear cell and non-clear cell samples.

None of the training models were able to provide precision or recall scores of 1 indicating both false positive and false negative classifications in all models. Radial SVM had the greatest success in classifying clear cell and non-clear cell samples according to the F_1 metric.

All five training models misclassified the same 11 samples. Nine of these were clear cell samples (4, 13, 21, 29, 30, 47, 48, 54, 59) and two were non-clear cell samples (69 and 157). The radial SVM model misclassified one non-clear cell sample (85) that no other models misclassified. The random forest model misclassified a unique clear cell sample (57). Furthermore, the KNN model misclassified three clear cell samples that no other model misclassified (25, 42 and 55). The most commonly misclassified non-clear cell samples were of endometrioid and HGSOC histology.

The random forest model provided the greatest test classification performance according to the F_1 metric. This model made one false negative classification of clear cell sample 4. However, an improvement in test classification performance when compared with training classification performance is unusual. It is noted here that the training and test data were drawn from

separate data sets.

Linear SVM, polynomial SVM and KNN models all provided models with equal predictive capabilities at a threshold of $t > 0.5$; precision and recall for these models indicated both false positive and negative misclassifications. In each of these models, one false positive and one false negative classification was made for clear cell sample 4 and non-clear cell sample 46.

The radial SVM model performed the worst when classifying the test samples; precision and recall values indicate a greater frequency of false positive misclassifications than false negative misclassifications. This leads to the model having the lowest Jaccard index and Fowlkes Mallows index. The test model made one false negative misclassification of clear cell sample 4 and two false positive misclassifications for non-clear cell samples 46 and 63.

The most common histology of non-clear cell samples misclassified was HGSOC; no samples of any other histology were misclassified in the test models.

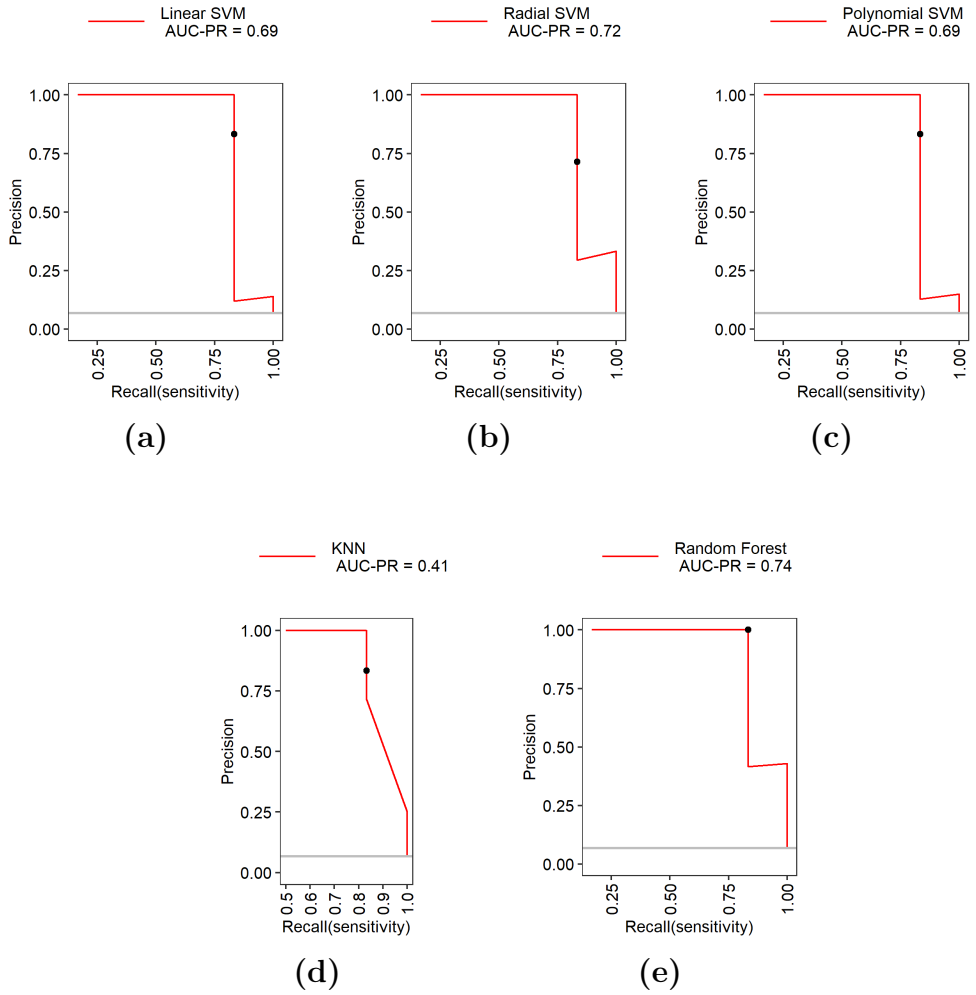


Figure 3.4: **(a)** Linear SVM precision-recall curve. **(b)** Radial SVM precision-recall curve. **(c)** Polynomial SVM precision-recall curve. **(d)** KNN precision-recall curve. **(e)** Random Forest precision-recall curve. The precision and recall of classifiers at a threshold of $t > 0.5$ are represented by black points.

Figure 3.4 represents the precision-recall curves for each of the five models tested for distinguishing between clear cell and non-clear cell samples.

Figures 3.4a and 3.4c present the precision-recall curves for linear and polynomial SVM models, respectively. The precision-recall curves show that it was possible, if considering a threshold $t^* \neq 0.5$, to classify these samples with a precision of 1 for a recall of 0.83333. For the linear SVM, a threshold

of $0.571 < t^* < 0.617$ provided this improved classification. For polynomial SVM, a threshold of $0.546 < t^* < 0.622$ provided this improved classification.

Figure 3.4d presents the precision-recall curve for the KNN model. This model had an area under the precision-recall curve of 0.41. This precision recall curve indicated that the threshold of $t^* = 0.5$ did not provide the optimal classification results available; the optimal model had a precision of 1 and recall of 0.83333 for a threshold between $0.67 < t^* < 0.83333$.

The radial SVM model's precision recall curve is presented in figure 3.4b and had an area under the precision-recall curve of 0.72. Similarly to linear and polynomial SVM models, for a threshold $t^* \neq 0.5$, the model was able to distinguish between clear cell and non-clear samples with a greater precision. A threshold between $0.754 < t^* < 0.883$ provided a precision of 1 with the corresponding recall of 0.83333; indicating that this model could reduce the number of false positive misclassifications to 0.

The precision-recall curve for the random forest model is presented in figure 3.4e with an area under the precision-recall curve of 0.74. Based on this precision-recall curve, it is clear that the random forest model at a threshold of $t^* = 0.5$ provided the optimal model in terms of maximising the F_1 score; it achieved a precision of 1 and recall of 0.83333.

When comparing each of the optimal models when varying the threshold for positive classifications, each of the models produced one false negative classification for clear cell sample 4.

3.3 Classifying Patients with Mucinous EOC

3.3.1 Gene Signature

Twenty five of the 70 statistically significant genes found between mucinous and non-mucinous EOC samples formed the gene signature used to distinguish mucinous EOC from non-mucinous EOC. These genes can be found in table 3.6, below.

Table 3.6: Gene signature for classifying mucinous vs non-mucinous samples.

Probe ID	Gene Symbol	Gene Name
209956_s_at	CAMK2B	Calcium/calmodulin dependent protein kinase II beta
219505_at	CECR1	Cat eye syndrome chromosome region, candidate 1
228825_at	PTGR1	Prostaglandin reductase 1
219786_at	TESMIN	Testis expressed metallothionein like protein
230527_at	LOC101926907	Uncharacterized LOC101926907
219277_s_at	OGDHL	Oxoglutarate dehydrogenase-like
209309_at	AZGP1	Alpha-2-glycoprotein 1, zinc-binding

Continued on next page

Table 3.6 – continued from previous page

Probe ID	Gene Symbol	Gene Name
213059_at	CREB3L1	cAMP responsive element binding protein 3 like 1
219796_s_at	CDHR5	Cadherin related family member 5
224939_at	NUFIP2	NUFIP2, FMR1 interacting protein 2
37005_at	MINOS1-NBL1///NBL1	MINOS1-NBL1 readthrough///neuroblastoma 1, DAN family BMP antagonist
208063_s_at	CAPN9	Calpain 9
204607_at	HMGCS2	3-hydroxy-3-methylglutaryl-CoA synthase 2
204667_at	FOXA1	Forkhead box A1
210015_s_at	MAP2	Microtubule associated protein 2
1558622_a_at	ZNF548	zinc finger protein 548
230935_at	HAGLROS	HAGLR opposite strand (non-protein coding)
203564_at	FANCG	Fanconi anemia complementation group G
242915_at	ZNF682	zinc finger protein 682
231939_s_at	BDP1	B double prime 1, subunit of RNA polymerase III transcription initiation factor IIIB
235188_at	PCNX4	Pecanex homolog 4 (Drosophila)
1559746_a_at	ZNF718///ZNF595	Zinc protein finger 718///zinc protein finger 595
213496_at	PLPPR4	Phospholipid phosphatase related 4
209125_at	KRT6A	Keratin 6A
231070_at	IYD	iodotyrosine deiodinase

3.3.2 Principal Component Analysis

As a form of exploratory data analysis, principal component analysis was employed to determine whether linear combinations of the gene signature presented in table 3.6 could separate the training mucinous and non-mucinous samples in two-dimensions.

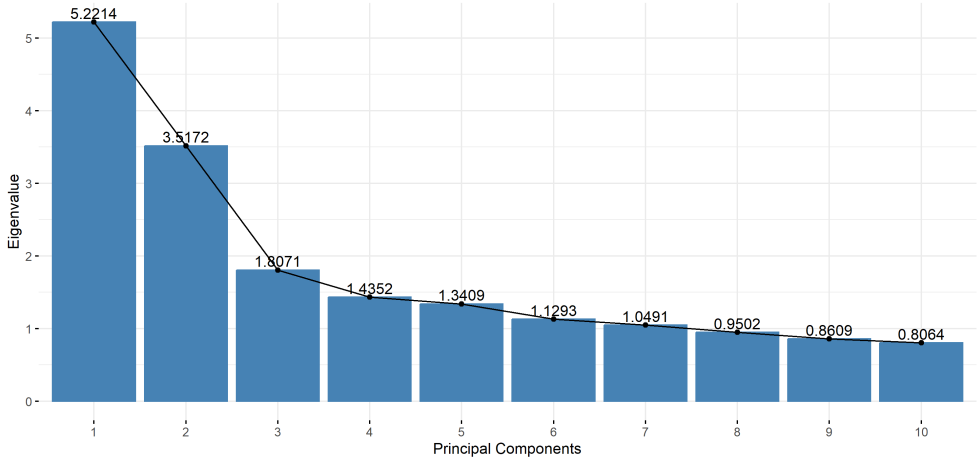


Figure 3.5: Screeplot of the first ten principal components for principal component analysis of mucinous EOC gene signature.

Figure 3.5 displays the screeplot produced for the first ten principal components. With eigenvalues greater than 1, the first seven principal compo-

nents account for more variance than any of the original 25 genes. If this analysis was to consider principal components that provide a cumulative variance of approximately 95%, 21 principal components would be retained. To reduce the retained variance below 95% would not be suitable in a diagnostic context. Therefore, once visualisation of the data in two-dimensions was complete, the original 25 genes were retained for further analysis with no pre-processing.

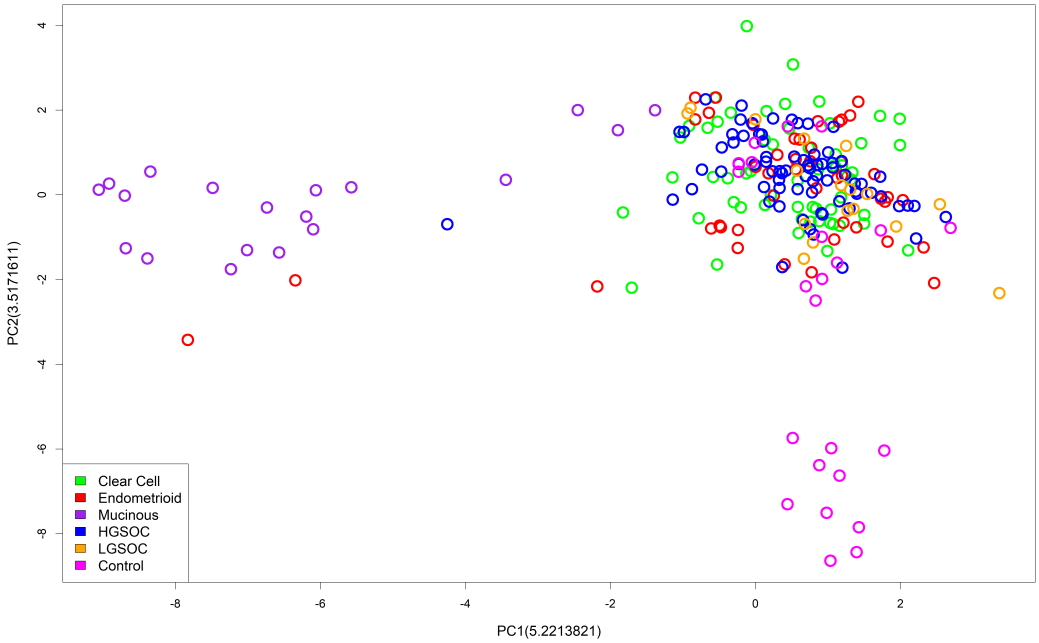


Figure 3.6: Two dimensional principal component plot of samples using mucinous gene signature.

Figure 3.6 displays the samples in two dimensions based on the first two principal components. It indicated that the linear combination of the 25 genes representing the first principal component could separate a large number of mucinous samples and non-mucinous samples into separate clusters. Thus, supporting the potential ability of the suggested gene signature in distinguishing mucinous from non-mucinous EOC. This figure shows that two endometrioid and one HGSOC samples may have had greater similarities in gene expression to the 19 mucinous samples than to all other non-mucinous samples. However, the principal component plot still suggests an ability to separate mucinous and non-mucinous cases.

3.3.3 Classification Model Training and Testing

Mucinous was described as the positive class.

Linear SVM

The optimal cost parameter was $C = 1$ with the greatest mean precision-recall AUC score of 0.41021. This model used a total of 15 support vectors; 7 of which were mucinous samples and 8 were non-mucinous samples. The model predicted all samples correctly.

Radial SVM

The optimal parameters for radial SVM were C=1.25 and $\gamma = 0.01$. These parameters achieved the highest mean precision-recall AUC of 0.42417. This model used 48 support vectors; 6 mucinous and 42 non-mucinous. This model had three false positive classifications.

Polynomial SVM

The optimal parameters for maximising the precision-recall AUC score were $C = 1$, $scale = 0.01$ and $degree = 3$. The maximum mean precision-recall AUC was 0.41792. The number of support vectors used in this model was 36; 5 were mucinous samples and 31 were non-mucinous samples. This model had three false positive classifications.

KNN

Fifteen nearest neighbours was selected as the optimal parameter value of k when using Euclidean distance. This optimal parameter achieved an optimal mean precision-recall AUC score of 0.20729. This training model had a total of 4 false negative and 3 false positive classifications.

Random Forest

A random forest model using 500 trees tried one random gene at each split. This achieved the maximised mean precision-recall AUC of 0.38019. This model had nine false negative classifications.

Table 3.7: Table providing comparisons of Jaccard index (JI), Fowlkes Mallows index (FMI), precision, recall, specificity and F_1 score for the mucinous training classification models.

Classifier	JI	FMI	Precision	Recall	Specificity	F_1 Score
Linear SVM	**1.00000**	**1.00000**	1.00000	1.00000	1.00000	**1.00000**
Radial SVM	0.86364	0.92932	0.86364	1.00000	0.98636	0.92683
Polynomial SVM	0.86364	0.92932	0.86364	1.00000	0.98636	0.92683
KNN	0.68182	0.81111	0.83333	0.78947	0.98636	0.81081
RF	0.52632	0.72548	1.00000	0.52632	1.00000	0.68966

Table 3.8: Table providing comparisons of Jaccard Index (JI), Fowlkes Mallows Index (FMI), precision, recall, specificity, F_1 score and PR-AUC for mucinous test predictions.

Classifier	JI	FMI	Precision	Recall	Specificity	F_1 Score	PR-AUC
Linear SVM	**1.00000**	**1.00000**	**1.00000**	1.00000	**1.00000**	**1.00000**	0.86
Radial SVM	0.77778	0.88192	0.77778	1.00000	0.97561	0.87500	0.74
Polynomial SVM	0.77778	0.88192	0.77778	1.00000	0.97561	0.87500	0.86
KNN	0.70000	0.83666	0.70000	1.00000	0.96341	0.82353	0.49
RF	0.50000	0.67612	0.80000	0.57143	0.98780	0.66667	0.76

Tables 3.7 and 3.8 present the Jaccard index, Fowlkes Mallows index, precision, recall, specificity and F_1 score for each of the five training and test models for classifying samples with mucinous and non-mucinous EOC.

The linear SVM model provided the best classification performance of all five models when applied to the training data according to the F_1 metric. The linear SVM model correctly classified all mucinous and non-mucinous samples.

Trained radial and polynomial SVM models achieved equivalent classification performances. They both made three false positive classifications and both misclassified the same non-mucinous samples (82, 90 and 166).

The KNN model made four false negative classifications and 3 false positive classifications. The samples misclassified by the KNN model were (1, 3, 6, 10, 82, 90 and 166). This model misclassified the same non-mucinous samples as the radial and polynomial SVM.

The random forest model made nine false negative misclassifications. The samples misclassified in the training of this model were 1, 2, 3, 4, 5, 6, 7, 10 and 19. Although this model did not make any false positive misclassifications, overall this model performed the worst of all five models according to the F_1 metric.

Similarly to the the training model abilities, for a threshold of $t > 0.5$, the linear SVM had the greatest test classification performance and correctly classified all samples.

Both radial and polynomial SVM models at a threshold of $t > 0.5$ were able to classify the test samples with equal success. Both models made no false negative classifications but made two false positive classifications.

For the KNN model, when considering the probabilities/scores of a test sample being classified as a member of the positive class, a number of samples had a probability/score of exactly 0.5 for being classified as both mucinous and non-mucinous histology. Therefore, the threshold taken to classify a sample as the positive class was $t \geq 0.5$. Hence, for a threshold $t \geq 0.5$, three false positive classifications were made.

The random forest model at a threshold of $t > 0.5$ had the worst test classification performance of all five models indicated by the F_1 metric. This model made three false negative classifications and one false positive classification.

The most commonly misclassified test sample was 16. Radial SVM, polynomial SVM and KNN models all misclassified sample 54. The KNN model was the only model to misclassify the non-mucinous sample 37. Furthermore, the random forest model was the only model to misclassify the mucinous samples 1, 3 and 5.

The only non-mucinous histologies that were misclassified as mucinous samples were endometrioid and HGSOC.

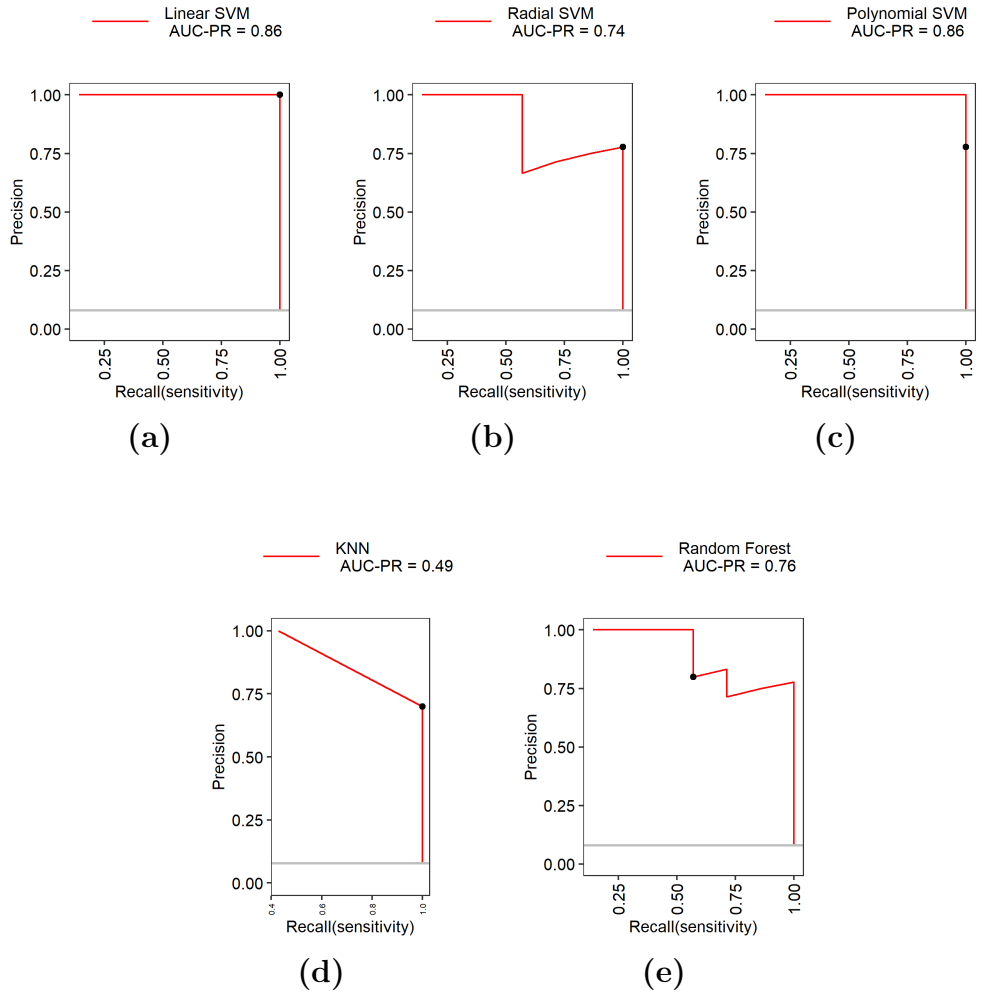


Figure 3.7: **(a)** Linear SVM precision-recall curve. **(b)** Radial SVM precision-recall curve. **(c)** Polynomial SVM precision-recall curve. **(d)** KNN precision-recall curve. **(e)** Random Forest precision-recall curve. The precision and recall of classifiers at a threshold of $t > 0.5$ are represented by black points.

Figure 3.7 displays the five precision-recall curves corresponding to the five classification models linear SVM, radial SVM, polynomial SVM, KNN and random forest, respectively. Based on the threshold $t > 0.5$ as discussed previously, the linear SVM model was the optimal model for classification due to all of the training and test samples being correctly classified.

Figure 3.7c represents the precision-recall curve for the polynomial SVM

model. The threshold $t > 0.5$, represented by the black point, was not the optimal classification result this model could achieve; by the precision-recall curve it is clear that both a precision and recall of 1 could be achieved. This optimal precision and recall was achieved at threshold values of $0.5793 < t^* < 0.5845$. Figure 3.7b displays the precision-recall curve for the radial SVM model when applied to the test data. There was no threshold value of t^* that would provide both a precision and recall of 1 when classifying samples for this model. Instead, the optimal performance for this model was found at the threshold value of $t^* = 0.5$ with a precision, recall and F_1 score of 0.77778, 1 and 0.875, respectively. Hence, the best performance of the radial SVM model had two false positive classifications. The KNN model's precision-recall curve is presented in figure 3.7d. This model only achieved a precision of 1 when recall was 0.4. The best performance for this model was found for the threshold $t \geq 0.5$ making three false positive classifications. Figure 3.7e presents the precision-recall curve for the random forest classification model. At no threshold value of t^* could this model reach both a precision and recall of 1. However, the model at a threshold of $t > 0.5$ did not provide the best classification performance this model could achieve. The best performance for the random forest model was found for a threshold value of $0.208 < t^* < 0.314$; these threshold values led to two false positive classifications but no false negative classifications.

3.4 Classifying Patients with LGSOC

3.4.1 Gene Signature

Eighteen genes were taken forward as a gene signature for distinguishing LGSOC samples from non-LGSOC samples. These genes are listed in table 3.9, below.

Table 3.9: Gene signature for classifying LGSOC vs non-LGSOC samples.

Probe ID	Gene Symbol	Gene Name
227999_at	PWWP2B	PWWP domain containing 2B
205053_at	PRIM1	Primase (DNA) subunit 1
212592_at	JCHAIN	Joining chain of multimeric IgA And IgM
204610_s_at	CCDC85B	Coiled-coil domain containing 85B
200887_s_at	STAT1	Signal transducer and activator of transcription 1
201985_at	KIAA0196	KIAA0196
211959_at	IGFBP5	Insulin like growth factor binding protein 5
224603_at	SNHG16///SNORD1C///SNORD1A	Small nucleolar RNA host gene16///
		Continued on next page

Table 3.9 – continued from previous page

Probe ID	Gene Symbol	Gene Name
		small nucleolar RNA, C/D box 1C/// small nucleolar RNA, C/D box 1A
214096_s_at	SHMT2	Serine hydroxymethyltransferase 2
223274_at	TCF19	Transcription factor 19
205582_s_at	GGT5	Gamma-glutamyltransferase 5
204441_s_at	POLA2	DNA polymerase alpha 2, accessory subunit
208012_x_at	SP110	SP110 nuclear body protein
220091_at	SLC2A6	Solute carrier family 2 member 6
212479_s_at	RMND5A	Required for meiotic nuclear division 5 homolog A
222452_s_at	GPBP1L1	GC-rich promoter binding protein 1 like 1
226310_at	RICTOR	RPTOR independent companion of MTOR complex 2
205329_s_at	SNX4	Sorting nexin 4

3.4.2 Principal Component Analysis

Principal component analysis was applied as exploratory data analysis to the gene signature presented in table 3.9. The aim was to visualise how these genes may separate the samples in two-dimensions.

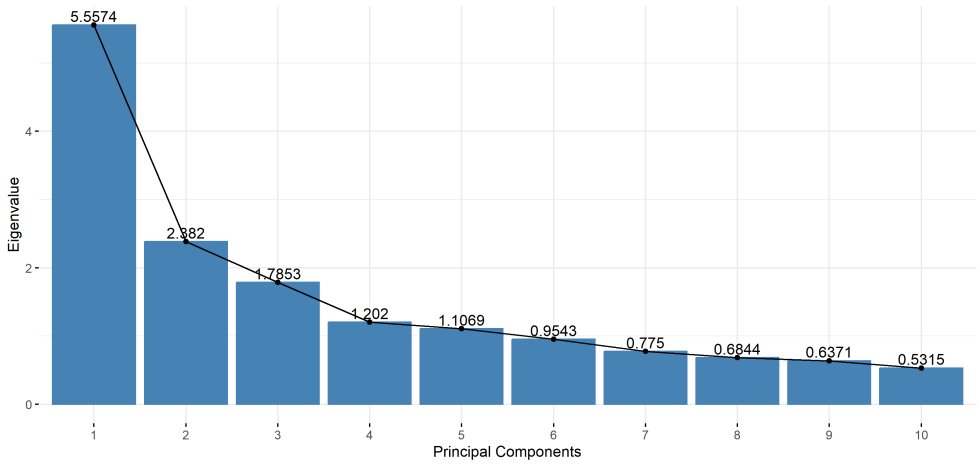


Figure 3.8: Screeplot of the first ten principal components for principal component analysis of the LGSOC gene signature.

The screeplot presented in figure 3.8 signals that of the first ten principal components, the first five account for greater variance than any of the original 18 genes in the gene signature. However, 14 of the 18 principal components would be required to provide a cumulative variance proportion of 95%. Less

than this number of principal components would not retain a suitable amount of variance for diagnostic purposes. Therefore, principal component analysis was not used to pre-process data prior to classification model application.

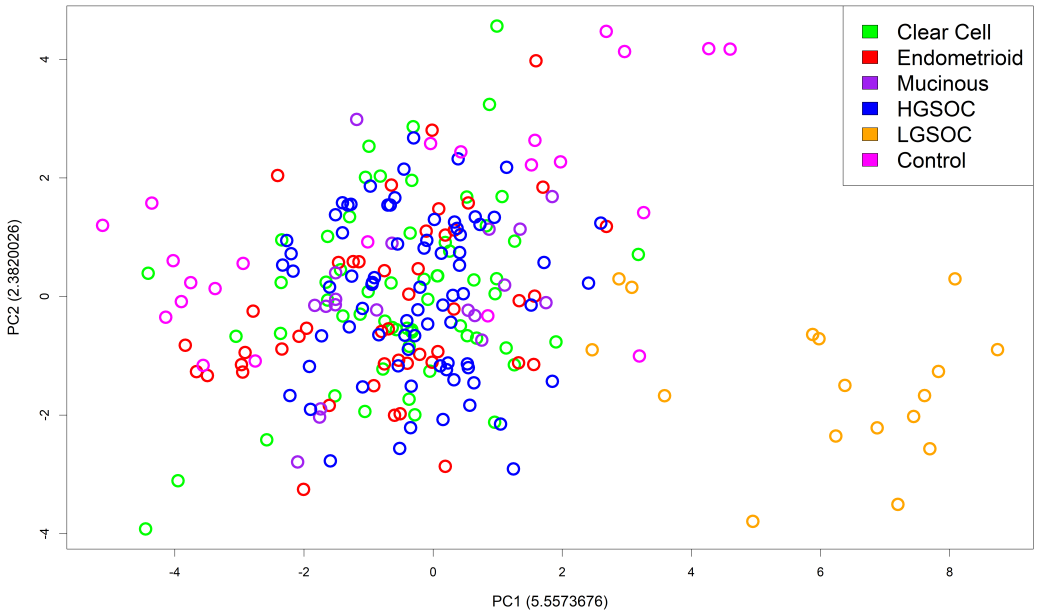


Figure 3.9: Two dimensional principal component plot of samples using LGSOC gene signature.

Figure 3.9 presents all 239 training samples with respect to the first two principal components. This figure shows that at least 13 LGSOC samples could be separated well from all other samples based on the first linear combination of the LGSOC gene signature represented by the first principal component. It is clear that non-LGSOC samples form a cluster to the left hand side of this principal component plot. Whereas, there is a cluster of LGSOC samples to the right. Four LGSOC samples were less well separated from non-LGSOC samples. One particular sample that was not separated from LGSOC was a normal ovary sample. Furthermore, the samples closest to the three LGSOC samples not separated from other EOC subtypes were of histology clear cell and HGSOC. Although perfect separation was not seen, it did support the hypothesis that the chosen gene signature had potential to distinguish between LGSOC and non-LGSOC cases.

3.4.3 Classification Model Training and Testing

LGSOC was described as the positive class.

Linear SVM

A cost parameter of $C = 0.1$ supplied the optimal model for classifying LGSOC vs non-LGSOC with respect to the chosen gene signature. This C provided the optimal mean precision-recall AUC score 0.34875. This model used 25 support vectors; 3 were LGSOC samples and 22 are non-LGSOC samples. This final model made two false positive classifications.

Radial SVM

A model with cost parameters $C = 5$ and $\gamma = 0.07$ was optimal for classifying LGSOC vs non-LGSOC samples based on the maximum mean precision-recall AUC 0.35292. For this model 53 support vectors were used; 14 of these support vectors were LGSOC samples and 39 were non-LGSOC samples. This model correctly classified all training samples.

Polynomial SVM

A model with parameters $C = 0.25$, $degree = 2$ and $scale = 0.1$ was optimal for classifying LGSOC samples vs non-LGSOC samples. These parameters were chosen based on the optimal mean precision-recall AUC score 0.35083. There were 22 support vectors used in this model; 4 LGSOC samples and 18 non-LGSOC samples. This training model had one false positive classification.

KNN

For KNN classification of LGSOC samples, five nearest neighbours was selected as the optimal k when using a Euclidean distance with the maximum mean precision-recall AUC of 0.15917. This model had four false negative classifications.

Random Forest

A random forest model using 500 trees tried one random gene at each split. The maximised mean precision-recall AUC was 0.30358. This training model

had four false negative classifications.

Table 3.10: Table providing comparisons of Jaccard Index (JI), Fowlkes Mallows Index (FMI), precision, recall, specificity and F_1 score for the LGSOC training classification models.

Classifier	JI	FMI	Precision	Recall	Specificity	F_1 Score
Linear SVM	0.89474	0.94591	0.89474	1.00000	0.99099	0.94444
Radial SVM	**1.00000**	**1.00000**	1.00000	1.00000	1.00000	**1.00000**
Polynomial SVM	0.94444	0.97183	0.94444	1.00000	0.99550	0.97143
KNN	0.76471	0.87447	1.00000	0.76471	1.00000	0.86667
RF	0.76471	0.87447	1.00000	0.76471	1.00000	0.86667

Table 3.11: Table providing comparisons of Jaccard Index (JI), Fowlkes Mallows Index (FMI), precision, recall, specificity, F_1 score and PR-AUC for LGSOC test predictions.

Classifier	JI	FMI	Precision	Recall	Specificity	F_1 Score	PR-AUC
Linear SVM	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.67
Radial SVM	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.67
Polynomial SVM	0.75000	0.86603	0.75000	1.00000	0.98837	0.85714	0.67
KNN	0.75000	0.86603	0.75000	1.00000	0.98837	0.85714	0.24
RF	0	NA	NA	0	1.00000	0	0.67

Tables 3.10 and 3.11 present the evaluation metrics for the training and testing of each of the five classification models.

The model with the greatest classification ability when applied to the training data was radial SVM; an F_1 metric of 1 indicates no misclassifications.

Both trained linear and polynomial SVM models were able to achieve a recall of 1 indicating no false negative classifications. However, the polynomial and linear SVM models made one and two false positive classifications, respectively. The polynomial SVM had a greater F_1 score than the linear SVM model.

Trained KNN and random forest models had a precision of 1 signifying no false positive classifications. However, both models made four false negative classifications.

The linear and polynomial SVM models misclassified non-LGSOC sample 36. In addition, linear SVM misclassified the non-LGSOC sample 128. The

KNN and random forest models both misclassified LGSOC samples 1, 2, 3 and 4. The non-LGSOC subtypes most commonly misclassified as LGSOC were clear cell and mucinous.

At a threshold of $t > 0.5$, the linear and radial SVM correctly classified all test samples. For this same threshold, the polynomial SVM and KNN model achieved equivalent test classification performances achieving a recall of 1 indicating no false negative classifications. However, their precision was 0.75 as both models misclassified the non-LGSOC test sample 20. The random forest classifier at a threshold of $t > 0.5$ misclassified all test LGSOC samples. The non-LGSOC subtype most commonly misclassified as LGSOC for test data was mucinous.

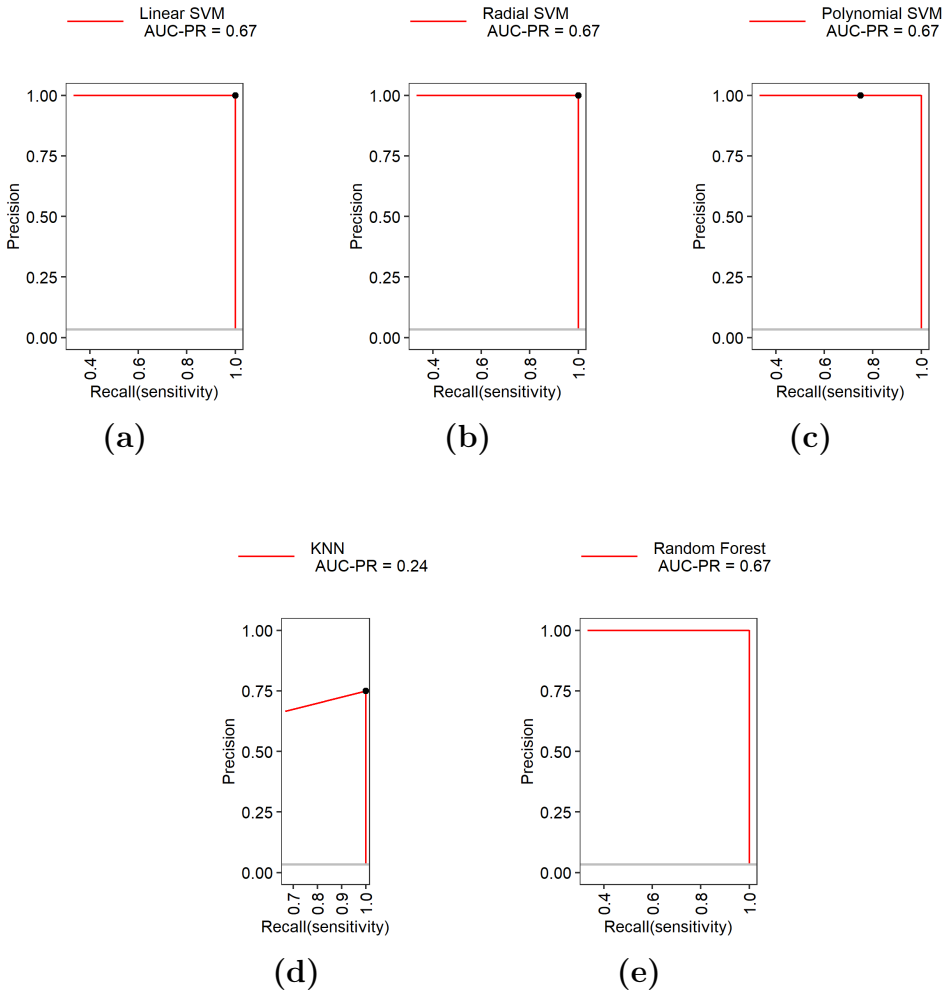


Figure 3.10: **(a)** Linear SVM precision-recall curve. **(b)** Radial SVM precision-recall curve. **(c)** Polynomial SVM precision-recall curve. **(d)** KNN precision-recall curve. **(e)** Random Forest precision-recall curve. The precision and recall of classifiers at a threshold of $t > 0.5$ are represented by black points.

Figure 3.10 displays the precision-recall curves for all five classification models when applied to test data.

The linear and radial SVM models, whose precision-recall curves are shown in figures 3.10a and 3.10b, provided perfect classification of the test data. At the threshold $t > 0.5$, the precision and recall of both models were 1 and this is represented by the black point in both figures.

At the standard threshold of $t > 0.5$, the polynomial SVM model did not provide a perfect classification of the test data. However, the precision-recall curve for this model found in figure 3.10c indicates that it could be achieved. For a threshold value in the range of $0.5367 < t^* < 0.9131$, all samples could be correctly classified.

The random forest model for a threshold in which $t^* \neq 0.5$ was also able to achieve perfect precision and recall. The value of precision and recall at $t^* = 0.5$ cannot be represented in figure 3.10e due to a recall value of 0 and a precision of $\frac{0}{0}$. For a threshold value in the range of $0.238 < t^* < 0.38$, all samples were correctly classified.

The KNN model could not achieve perfect classification for any threshold value of t^*. The best performance for this model was found for $t \geq 0.5$ due to one sample of LGSOC having a probability/score of 0.5 for being classified as LGSOC or non-LGSOC histology. Therefore, the best performance for the KNN model made one false positive classification of sample 20.

3.5 Classifying Endometrioid vs Clear Cell Patients

3.5.1 Gene Signature

Eleven of 12 genes found to be differentially expressed between endometrioid and non-endometrioid samples were used in the gene signature for differentiating between endometrioid and clear cell EOC samples. These genes are listed in table 3.12, below.

Table 3.12: Gene signature for classifying endometrioid vs clear cell samples.

Probe ID	Gene Symbol	Gene Name
201051_at	ANP32A	acidic nuclear phosphoprotein 32 family member A
212922_s_at	SMYD2	SET and MYND domain containing 2
212533_at	WEE1	WEE1 G2 checkpoint kinase
200907_s_at	PALLD	palladin, cytoskeletal associated protein
227722_at	RPS23	ribosomal protein S23
227525_at	GLCCI1	glucocorticoid induced 1
219036_at	CEP70	centrosomal protein 70
206378_at	SCGB2A2	secretoglobin family 2A member 2
214444_s_at	PVR	poliovirus receptor
232878_at	NR2F2-AS1	NR2F2 antisense RNA 1

Continued on next page

Table 3.12 – continued from previous page

Probe ID	Gene Symbol	Gene Name
202047_s_at	CBX6	chromobox 6

3.5.2 Principal Component Analysis

Principal component analysis was applied to the training samples based on the gene signature presented in table 3.12. Principal component analysis was used as a form of exploratory data analysis to visualise the training samples in two-dimensions. Here the two dimensions are represented by linear combinations of the original 11 genes in the gene signature.

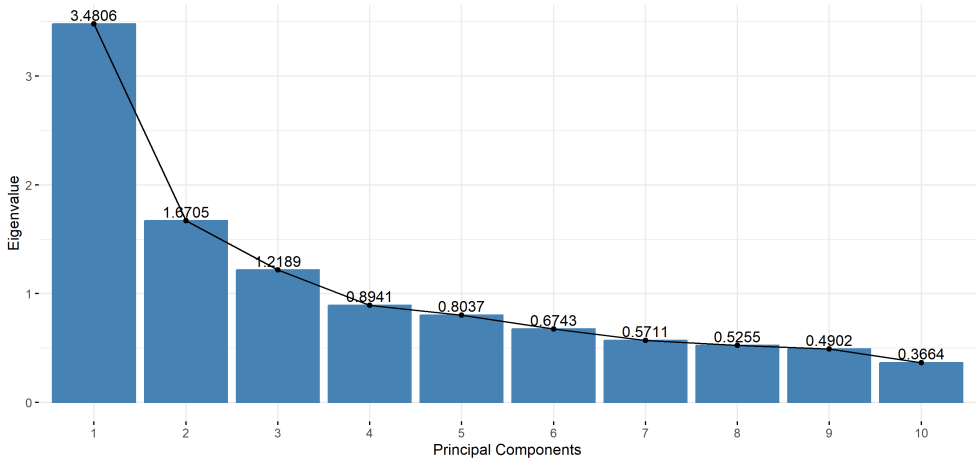


Figure 3.11: Screeplot of the first ten principal components for principal component analysis of clear cell versus endometrioid EOC gene signature.

Figure 3.11 displays the screeplot for the first ten principal components based on the gene signature suggested. It reveals that the first three principal components provided more variation than any of the original 11 genes due to eigenvalues being greater than 1. However, when considering the number of eigenvalues needed to allow 95% of the variance to be retained, ten of the 11 principal components would have to be retained. It would not be suitable in this context to reduce the variation to less than this level. Therefore, principal component analysis was used to visualise the training samples in less dimensions but not for pre-processing of data for classification.

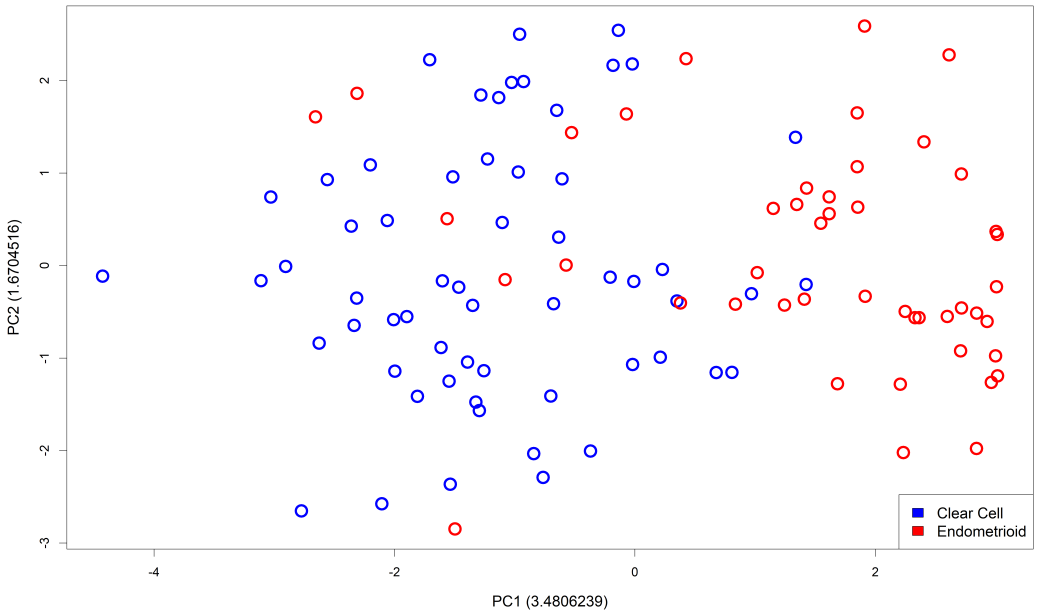


Figure 3.12: Two dimensional principal component plot of samples using clear vs endometrioid gene signature.

Figure 3.12 provides a visualisation of both clear cell and endometrioid samples when represented in two dimensions based on the first two principal components. It is clear that based on the linear combination of these 11 genes represented by the first principal component, there is a degree of separation between the clear cell and endometrioid training samples. It suggests that the use of this gene signature as a method of distinguishing between clear cell and endometrioid EOC cases has potential.

3.5.3 Classification Model Training and Testing

For each of the models, endometrioid EOC was designated as the positive class.

Linear SVM

Cross validation indicated that the optimal parameter C (cost) was $C = 0.01$ with the highest mean precision-recall AUC score of 0.74388. This model used a total of 72 support vectors; 40 of which were clear cell samples and 32 were endometrioid samples. The model had 6 false positive and 9 false negative classifications.

Radial SVM

Cross validation determined that the optimal parameters for radial SVM were $C = 1$ and $\gamma = 0.09$. These parameters had the highest mean precision-recall AUC of 0.76523. This model used 66 support vectors; 38 were clear cell samples and 28 were endometrioid samples. This training model had a total of 2 false positive and 5 false negative classifications.

Polynomial SVM

Following cross validation, the optimal parameters for maximising the mean precision-recall AUC were $C = 2$, $scale = 0.1$ and $degree = 3$. The maximum mean precision-recall AUC based on these parameters was 0.75263. The number of support vectors used in this model was 47; 26 were clear cell samples and 21 were endometrioid samples. The training model had only 1 false negative classification.

KNN

Cross validation selected 13 nearest neighbours as the optimal model parameter when using Euclidean distance. Thirteen nearest neighbours gave an optimal mean precision-recall AUC of 0.60735. This training model had 5 false positive and and 9 false negative classifications.

Random Forest

The optimal random forest model for this given gene signature tested one random gene at each split. This model gave a maximum mean precision-recall AUC of 0.76224. This model had 5 false positive and 8 false negative classifications.

Table 3.13: Table providing comparisons of Jaccard index (JI), Fowlkes Mallows index (FMI), precision, recall, specificity and F_1 score for clear cell vs endometrioid training classification models.

Classifier	JI	FMI	Precision	Recall	Specificity	F_1 Score
Linear SVM	0.71154	0.83193	0.86047	0.80435	0.89831	0.83146
Radial SVM	0.85417	0.92187	0.95349	0.89130	0.96610	0.92135
Polynomial SVM	**0.97826**	**0.98907**	**1.00000**	**0.97826**	**1.00000**	**0.98901**
KNN	0.72549	0.84178	0.88095	0.80435	0.91525	0.84091
RF	0.74510	0.85442	0.88372	0.82609	0.91525	0.85393

Table 3.14: Table providing comparisons of Jaccard index (JI), Fowlkes Mallows index (FMI), precision, recall, specificity, F_1 score and PR-AUC for clear cell vs endometrioid test predictions.

Classifier	JI	FMI	Precision	Recall	Specificity	F_1 Score	PR-AUC
Linear SVM	0.33333	0.57735	1.00000	0.33333	1.00000	0.50000	0.78
Radial SVM	0.66667	0.81650	1.00000	0.66667	1.00000	0.80000	0.83
Polynomial SVM	**1.00000**	**1.00000**	1.00000	**1.00000**	1.00000	**1.00000**	0.83
KNN	0.16667	0.40825	1.00000	0.16667	1.00000	0.28571	0.83
RF	0.66667	0.81650	1.00000	0.66667	1.00000	0.80000	0.83

Tables 3.13 and 3.14 provide the evaluation metrics for the training and test models for classifying clear cell and endometrioid samples. For the trained models, polynomial SVM had the greatest ability to classify these two types of samples shown by the greatest F_1 score. This model was able to produce no false positives and only one false negative. Whereas, the linear SVM model had the worst training classification abilities with the lowest F_1 score.

The only sample misclassified by all five models in training was endometrioid sample 69; this was the only misclassified sample for polynomial SVM. The clear cell samples incorrectly classified in linear SVM, radial SVM, KNN and random forest models were 30, and 59. Whereas, the endometrioid samples incorrectly classified by these four models were 61, 63, 65 and 71.

Clear cell sample 13 was only misclassified by the linear SVM model. Both clear cell samples 49 and 58 were only misclassified by the random forest model. In addition, the endometrioid sample 67 was only misclassified by the random forest model.

Linear SVM and KNN shared the greatest number of similar sample misclassifications when comparing pairs of classification training models (clear cell samples 21, 30, 48, 54 and 59 and endometrioid samples 61, 62, 63, 64,

65, 66, 69, 71 and 84).

Similarly to the training models, polynomial SVM was the optimal model for classification of test endometrioid and clear cell samples. In this case, at a standard threshold of $t > 0.5$, the polynomial SVM was able to classify endometrioid and clear cell samples with a test F_1 score of 1. The KNN model provided the worst classification of endometrioid and clear cell samples when applied to test data at a threshold $t > 0.5$; no false positive classifications were made but five false negative classifications were made. The linear SVM model made no false positive test classifications at a threshold of $t > 0.5$. However this model did make four false negative classifications. Both the radial SVM and random forest models for a threshold of $t > 0.5$ made no false positive test classifications but both models did make two false negative test classifications each.

The test sample most commonly misclassified was endometrioid sample 11. Similarly to the training data, linear SVM and KNN models misclassified the most similar sets of test samples (endometrioid samples 8, 9, 10 and 11). Furthermore, the random forest model was the only model to misclassify endometrioid test sample 7.

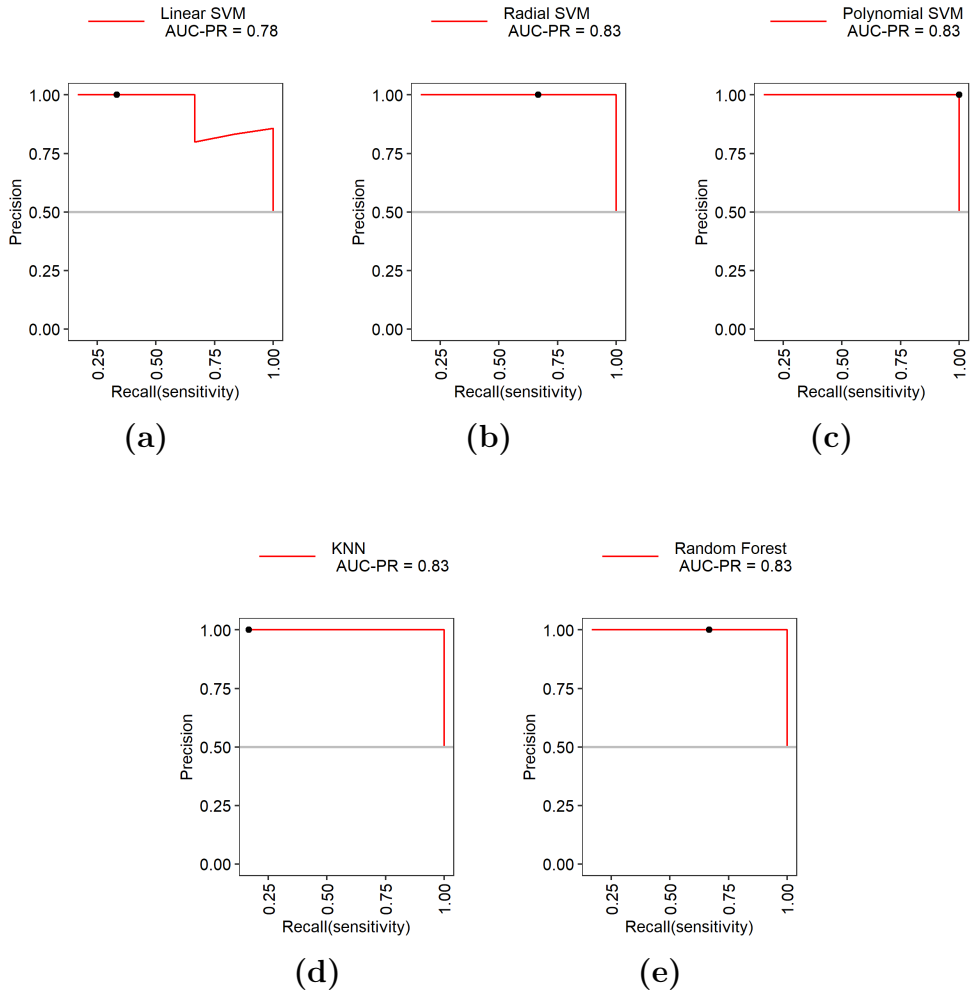


Figure 3.13: **(a)** Linear SVM precision-recall curve. **(b)** Radial SVM precision-recall curve. **(c)** Polynomial SVM precision-recall curve. **(d)** KNN precision-recall curve. **(e)** Random Forest precision-recall curve. The precision and recall of classifiers at a threshold of $t > 0.5$ are represented by black points.

Figure 3.13 displays the five precision-recall curves corresponding to the five classification models tested. As previously discussed, the polynomial SVM model was able to correctly classify all clear cell and endometrioid samples. This perfect classification with a precision and recall of 1 is represented by a black point and can be seen in figure 3.13c.

Figures 3.13b, 3.13d and 3.13e represent the precision-recall curves of the

radial SVM, KNN and random forest models, respectively. With the precision and recall value for the threshold $t > 0.5$ presented as a black point in the figures, it is clear that for some threshold $t > t^*$ where $t^* \neq 0.5$, a precision and recall of 1 could be achieved. For a threshold value of $0.1581 < t^* < 0.2015$, an F_1 score of 1 could be found for radial SVM. Moreover, a threshold value of $0.091 < t^* < 0.181$ gave an F_1 score of 1 for the KNN model. Finally, for the random forest model, a threshold value of $0.26 < t^* < 0.453$ gave an F_1 score of 1.

The linear SVM, whose precision-recall curve is represented by figure 3.13a, was the only classifier model that did not present a perfect classifier for any threshold t^*. The optimal classification performance for the linear SVM model was found for a threshold value $0.0225 < t^* < 0.0234$ and led to no false negative classifications and one false positive classification. The one clear cell sample misclassified as endometrioid was sample 4. This classifier had a precision of 0.85714 and a recall of 1.

Chapter 4

Discussion

4.1 Overview of Ovarian Cancer

A large number of studies have previously focused on accurately distinguishing between healthy samples or benign ovarian tumours and malignant ovarian tumours (Sans et al., 2019; Boylan et al., 2017; Han et al., 2018; Ke et al., 2015; Barnabas et al., 2019; Kawakami et al., 2019; Acharya et al., 2014; Lu et al., 2020b; Enroth et al., 2019; Aramendía-Vidaurreta et al., 2016; Suryawanshi et al., 2013).

Once diagnosis of OC is complete, standard treatments are available for patients including platinum-based chemotherapy (Lawrie et al., 2015; Kessous et al., 2017) and debulking surgery (Coleman et al., 2013; du Bois et al., 2009). However, due to the varying histological subtypes of EOC presenting diverse treatment responses (Karabuk et al., 2013; Alexandre et al., 2010; Bamias et al., 2010; Simons et al., 2015; Shu et al., 2015; Sugiyama et al., 2000; Itamochi et al., 2008; Schmeler et al., 2008; Grabowski et al., 2016; Tomasi Cont, 2015; Oseledchyk et al., 2017), the standard treatments available may not be suitable for all EOC histotypes. Furthermore, the five histotypes present different prognoses (Mackay et al., 2010; Zaino et al., 2011; Bamias et al., 2010; Alexandre et al., 2010; Sugiyama et al., 2000; Gilks et al., 2008; Bodurka et al., 2012; Peres et al., 2019; Lan & Yang, 2019; Simons et al., 2017) and common diagnosis stage (Kurman & Shih, 2016; Torre et al., 2018; Alcázar et al., 2013; Köbel et al., 2010; Shu et al., 2015; Zhu et al., 2021; Winterhoff et al., 2016; Brown & Frumovitz, 2014). Hence, the ability to differentiate between the five histological subtypes of EOC is critical in improving time to treatment, treatment choices, treatment response, overall survival rates and prognosis of patients.

4.2 Prior Classification Studies of EOC Subtypes

Prior studies aiming to distinguish between histotypes of EOC have used either biological biomarkers or imaging features to train classification models. Both Assem et al. (2018) and Dieters-Castator et al. (2019) have used biomarkers to predict whether patients' histology diagnoses were HGSOC or endometrioid, achieving 96% and 99.2% accuracy, respectively. An accuracy of 96% has also been reported by Woodbeck et al. (2019) when classifying patients by endometrioid or mucinous histology using biomarkers. Schwartz et al. (2002) proposed a 158 gene signature to distinguish between clear cell and non-clear cell patients achieving perfect classification of non-clear cell patients and misclassifying a single clear cell patient as non-clear cell histology. Further studies have considered the classification of more than two histotypes; a 93% accuracy was achieved when multiclass classification was performed on all five histotypes of EOC (Köbel et al., 2016). The classification of clear cell EOC, HGSOC, LGSOC and serous borderline tumours achieved a classification accuracy of 85% (Klein et al., 2019). The individual classification accuracies for HGSOC, clear cell, endometrioid and mucinous when using a 32 blood peripheral marker were 75.8%, 67.7%, 55.6% and 96%, respectively (Kawakami et al., 2019). In addition, when classifying clear cell and endometrioid patients against serous patients, a sensitivity and specificity of 86% and 79% were achieved (Suryawanshi et al., 2013). With HGSOC being the prevalent subtype of EOC, a number of studies focus on classifying HGSOC versus non-HGSOC patients. Qian et al. (2020) achieved one of the highest sensitivity and specificity scores of 0.88 and 0.97, respectively.

4.3 Current Study Purpose

This study had 2 main focuses: the detection of EOC and the diagnosis of EOC subtype. The aim was to determine whether genes that indicated presence or non-presence of EOC could also be used to identify a patient's diagnosed EOC subtype.

This study observed the existence of uniquely differentially expressed and dys-regulated probes when comparing each of the five EOC histological subtypes with normal ovaries. Furthermore, a number of these probes were not only found to hold uniquely differential expression relating to one subtype of EOC and normal ovaries, but also between this one subtype and all four other subtypes of EOC. Subsets of these unique probes were selected to form

signatures used to distinguish their related subtype from all other EOC subtypes.

Linear support vector machine, radial basis function support vector machine (radial SVM), polynomial support vector machine, K-nearest neighbour (KNN) and random forest models were employed to perform binary classifications for the following cases

1. Clear cell versus non-clear cell EOC.
2. Mucinous versus non-mucinous EOC.
3. LGSOC versus non-LGSOC.
4. Clear cell versus endometrioid EOC

4.4 Case 1: Clear Cell versus Non-Clear Cell EOC

4.4.1 Classification Results Overview

For the classification of clear cell and non-clear cell EOC, a gene signature of length 15 was selected and used to train and test the five classification models. The 15 genes (probe IDs) were GABARAPL1 (208868_s_at), CYS1 (228739_at), HAVCR1 (207052_at), LINC00472 (220324_at), RNF43 (218704_at), SLC45A4 (225597_at), HK1 (200697_at), SLC38A4 (220786_s_at), CYTIP (209606_at), C6orf132 (241455_at), LOC101930123///EEFK2 (225545_at), DYNLRB1 (217917_s_at), NPLOC4 (217796_s_at), MACROD2 (1563209_a_at) and FGB (216238_s_at).

For the classification of samples using this gene (probe) signature, firstly, a sample with a probability or score of greater than 0.5 was classified as clear cell histology. For this probability/score threshold, the random forest model provided the greatest test classification performance. This model performance achieved a test precision, recall and F_1 score of 100%, 83.33% and 90.909%, respectively. Considering these percentages in terms of classifications, all non-clear histology samples were correctly identified as non-clear cell histology. Whereas, one of six clear cell patient samples was incorrectly identified as non-clear cell histology. This is a comparable result to Schwartz et al. (2002) whose study used 153 genes in comparison to this study using 15 genes.

This study examined how changing the probability threshold for predicting a patient to have clear cell EOC could improve the performance of the other four classification models.

The linear SVM, radial SVM, polynomial SVM and KNN model performances could be improved to match that of the random forest model by varying the probability/score threshold described above. Hence, a range of probability/score values, t^*, for which a sample must exceed could be identified for each model that allowed for reduction in misclassifications. The result of this means that each model only misclassified one clear cell sample that is wrongly identified as non-clear cell EOC histology. The probability/score threshold values that allowed for this classification result for the linear SVM, radial SVM, polynomial SVM and KNN models were $0.571 < t^* < 0.617$, $0.754 < t^* < 0.883$, $0.546 < t^* < 0.622$ and $0.67 < t^* < 0.83333$, respectively.

Examining the misclassified sample for each of the models for their optimal probability/score threshold finds that it is the same clear cell EOC sample wrongly identified as non-clear cell EOC for all five models. One suggestion for this could be that this clear cell sample was previously misdiagnosed.

Regarding the trained models and their performances, the best performance on trained data is the radial SVM model. This model misclassified 13 samples in total and due to it misclassifying a smaller number of clear cell samples compared with all other trained models, it had a greater F_1 score, Jaccard index and Fowlkes Mallows index. However, when this trained model was tested on independent data, the radial SVM model had the worst classification performance of all five models. This could be due to the number of support vectors used. Linear and polynomial SVM used 87 and 88 support vectors, respectively. Whereas, the radial SVM model used 112 which may have reduced the models generalization ability and led to over-fitting of the training model to the training data. Both linear and polynomial SVM perform equivalently on the test and training data which may indicate that the addition of one clear cell support vector between the two models does not affect the classification performance of the models or the decision boundary significantly.

4.4.2 Clear Cell Gene Signature

A number of the genes used to form the gene (probe) signature for distinguishing between clear cell and non-clear cell EOC patient samples have previously been associated with ovarian cancer. A study concluded that higher expression levels of GABARAPL1 in OC was associated with reduced survival rates and increased risk of death (Chen et al., 2021d). HAVCR1's diagnostic ability in clear cell ovarian cancer has been indicated (Lin et al., 2007). LINC00472 was observed as highly expressed in early stage, low grade ovarian tumours (Fu et al., 2016). Reports found RNF43 to be frequently mutated in mucinous

EOC (Ryland et al., 2013; Zou et al., 2013). EEF2K was found to be up-regulated in primary and drug resistant ovarian cancer (Erdogan et al., 2021). This current study observed that GABARAPL1 (208868_s_at), HAVCR1 (207052_at), LINC00472 (220324_at) and LOC101930123///EEFK2 (225545_at) were up-regulated in clear cell EOC compared with normal ovary samples and were also differentially expressed in clear cell EOC when compared with endometrioid EOC, mucinous EOC, HGSOC and LGSOC. Based on the conclusions made by Chen et al. (2021d), this may suggest that GABARAPL1 reduces survival rates of patients diagnosed with clear cell EOC. This study provides further indication that HAVCR1 (207052_at) may hold diagnostic ability for clear cell EOC. Furthermore, clear cell EOC characteristics include more frequent early stage, low grade tumour which provides agreement with conclusions made by Fu et al. (2016) about LINC00472 expression in OC. Conclusions made by Erdogan et al. (2021) with respect to LOC101930123///EEFK2, also described as LOC101930123 or EEFK2, expression in primary and drug resistant OC may also be supported by this study as clear cell EOC are often characterised as having drug resistance. Results from this study indicated down-regulation of the gene (probe) RNF43 (218704_at) in clear cell EOC when compared with normal ovaries and differential expression between clear cell EOC and endometrioid EOC, mucinous EOC, HGSOC and LGSOC. This study does not find significant differences in RNF43 between mucinous EOC and normal ovaries or other EOC subtypes apart from clear cell.

Many of the genes included in the signature for distinguishing clear cell EOC from non-clear cell EOC were found to have associations with other cancers.

Low expression levels of GABARAPL1 have been detected in hepatocellular, lymph node positive breast and prostate cancers (Liu et al., 2014a; Berthier et al., 2010; Su et al., 2017). These expression levels of GABARAPL1 were indicative of poor prognosis in lymph node positive breast and prostate cancer (Berthier et al., 2010; Su et al., 2017).

High expression levels of CYS1 been associated with worse prognosis in (clear cell) renal cell carcinoma and was used to form a gene signature for prognosis prediction with six other genes (Xia et al., 2023). However, in pancreatic ductal adenocarcinoma CYS1 was observed to have greater expression in normal tissue when compared to tumour tissue (Jiang et al., 2008).

HAVCR1 has previously been suggested as a possible marker for renal cell carcinoma (Lin et al., 2007; Han et al., 2005; Scelo et al., 2018). High expression levels of HAVCR1 in colorectal cancer were found to be associated with improved disease-free survival (Wang et al., 2013).

LINC00472 was discovered in greater expression levels in less aggres-

sive tumours and was associated with improved prognosis for breast cancer (Shen et al., 2015). In contrast to the high expression of LINC00472 in breast cancer, LINC00472 was observed as down-regulated in colorectal cancer (Ye et al., 2018). However, similarly to its association in breast cancer, LINC00472's increased expression was suggested to be associated with reduced tumour growth and development (Ye et al., 2018).

RNF43 has been described as having a tumour suppressing role in gastric cancer (Niu et al., 2015).

SLC45A4 has opposing effects in patients with pancreatic ductal adenocarcinoma and osteocarcinoma. In pancreatic ductal adenocarcinoma, it is found more highly expressed in TP53 mutated tumours and is associated with poor clinical outcome (Chen et al., 2021c). Whereas, in osteocarcinoma, this gene was proposed to be favourable in prognosis of patients and was also included in a 17 gene signature that could accurately predict the outcomes of patients with osteosarcoma (Yang et al., 2022). Moreover, up-regulation of SLC45A4 was observed in both primary and metastatic melanoma compared with normal samples; for this study no significant association with prognosis was reported (Xie et al., 2021).

High expression levels of HK1 have been associated with poor prognosis in both colorectal cancer and advanced stage gastric cancer (He et al., 2016; Gao et al., 2015).

SLC38A4 has displayed low expression levels in hepatocellular cancer; this low expression has been associated with poor prognosis (Li et al., 2021b).

CYTIP was selected to be in a gene signature for classification of triple positive breast cancer subtypes; the gene signature achieved an accuracy of approximately 91.4% (Ben Azzouz et al., 2021).

A study found that expression of EEF2K in lung cancer and advanced stage stomach adenocarcinoma may also be associated with poor prognosis (Bircan et al., 2018; Jiang et al., 2021a).

A study has considered DYNLRB1's role in colorectal cancer and concluded that inhibition of this gene may be useful for the suppression and prevention of tumour cell growth (Jin et al., 2013).

NPLOC4 up-regulation was identified in lung squamous cell carcinoma and (clear cell) renal cell carcinoma with associations to poor overall survival (Wang et al., 2023; Yoshino et al., 2020).

Over-expression of MACROD2 was associated with tamoxifen-resistant tumours and poor overall survival for primary tumours in breast cancer (Mohseni et al., 2014). In cervical cancer research, MACROD2 has been described as a hotspot for HPV integration on the cervix which may indicate MACROD2 is involved in the development of cervical cancer (Zhao et al., 2023).

FGB plasma level was under-expressed in earlier stages of renal cell carcinoma post-neoadjuvant chemotherapy using the treatment sunitinib when compared with patients not receiving this treatment (Dzik et al., 2017). FGB plasma level has further been associated with post-neoadjuvant chemotherapy in breast cancer; higher expression levels had significant associations with poor prognosis (Mei et al., 2017). Furthermore, FGB was over-expressed in breast cancer with reduced expression leading to the inhibition of breast tumour growth (Liu et al., 2022a). Similarly, high pre-operative plasma FGB levels were detected in esophageal cancer and was associated with poor prognosis (Lv et al., 2018). Conversely, FGB was found to have lower pre-operative plasma levels in penile cancer (Ma et al., 2017). The associations between low levels of FGB and prognosis agree with Lv et al. (2018); low levels associated with significantly longer cancer specific survival and improved prognosis (Ma et al., 2017).

Limited information of studies on ovarian cancer that focus on the role of genes CYS1 (228739_at), SLC45A4 (225597_at), HK1 (200697_at), SLC38A4 (220786_s_at), CYTIP (209606_at), C6orf132 (241455_at), DYNLRB1 (217917_s_at), NPLOC4 (217796_s_at), MACROD2 (1563209_a_at) and FGB (216238_s_at) exist. This study implies that these genes may have an ability to identify whether a patient has clear cell EOC. Due to their associations with other cancers, there is plausability that these genes also have roles in the development of OC. Furthermore, there are limited studies on the role of C6orf132 (241455_at) in cancer. This study provides indication that C6orf132 may play a role in OC and have abilities to differentiate clear cell EOC from non-clear cell EOC.

4.5 Case 2: Mucinous versus Non-mucinous EOC

4.5.1 Classification Results Overview

The gene signature selected for distinguishing between mucinous and non-mucinous patients samples was of length 25. The genes (probes) included in this signature were CAMK2B (209956_s_at), CECR1 (219505_at), PTGR1 (228825_at), TESMIN (219786_at), LOC101926907 (230527_at), OGDHL (219277_s_at), AZGP1 (209309_at), CREB3L1 (213059_at), CDHR5 (219796_s_at), NUFIP2 (224939_at), MINOS1-NBL1///NBL1 (37005_at), CAPN9 (208063_s_at), HMGCS2 (204607_at), FOXA1 (204667_at), MAP2 (210015_s_at), ZNF548 (1558622_a_at), HAGLROS (230935_at), FANCG (203564_at), ZNF682 (242915_at), BDP1 (231939_s_at), PCNX4 (235188_at), ZNF718///ZNF595 (1559746_a_at), PLPPR4 (213496_at), KRT6A (209125_at)

and IYD (231070_at).

A sample with a probability/score of greater than 0.5 was the first threshold evaluated as predicting a patient as having a mucinous EOC tumour. The linear SVM model achieved the best test classification performance with 100% precision, recall and F_1 score. It is also worth noting that when the model was trained, the linear SVM model was also able to classify all samples with 100% precision and recall. This signifies that this model was able to completely distinguish between the mucinous and non-mucinous patient samples. With limited research regarding binary classification of mucinous EOC versus non-mucinous EOC using genomics, a direct comparison of results is not feasible. We will consider Qian et al. (2020) as a benchmark, who achieved an F_1 score of approximately 93% in classifying HGSOC versus non-HGSOC using MRI imaging. When comparing the result by Qian et al. (2020) with this study result for mucinous versus non-mucinous EOC it is clear that the gene signature chosen has predictive capabilities for this specific EOC subtype. Moreover, the capabilities of this gene signature may indicate that more research in genomics for classification of EOC histotypes is needed as it outperforms MRI imaging. This may however, be due to the sample size of the training and testing data. A larger sample size would be needed to determine whether this statement holds.

Further investigation on the effect of varying the probability/score threshold for predicting a patient as having a mucinous EOC tumour on the classification performance of radial SVM, polynomial SVM, KNN and random forest models took place.

The only other model that achieved 100% test precision and recall for a varied threshold value was the polynomial SVM. When prediction of a sample having mucinous EOC was based on a probability/score of greater than t^*, where $0.5793 < t^* < 0.5845$, the two non-mucinous EOC patient samples originally incorrectly classified as mucinous EOC were corrected providing 100% classification precision and recall.

A perfect classification could not be achieved by radial SVM, KNN or random forest models. The best classification performance achieved by the radial SVM model was in fact found using the original probability/score threshold. This best performance achieved a test precision, recall and F_1 score of 77.78%, 100% and 87.5%, respectively. The misclassifications made here were two non-mucinous EOC samples being identified as mucinous EOC histology. Similarly, the best test performance for the random forest model also misclassified two non-mucinous samples as mucinous histology, achieving precision, recall and F_1 scores of 77.78%, 100% and 87.5%, respectively. This improved performance for the random forest model was found at a threshold value of $0.208 < t^* < 0.314$.

Interestingly, for the KNN model the best classification performance was observed for a probability/score threshold of $t \geq 0.5$. This was due to four mucinous EOC samples and three non-mucinous EOC samples having an equal probability/score $t = 0.5$ of being classified as mucinous or non-mucinous histology. Therefore, the threshold was altered to become $t \geq 0.5$ such that all samples with a probability/score of 0.5 were classified as mucinous EOC histology. The best test performance of the KNN model misclassified three non-mucinous EOC samples as mucinous EOC histology. The histologies incorrectly identified as mucinous EOC tumours in these models were endometrioid and HGSOC.

The linear SVM model was able to provide perfect classification of both training and test data when using the suggested gene signature. For this model, the hyperplane was orientated based on 15 support vectors. As none of the support vectors were misclassified, it is clear that for perfect classification of both training and test data only a small number of margin violations were needed. Implications of this are that based on the gene signature described above, mucinous and non-mucinous EOC samples can be linearly separated.

4.5.2 Mucinous Gene Signature

Out of the genes investigated here CAMK2B, CECR1, OGDHL, AZGP1, FOXA1, HAGLROS, FANCG and BDP1 have previously been discussed with regards to ovarian cancer.

CAMK2B in high expression has been associated with improved prognosis for ovarian cancer (Jia et al., 2022); this current study found that CAMK2B is down-regulated in mucinous EOC which may indicate CAMK2B being a negative prognostic factor in mucinous EOC.

The significantly increased expression of CECR1 has also been detected in serous EOC and was associated with improved prognosis (Gao et al., 2022). In contrast, this current study found no significance in CECR1 expression in HGSOC or LGSOC when compared to normal ovaries. However, this study did find down-regulation of CECR1 in mucinous EOC.

Methylation of OGDHL has previously been discussed in ovarian cancer (Hoque et al., 2008).

AZGP1 presented no expression in normal ovaries when northern blotting was performed (Freije et al., 1991). Mucinous EOC was found to have greater expression of AZGP1 in this study compared with normal ovaries which may support the previous conclusion made by Freije et al. (1991).

Similarly to this current study, high FOXA1 expression has been associated with mucinous EOC (Sheta et al., 2021).

Significant up-regulation of HAGLROS in ovarian cancer has previously been associated with disease stage, prognosis and tumour size (Yang et al., 2019). HAGLROS was observed as down-regulated in mucinous EOC compared with normal ovaries. Early stage diagnosis is observed in greater frequency for mucinous EOC which may indicate agreement with the conclusion made by Yang et al. (2019).

The expression of FANCG was significantly increased in platinum-sensitive ovarian cancer patients than those with platinum-resistance (Xing et al., 2020). This current study found significantly lower expression levels of FANCG in mucinous EOC than in normal ovaries; platinum-resistance is frequently observed in mucinous EOC tumours which provides agreement with the conclusion made by Xing et al. (2020).

A decrease in BDP1 expression for serous EOC was observed, with high expression levels of BDP1 associated with later stage of development and worse prognosis (Cabarcas-Petroski et al., 2023). Although the current study did not find significant expression level differences between serous EOCs and normal ovaries, the current study observed that BDP1 had lower expression levels in mucinous EOC. As mucinous EOC is more frequently diagnosed in earlier stages, this study provides confirmation on this conclusion in the study by Cabarcas-Petroski et al. (2023).

Many of the genes selected in this study for distinguishing between mucinous and non-mucinous EOC have been associated with a variety of cancers. They are: CAMK2B, CECR1, LOC101926907, OGDHL, AZGP1, CREB3L1, CDHR5, NUFIP2, CAPN9, HMGCS2, FOXA1, MAP2, HAGLROS, FANCG, BDP1, PLPPR4, KRT6A and IYD.

Lower expressions of CAMK2B were found in breast cancer, pancreatic adenocarcinoma and kidney renal papillary cell carcinoma (Kim et al., 2011; Luo et al., 2022; Jia et al., 2022). This lower expression was associated with poor survival rates (Luo et al., 2022); high expression in kidney renal papillary cell carcinoma was associated with lower tumour grade and low metastasis (Jia et al., 2022).

CECR1 was found in significantly higher expression levels for breast tumour tissue than normal breast tissue (Aghaei et al., 2010). Increased expression levels were observed between grade III versus grade II and post-menopausal versus pre-menopausal breast cancer (Aghaei et al., 2005). Another study found that elevated levels of CECR1 indicated improved survival rates in patients with breast cancer, lung adenocarcinoma, sarcoma, melanoma, kidney cancer and pancreatic cancer (Wang et al., 2021).

LOC101926907 has been included in a signature to diagnose patients with acute myeloid leukemia along with 15 other probe sets and patient age (Angelakis et al., 2023).

A range of studies have taken place to determine whether OGDHL is associated with various cancers. Common conclusions made in papers were that OGDHL expression was suppressed in tumours due to promoter hypermethylation in cancers such as colorectal cancer and hepatocellular carcinoma (Fedorova et al., 2015; Khalaj-Kondori et al., 2020; Dai et al., 2020). OGDHL was observed to be down-regulated in pancreatic, liver, clear cell renal cell and hepatocellular cancer and this expression was associated with a poor prognosis (Liu et al., 2019a; Jiao et al., 2019; Hu et al., 2019; Dai et al., 2020).

Down-regulation of AZGP1 in advanced breast cancer tissue was found when compared with normal breast tissue; this low expression was associated with a worse prognosis in breast cancer patients (Parris et al., 2014). Furthermore, Stavnes et al. (2013) found that AZGP1 was found in significantly higher expression levels in breast cancer tissue when compared with serous ovarian cancer. Increased expression levels of AZGP1 in HPV-positive oropharyngeal squamous cell carcinoma compared with normal tissue was found to be related to improved survival and recurrence-free survival in patients (Poropatich et al., 2019). Expression was also found to be higher in prostate cancer than normal tissue samples (Cao et al., 2019); both biochemical relapse-free and metastasis-free survival were shorter for low levels of AZGP1 (Zhang et al., 2017a; Jung et al., 2014).

CREB3L1 in high expression was found to have negative association with progression-free survival in anaplastic thyroid carcinoma (Pan et al., 2022). In triple negative breast cancer, CREB3L1 was able to determine the response of these tumours to doxorubicin-base chemotherapy; tumours with a greater sensitivity to this treatment had greater expression levels of CREB3L1 (Denard et al., 2018). It was discussed that up-regulation of CREB3L1 was mostly in low-medium grade tumours whereas repressed expression was found in higher grade tumours in breast cancer (Ward et al., 2016).

High expression levels of CDHR5 in pancreatic ductal adenocarcinoma were associated with shorter overall survival and was an independent prognostic factor (Gao et al., 2020). In contrast, CDHR5 was found to be down-regulated when comparing colorectal tumour tissue with normal tissue and was found in patients with worse prognosis (Losi et al., 2011; Beck et al., 2021); a similar conclusion was made when considering CDHR5 expression in hepatocellular carcinoma (Ding et al., 2020).

Higher expressions of NUFIP2 in upper tract urothelial carcinoma were significantly associated with improved survival (Fu et al., 2022).

CAPN9 was observed in low expression levels in gastric cancer which was associated with poor prognosis and advanced stage (Peng et al., 2016).

Low CAPN9 expression was associated with poor prognosis in breast cancer patients who had received endocrine therapy (Davis et al., 2014).

HMGCS2 was associated with improved prognosis in hepatocellular carcinoma (Tang et al., 2022; Xu et al., 2021). Reduction of this gene in colorectal cancer was found to promote tumour growth through angiogenesis (Zou et al., 2019). Low levels were also found in prostate cancer which led to both shorter disease-free and biochemical recurrence-free survival (Wan et al., 2019).

FOXA1 was highly expressed in hepatocellular carcinoma and gastric cancer which implied poor prognosis in patients (Liu et al., 2022b; Dai et al., 2021).

High expression of MAP2 was described as predictive of good prognosis in patients with glioma (Yi et al., 2018). It is suggested that an increase of MAP2 expression inhibits melanoma development (Song et al., 2010).

Greater expression levels of HAGLROS in nephroblastoma, laryngeal squamous cell carcinoma, diffuse large B-cell lymphoma, lung carcinoma and gastric cancer were associated with tumour progression (Li et al., 2022; Ma et al., 2022; Chen et al., 2018; Shu et al., 2022; Wang et al., 2019b).

Loss of copy numbers of FANCG which can lead to changes in gene expression level was found frequently in head and neck squamous cell carcinoma which may indicate FANCG differential expression in this cancer (Türke et al., 2017).

Expressions of BDP1 were significantly decreased in lymphoma with correlations being apparent between BDP1 expression and clinical outcomes in activated B-cell diffuse large cell carcinoma (Cabarcas-Petroski & Schramm, 2022).

High PLPPR4 expression levels were identified in peritoneal metastasis of gastric cancer cells and associated with poor prognosis (Zang et al., 2020). PLPPR4 has been used to form a three gene signature for predicting prognosis of patients with gastric cancer with high expression indicating high risk of poor prognosis (Wang et al., 2022b).

KRT6A has been identified as up-regulated in cancers such as bladder cancer and lung adenocarcinoma (Chen et al., 2022; Yang et al., 2020). In bladder cancer KRT6A expression was found to promote tumour growth and development (Chen et al., 2022). In lung adenocarcinoma, the high expression levels of KRT6A were associated with aggressive stage and lymph node positive tumours (Yang et al., 2020).

IYD has been found in greater expression in breast cancer tissue and these expression levels were associated with prognosis of these patients (Qian et al., 2023). Over-expression of IYD in hepatocellular carcinoma was described as having a tumour suppressing role (Lu et al., 2020a).

In conclusion: PTGR1, TESMIN, LOC101926907, CREB3L1, CDHR5,

NUFIP2, MINOS1-NBL1///NBL1, CAPN9, HMGCS2, MAP2, ZNF548, ZNF682, PCNX4, ZNF718///ZNF595, PLPPR4, KRT6A and IYD were all genes used within the mucinous gene signature that had limited information on their roles in ovarian cancer. Based on the results of classification using linear SVM, this study suggests that these genes combined with CAMK2B, CECR1, OGDHL, AZGP1, FOXA1, HAGLROS, FANCG and BDP1 could be of interest in further study for diagnosis of mucinous EOC. The prior associations of LOC101926907, CREB3L1, CDHR5, NUFIP2, CAPN9, HMGCS2, MAP2, , PLPPR4, KRT6A and IYD in various other cancers does suggest that it is plausible for these genes to also play roles in ovarian cancer and further research is needed to address this. Of the genes with limited associations to cancers such as PTGR1, TESMIN, MINOS1-NBL1///NBL1, ZNF548, ZNF682, PCNX4 and ZNF718///ZNF595, further consideration for the roles these genes may play in ovarian cancer could be of interest. It may be possible that these genes have roles specific to ovarian cancer.

4.6 Case 3: LGSOC versus non-LGSOC

4.6.1 Classification Results Overview

The signature proposed for distinguishing LGSOC patients from non-LGSOC patients consisted of 18 genes. The genes (probe IDs) included were PWWP2B (227999_at), PRIM1 (205053_at), JCHAIN (212592_at), CCDC85B (204610_s_at), STAT1 (200887_s_at), KIAA0196 (201985_at), IGFBP5 (211959_at), SNHG16///SNORD1C///SNORD1A (224603_at), SHMT2 (214096_s_at), TCF19 (223274_at), GGT5 (205582_s_at), POLA2 (204441_s_at), SP110 (208012_x_at), SLC2A6 (220091_at), RMND5A (212479_s_at), GPBP1L1 (222452_s_at), RICTOR (226310_at) and SNX4 (205329_s_at).

When evaluating this gene signature's ability to distinguish between LGSOC and non-LGSOC patient samples in test data, both the linear and radial SVM model presented classification performances with 100% precision, recall and F_1 scores when samples were predicted to be LGSOC for a probability/score greater than 0.5. Therefore, both models were able to completely distinguish between samples of LGSOC and non-LGSOC histology. With limited research regarding binary classification of LGSOC versus non-LGSOC using genomics, we will consider Qian et al. (2020) as a benchmark. Qian et al. (2020) achieved an F_1 score of approximately 93% in classifying HGSOC versus non-HGSOC using MRI imaging. When comparing the result by Qian

et al. (2020) with this study result for LGSOC versus non-LGSOC, the F_1 score in this study would indicate excellent predictive qualities of the chosen gene signature. Similarly, it would indicate the use of genomics outperforming MRI imaging in EOC subtype prediction. However, the sample size of this testing data may have an affect of this benchmark comparison. With only three LGSOC samples, for further validation of this gene signature a much larger number of LGSOC samples would be needed.

Exploration into the effect of different probability/score threshold values t^* on the classification performance in polynomial SVM, KNN and random forest models took place.

Both polynomial and random forest models achieved 100% test precision, recall and F_1 scores when fluctuating the probability/score threshold that classified samples as LGSOC. This perfect classification of the two groups was found in the polynomial and random forest models for $0.5367 < t^* < 0.9131$ and $0.238 < t^* < 0.38$, respectively.

The KNN model never achieved a perfect classification performance for any probability/score threshold. The best test classification performance for the KNN model was found when samples with probabilities/score greater than or equal to 0.5 were predicted to be LGSOC tumours. One sample of LGSOC had an equal chance of being classified as LGSOC and non-LGSOC histology. Hence, the threshold $t \geq 0.5$ was employed. With this threshold, one non-LGSOC sample was incorrectly identified as LGSOC histology and the model achieved a test precision, recall and F_1 score of 75%, 100% and 85.71%, respectively.

The radial SVM model using 53 support vectors correctly classified all patient samples for both training and testing. In comparison, the linear SVM model misclassified two non-LGSOC samples when training the model and correctly classified all testing samples. Hence, for the linear SVM model it indicates that allowing for misclassifications in the training stage allowed for better generalization of the model for testing. Whereas, even with the use of an increased number of support vectors in the radial SVM model, this did not affect the generalization of the model.

The most commonly misclassified samples were of mucinous histology using this gene signature.

4.6.2 LGSOC Gene Signature

Previous studies discuss the genes JCHAIN, STAT1, IGFBP5, SNHG16, SHMT2, POLA2 and RMND5A and their possible roles or expression in ovarian cancer. JCHAIN's high expression levels in ovarian cancer has been associated with an increased overall survival (Zou et al., 2022). This current

study found that JCHAIN was down-regulated in LGSOC which may indicate that JCHAIN has a negative effect of LGSOC patient survival. High expression of STAT1 was observed in HGSOC (Liu et al., 2020a). This current study indicated that low expression levels of STAT1 are found in LGSOC. Furthermore, a study found that STAT1 levels were found to be greater in patients with platinum-resistant ovarian cancer (Stronach et al., 2011) which may suggest that STAT1 is a gene not related to platinum-resistance in LGSOC patients. Down-regulation of IGFBP5 has been reported in ovarian cancer (Hwang et al., 2016; Tamir et al., 2016) which is in agreement with this current study where IGFBP5 was found to be down-regulated in LGSOC. SNHG16 has previously been reported as significantly highly expressed in ovarian cancer tissue with associations to high grade tumours (Yang et al., 2018c). In contrast, the probe 224603_at representing SNHG16///SNORD1C///SNORD1A was up-regulated in LGSOC when expression was analysed. High SHMT2 expression has also been reported in ovarian cancer when compared with normal tissue (Lee et al., 2014); this current study determines that SHMT2 was significantly up-regulated in LGSOC when compared with normal ovaries. POLA2 has also been proposed as prognostic in serous ovarian cancer and was expressed in low levels (Willis et al., 2016). This current study found the POLA2 was down-regulated in LGSOC samples compared with normal ovary samples indicating that these low levels of POLA2 may be specific to LGSOC. RMND5A has been found with over-expression in ovarian cancer tissue when compared with normal tissue (Li et al., 2008). However, the current study found that RMND5A was down-regulated in LGSOC.

Studies also associated the genes used in the signature for distinguishing between LGSOC and non-LGSOC samples with other cancers as discussed below.

PWWP2B is down-regulated in recurrent gastric cancer patients and has been suggested as predictive of this recurrence (Sohn et al., 2021).

High expression of PRIM1 was correlated with worse prognosis and implied to be independently prognostic in hepatocellular carcinoma (Zhang et al., 2020).

JCHAIN was found to be significantly differentially expressed between patients with varying responses to treatment of B-cell lymphoblastic leukemia with significant association to poor overall survival (Cruz-Rodriguez et al., 2016).

CCDC85B was found in greater expression in non-small cell lung cancer tissue and colorectal cancer than normal tissue (Feng et al., 2019; Wang et al., 2018). It was associated with non-small cell lung cancer progression due to its tumour promotion characteristics (Feng et al., 2019).

High expression of KIAA0196 was observed specifically in poorly differentiated hepatocellular carcinoma tissue (Huang et al., 2017). Increased expression of KIAA0196 was identified in prostate cancer (Porkka et al., 2004).

IGFBP5 was down-regulated in kidney renal papillary carcinoma controlled to non-tumour tissue with low expression being associated with longer survival in patients (Wang et al., 2019a). Over-expression of this gene in both urothelial cancers of the upper urinary tracts and bladder was associated with poor prognosis as well as advanced stage and grade (Liang et al., 2013).

SNHG16 has been observed as up-regulated in colorectal carcinomas, hepatocellular carcinoma, glioma and non-small cell lung cancer with a role of promoting tumour growth (Li et al., 2019; Chen et al., 2019; Yang et al., 2018a; Han et al., 2019).

SHMT2 expression was found higher in lung adenocarcinoma, gastric cancer, esophageal cancer and colorectal cancer than in normal tissue with a lower expression level improving a patients prognosis (Luo et al., 2021; Liu et al., 2019b).

Over-expression of TCF19 was significantly associated with hepatocellular carcinoma tumour growth promotion (Zeng et al., 2019). TCF19 was also over-expressed in colorectal cancer and suggested to lead to an increase in distant metastasis (Du et al., 2020).

GGT5 is also highly expressed in lung adenocarcinoma and had associations with poor prognosis and enhanced drug resistance (Wei et al., 2020).

Reduction of POLA2 in lung cancer cells was associated with gemcitabine resistance (Koh et al., 2016).

RMND5A was highly expressed and associated with poor prognosis in pancreatic adenocarcinoma (Chen et al., 2021b).

A high expression of RICTOR is related to increased tumour growth and development; studies have shown that reduction in this expression inhibits this growth and development (Montero et al., 2012).

SNX4 has previously been included in a signature for predicting the recurrence-free survival of patients with prostate cancer (Cai et al., 2021).

Studies reporting associations with the three genes GPBP1L1, SP110 and SLC2A6 and cancers were limited. With limited studies also relating SP110, SLC2A6 and GPBP1L1 to ovarian cancer, this current study may indicate that further research into these genes is necessary to determine their roles specifically in LGSOC. Furthermore, limited studies associating PWWP2B, KIAA0196, TCF19, GGT5, RICTOR and SNX4 expression with ovarian cancer but abundant studies associating them with various other cancers may suggest that further research is needed to determine the roles of these genes in ovarian cancer. With a number of these genes being related to prognosis and tumour development, these genes could show important association with

LGSOC development or overall prognosis of patients.

4.7 Case 4: Endometrioid versus Clear Cell EOC

4.7.1 Classification Results Overview

For distinguishing between endometrioid and clear cell patient samples, an 11 gene signature was employed. The gene signature contained the genes (probe ID): ANP32A (201051_at), SMYD2 (212922_s_at), WEE1 (212533_at), PALLD (200907_s_at), RPS23 (227722_at), GLCCI1 (227525_at), CEP70 (219036_at), SCGB2A2 (206378_at), PVR (214444_s_at), NR2F2-AS1 (232878_at) and CBX6 (202047_s_at).

Firstly, considering a sample with a probability/score of greater than 0.5 as being predicted to be of endometrioid histology, the polynomial SVM model provided the best test predictive abilities with an F_1 score of 100%. Once again using Qian et al. (2020) as a benchmark for result comparison, this current study's model outperforms the classification performance achieved by Qian et al. (2020). Similarly to both mucinous versus non-mucinous EOC and LGSOC versus non-LGSOC, this could be due to sample size. A larger sample size for both training and testing would be needed to determine whether this classification achievement maintained its improved performance of Qian et al. (2020) classification results.

This study examined the effect of varying the probability/score threshold for predicting a patient's histology as endometrioid EOC for the four other classification models.

Ranges of probability/scores that allowed radial SVM, KNN and random forest models to achieve test F_1 scores of 100% were identified. The ranges of probability/score threshold t^* were $0.1581 < t^* < 0.2015$, $0.091 < t^* < 0.181$ and $0.26 < t^* < 0.453$ for linear SVM, KNN and random forest, respectively.

The linear SVM classification performance could be improved by adopting a different probability/score threshold. However, this improved model did not provide a test F_1 score of 100%. For a threshold range $0.0225 < t^* < 0.0234$, the model performance was improved to provide 85.71% precision and 100% recall. This improved result had one clear cell patients sample being misclassified as endometrioid histology. Interestingly, the misclassified clear cell sample was also the clear cell sample misclassified when differentiating between clear cell EOC and non-clear cell EOC histotype.

Overall, when considering both training and tested model performances, the polynomial SVM has the best classification performance for both cases. The use of 47 support vectors and a non-linear decision boundary of degree 3

provided only one clear cell sample misclassification in the training data and no misclassifications in the test data. This indicates that only one support vector violated both the margin and hyperplane to be misclassified out of all 47 support vectors. Furthermore, the linear SVM model had the worst classification performance on training data and was the only model unable to achieve perfect classification for any threshold. This implies that clear cell and endometrioid samples in both training and test data based on the gene signature could not be linearly separated.

4.7.2 Clear Cell versus Endometrioid Gene Signature

Of the 11 genes considered in this study six, SMYD2, WEE1, PALLD, PVR, NR2F2-AS1 and CBX6 have been previously discussed in ovarian cancer studies. SMYD2 was presented as a possible prognostic marker for clear cell EOC (Engqvist et al., 2020). In addition, its significant up-regulation in HGSOC compared with normal ovaries has been discussed (Kukita et al., 2019). In contrast to previous studies, this current study found that SMYD2 (represented by the probe 212922_s_at), was up-regulated in endometrioid EOC when compared with normal ovaries and was differentially expressed in endometrioid EOC when compared with clear cell EOC, mucinous EOC, HGSOC and LGSOC.

The implementation of WEE1 inhibitors to treat TP53 mutated, platinum-resistant ovarian cancer tumours is currently being trialled (Embaby et al., 2023). This suggests that WEE1 plays a role in the platinum-resistance of OC tumours. However, this study found WEE1 to be highly expressed in endometrioid tumours which is a subtype of EOC not characterized by platinum-resistance which may imply WEE1 has another role in this subtype of EOC.

High expression of PALLD has recently been associated with post-chemotherapy, recurrent HGSOC (Davidson et al., 2020). Due to lack of recurrent tumour representation in this current study, the high expression of PALLD in post-chemotherapy, recurrent HGSOC could not be evaluated. Instead, for assumed primary tumours, PALLD was detected in greater expression levels in endometrioid EOC when compared with normal ovaries and clear cell EOC, mucinous EOC, HGSOC and LGSOC.

PVR has been observed as up-regulated in clear cell EOC when compared with other EOC subtypes (Schwartz et al., 2002). This current study did not find any significant expression differences of PVR in clear cell EOC when compared with mucinous EOC, LGSOC and HGSOC; the EOC subtype found to have significant differential expression of PVR with all other EOC subtypes was endometrioid EOC.

NR2F2-AS1 was recently considered down-regulated in EOC when compared with normal ovaries (Salamini-Montemurri et al., 2023). NR2F2-AS1 was also found to be down-regulated in endometrioid EOC when compared with normal ovaries during this study, which implies slight agreement with previous studies.

A study observed lower expression of CBX6 in ovarian cancer than in normal ovaries (Hu et al., 2022). The current study detected down-regulation of CBX6 in endometrioid EOC compared with normal samples. Hu et al. (2022) found this lower expression when comparing only serous EOC with normal samples which suggests that further research may be needed for other subtypes of EOC.

In addition, ANP32A, SMYD2, WEE1, PALLD, RPS23, CEP70, SCGB2A2, PVR, NR2F2-AS1 and CBX6 all have previous associations with a variety of cancers other than ovarian.

ANP32A was observed to be highly expressed in hepatocellular carcinoma, glioma, colorectal cancer and oral squamous cell carcinoma (Tian et al., 2020; Xie et al., 2018; Yan et al., 2017; Shi et al., 2011; Velmurugan et al., 2016). ANP32A was concluded to be unfavourable in the prognosis of hepatocellular carcinoma (Tian et al., 2020) and associated with progression in glioma, colorectal cancer and oral squamous cell carcinoma (Xie et al., 2018; Yan et al., 2017; Shi et al., 2011; Velmurugan et al., 2016).

SMYD2 was found in high expression levels for breast cancer, acute lymphoblastic leukemia and gastric cancer (Li et al., 2018; Sakamoto et al., 2014; Komatsu et al., 2015). These levels were suggested to promote tumour growth and development in breast cancer (Li et al., 2018) and associated with poor prognosis in acute lymphoblastic leukemia and gastric cancer (Sakamoto et al., 2014; Komatsu et al., 2015). SMYD2 expression levels decreased in acute lymphoblastic leukemia patients who responded to chemotherapy (Sakamoto et al., 2014).

WEE1 inhibition is discussed as potential pathways for treatment of glioblastoma, osteosarcoma and acute myeloid leukemia (Mir et al., 2010; PosthumaDeBoer et al., 2011; Porter et al., 2012). In each of the three cancers, WEE1 is over-expressed and its inhibition may lead to apoptosis of cancer cells (Mir et al., 2010; PosthumaDeBoer et al., 2011; Porter et al., 2012).

PALLD was found in significantly lower expression levels for pancreatic ductal adenocarcinoma patients treated with chemotherapy first when compared with patients treated with surgery first (Sato et al., 2016). Similarly, PALLD has been studied in cutaneous melanoma treatment; significant differences in PALLD expression was found when comparing patients pre- and post-treatment (de Lima et al., 2013).

Over-expression of RPS23 was found in all stages of colorectal carcinoma (Lau et al., 2014).

CEP70 was detected in higher expression levels for breast cancer and pancreatic cancer compared with normal tissue and was suggested to promote tumour growth and development (Shi et al., 2017; Xie et al., 2016).

SCGB2A2 has been associated with the promotion of cancer cell growth and development (Picot et al., 2016).

High expression of PVR has been observed in pancreatic and prostate cancer (Nishiwada et al., 2015; Papanicolau-Sengos et al., 2019). This high expression in pancreatic cancer led to poorer post-operative prognosis and had independent prognostic value (Nishiwada et al., 2015).

Expression of NR2F2-AS1 vary dependent on cancer type. For example, it was observed as down-regulated in oral squamous cell carcinoma (Liang et al., 2022). Whereas, up-regulation of NR2F2-AS1 was observed in osteosarcoma and prostate cancer (Ye et al., 2022a; Fu et al., 2020). High levels in osteosarcoma were correlated with worse clinical stage and found to promote tumour growth (Ye et al., 2022a).

CBX6 was detected with high expression in hepatocellular carcinoma and was concluded to be unfavourable in patient prognosis, with high CBX6 expression significantly associated with larger tumour size (Zheng et al., 2017). Multiple studies found down-regulation of CBX6 in breast cancer tissue, where over-expression of this gene was suggested to improve prognosis and inhibit tumour development (Deng et al., 2019; Li et al., 2020; Liang et al., 2017).

Finally the gene GLCCI1 investigated here is not usually associated with cancer but is associated with asthma, its severity and response to inhaled corticosteroid treatments (Jiang et al., 2021b; Hu et al., 2016; Xun et al., 2019).

Thus, the vast majority of genes selected for the signature used to distinguish between clear cell and endometrioid EOC have previously been associated with various cancers. This provides foundation for their possible roles in ovarian cancer and possible diagnostic qualities. However, GLCCI1 has limited discussion on its association with cancer and further research may be needed.

4.8 Limitations of Study

This current study does not go without limitations. Firstly, the training data used within this study consisted of 239 samples in which imbalance was high across sample groups. Furthermore, the test data contained 89

samples also containing high imbalance between groups. To address this imbalance, stratified k-fold cross validation, precision-recall curves, metrics such as precision, recall, F_1 score, Fowlkes Mallows index and Jaccard index were used. An increase in sample size may provide further insights into the differences in expression between varying probes when comparing subtypes of EOC and normal ovary samples.

In addition, the results from each of the four classifications groups have suggested promising abilities of the selected gene (probe) signatures to distinguish between subtypes of epithelial ovarian cancer. However, results from tested models for classification will need much greater sample sizes to improve reliability of results. Particularly in cases such as classifying LGSOC versus non-LGSOC sample. Currently, the test data only contains three LGSOC samples compared with 86 non-LGSOC samples.

Furthermore, this study is based on the use of statistical and machine learning methodology in selecting appropriate genes/probes for classification of ovarian cancer samples. A limitation includes the lack of immunohistochemical analysis to determine whether differentially expressed genes that code for proteins indicate a change in protein levels. Furthermore, analysis of the co-expression of genes/probes was not evaluated within this study.

When considering the classification model training and parameter selection through stratified repeated 10-fold cross validation, the precision-recall area under the curve was selected as the metric to maximise for parameter tuning. Although this parameter tuning provided models that achieved good classification performances, the precision-recall area under the curve is less interpretative than the standard ROC-AUC due to the curve not having to start at 100% precision and 0% recall. Therefore, a better choice for parameter tuning may have been to find the parameter that maximised the F_1 score, as a balance of precision and recall was optimal in this study.

Another limitation is time. An increased time period for obtaining secondary data and data analysis could have allowed for greater sample size in both training and testing data. Additionally, time to obtain data on survival of patients diagnosed with these subtypes of ovarian cancer may have provided insight into the genes/probes selected in this study and whether they have an affect on the prognosis of patients with corresponding EOC histotypes.

4.9 Future Research

Based on a limitation discussed previously, an area of future work would be to collect further data to analyse and determine whether the gene (probe)

signatures proposed in this study maintain their ability to distinguish between epithelial ovarian cancer subtypes. The retrieval of survival data for patients with ovarian cancer would also be interesting to determine whether genes (probes) in these signatures could be associated with patient prognosis.

Another area for further research would be to determine whether gene signatures can be identified that can distinguish ovarian cancer patients by their stage of development.

It could be worthwhile to study particular patient samples of epithelial ovarian cancer that were frequently misclassified and determine whether substantial differences from other samples of the same diagnosed histology exist.

Chapter 5

Conclusion

Two questions were posed during this study. Firstly, can genes be used to detect the presence of ovarian cancer? If so, can subsets of these genes be used to diagnose a patient's EOC subtype? The results indicate that for classification of both mucinous versus non-mucinous EOC patient samples and LGSOC versus non-LGSOC patient samples, the selected gene/probe signatures were able to completely distinguish between the classes. The model with perfect classification ability for mucinous versus non-mucinous EOC samples was a linear SVM model with 15 support vectors determining the position of the decision boundary. It is concluded that the mucinous and non-mucinous EOC patient samples can be linearly separated with respect to the provided gene signature. Whereas, the classification model able to perfectly distinguish between LGSOC and non-LGSOC patient samples was a radial SVM model using 53 support vectors.

For classification of clear cell EOC versus non-clear cell EOC patient samples, no training model had the ability to completely distinguish between the two classes. The best performing model for classifying the training data was the radial SVM model using 112 support vectors. However, this large number of support vectors may indicate over-fitting of the data to the training model as the radial SVM model is the worst performing model when testing data is applied. The best performing model for a standard threshold $t > 0.5$ was the random forest model which misclassified one clear cell sample. This random forest model was built from 500 decision trees and randomly selected one of the 15 genes at each decision node split. The classification ability of this model is comparable to a prior study that uses a much larger number of genes for binary classification. Due to aims of reducing diagnosis time, a smaller gene signature used for diagnosis is preferable.

No training models were able to perfectly differentiate between clear cell and endometrioid EOC patient samples. The model with the best classifi-

cation performance for the training data was the polynomial SVM of degree 3, using 47 support vectors. For training data, this polynomial SVM model only misclassified one endometrioid sample. However, this polynomial SVM model was able to perfectly classify the test clear cell and endometrioid samples at a threshold $t > 0.5$. Therefore, the gene signature presented for classifying these EOC subtypes indicates an ability to distinguish the two.

Previous studies have combined the diagnosis of endometrioid and clear cell EOC due to their association with endometrioisis. However, endometrioid has a better response to platinum-based chemotherapy and a greater survival rate compared with clear cell. Hence, it is vital to differentiate between the two subtypes in order to provide treatments to patients that they will have the greatest chance of response to. In addition, it is noted that endometrioid is difficult to distinguish from other EOC subtypes. So this study provides further genes that could be utilised in distinguishing endometrioid EOC from other EOC subtypes, specifically clear cell EOC.

This study has provided three separate gene signatures that indicate diagnostic abilities when considering EOC histological subtypes clear cell, mucinous and LGSOC, respectively. A fourth gene signature indicates the ability to distinguish between clear cell and endometrioid EOC tumours. Further research into these four gene signatures is needed as they could provide a solution to current issues with the accuracy of EOC subtype diagnosis. Moreover, the three gene signatures with possible diagnostic capabilities indicated unique expression patterns in their corresponding EOC subtype. Genes with unique relationships with specific EOC subtypes may not only be valuable in diagnosis but for possible treatment paths.

The incentive for obtaining gene signatures with diagnostic ability in EOC histological subtypes is to improve the overall accuracy of diagnosis. EOC histotypes vary in features such as prognosis, common diagnosis stage and treatment response. Therefore, differing treatments should be provided based on a patient's EOC subtype diagnosis to prevent the current problems with platinum-resistance. With treatment provided that the patient's tumour is most susceptible to, an improvement in treatment response rates would be the result. Improvements in treatment response in turn would improve the prognosis of patients. In addition, the reduction in need for invasive diagnostic testing could reduce the time to diagnosis. Hence, the tumour could be diagnosed earlier, further improving patient prognosis.

Chapter 6

References

Abdulhadi, N., & Al-Mousa, A. (2021). Diabetes detection using machine learning classification methods. In *2021 International Conference on Information Technology (ICIT)* (pp. 350–354). doi:10.1109/ICIT52682.2021.9491788.

Abera, F. Z., & Khedkar, V. (2020). Machine learning approach electric appliance consumption and peak demand forecasting of residential customers using smart meter data. *Wireless Personal Communications*, *111*, 65–82. doi:10.1007/s11277-019-06845-6.

Abiko, K., Mandai, M., Hamanishi, J., Yoshioka, Y., Matsumura, N., Baba, T., Yamaguchi, K., Murakami, R., Yamamoto, A., Kharma, B. et al. (2013). Pd-l1 on tumor cells is induced in ascites and promotes peritoneal dissemination of ovarian cancer through ctl dysfunction. *Clinical cancer research*, *19*, 1363–1374. doi:10.1158/1078-0432.CCR-12-2199.

Acharya, U. R., Fernandes, S. L., WeiKoh, J. E., Ciaccio, E. J., Fabell, M. K. M., Tanik, U. J., Rajinikanth, V., & Yeong, C. H. (2019). Automated detection of alzheimer's disease using brain mri images– a study with various feature extraction techniques. *Journal of Medical Systems*, *43*, 302. doi:10.1007/s10916-019-1428-9.

Acharya, U. R., Sree, S. V., Kulshreshtha, S., Molinari, F., Koh, J. E. W., Saba, L., & Suri, J. S. (2014). Gynescan: An improved online paradigm for screening of ovarian cancer via tissue characterization. *Technology in Cancer Research & Treatment*, *13*, 529–539. doi:10.7785/tcrtexpress.2013.600273.

Aggarwal, C. C. (2014a). Decision trees: Theory and algorithms. In *Data*

classification: algorithms and applications (pp. 87–115). CRC Press. (1st ed.). doi:10.1201/b17320.

Aggarwal, C. C. (2014b). Distance metric learning for data classification. In *Data classification: algorithms and applications* (pp. 472–473). CRC Press. (1st ed.). doi:10.1201/b17320.

Aggarwal, C. C. (2014c). Probabilistic models for classification. In *Data classification: algorithms and applications* (pp. 69–70). CRC Press. (1st ed.). doi:10.1201/b17320.

Aggarwal, C. C. (2014d). Probabilistic models for classification. In *Data classification: algorithms and applications* (pp. 73–76). CRC Press. (1st ed.). doi:10.1201/b17320.

Aghaei, M., Karami-Tehrani, F., Salami, S., & Atri, M. (2005). Adenosine deaminase activity in the serum and malignant tumors of breast cancer: the assessment of isoenzyme ada1 and ada2 activities. *Clinical biochemistry*, *38*, 887–891. doi:10.1016/j.clinbiochem.2005.05.015.

Aghaei, M., Karami-Tehrani, F., Salami, S., & Atri, M. (2010). Diagnostic value of adenosine deaminase activity in benign and malignant breast tumors. *Archives of Medical Research*, *41*, 14–18. doi:10.1016/j.arcmed.2009.10.012.

Aghajanian, C., Blank, S. V., Goff, B. A., Judson, P. L., Teneriello, M. G., Husain, A., Sovak, M. A., Yi, J., & Nycum, L. R. (2012). Oceans: A randomized, double-blind, placebo-controlled phase iii trial of chemotherapy with or without bevacizumab in patients with platinum-sensitive recurrent epithelial ovarian, primary peritoneal, or fallopian tube cancer. *Journal of Clinical Oncology*, *30*, 2039–2045. doi:10.1200/JCO.2012.42.0505.

Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R. B., Aly, D. M., Almgren, P., Wessman, Y., Shaat, N., Spégel, P., Mulder, H., Lindholm, E., Melander, O., Hansson, O., Malmqvist, U., Lernmark, Å., Lahti, K., Forsén, T., Tuomi, T., Rosengren, A. H., & Groop, L. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes & Endocrinology*, *6*, 361–369. doi:10.1016/S2213-8587(18)30051-2.

Ahmed, A. A., Etemadmoghadam, D., Temple, J., Lynch, A. G., Riad, M., Sharma, R., Stewart, C., Fereday, S., Caldas, C., deFazio, A., Bowtell,

D., & Brenton, J. D. (2010). Driver mutations in tp53 are ubiquitous in high grade serous carcinoma of the ovary. *The Journal of Pathology*, *221*, 49–56. doi:https://doi.org/10.1002/path.2696.

Alcázar, J. L., Utrilla-Layna, J., Mínguez, J. A., & Jurado, M. (2013). Clinical and ultrasound features of type i and type ii epithelial ovarian cancer. *International Journal of Gynecologic Cancer*, *23*, 680–684. doi:10.1097/IGC.0b013e31828bdbb6.

Alexandre, J., Ray-Coquard, I., Selle, F., Floquet, A., Cottu, P., Weber, B., Falandry, C., Lebrun, D., & Pujade-Lauraine, E. (2010). Mucinous advanced epithelial ovarian carcinoma: clinical presentation and sensitivity to platinum-paclitaxel-based chemotherapy, the gineco experience. *Annals of Oncology*, *21*, 2377–2381. doi:10.1093/annonc/mdq257.

Alhogail, A., & Alsabih, A. (2021). Applying machine learning and natural language processing to detect phishing email. *Computers & Security*, *110*, 102414. doi:10.1016/j.cose.2021.102414.

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672. doi:https://doi.org/10.1016/j.compbiomed.2021.104672.

Ali, R. H., Kalloger, S. E., Santos, J. L., Swenerton, K. D., & Gilks, C. B. (2013). Stage ii to iv low-grade serous carcinoma of the ovary is associated with a poor prognosis: A clinicopathologic study of 32 patients from a population-based tumor registry. *International Journal of Gynecological Pathology*, *32*, 529–535. doi:10.1097/PGP.0b013e31827630eb.

Alsop, K., Fereday, S., Meldrum, C., DeFazio, A., Emmanuel, C., George, J., Dobrovic, A., Birrer, M. J., Webb, P. M., Stewart, C., Friedlander, M., Fox, S., Bowtell, D., & Mitchell, G. (2012a). Author correction. *Journal of Clinical Oncology*, *30*, 4180–4180. doi:10.1200/JCO.2012.47.3777.

Alsop, K., Fereday, S., Meldrum, C., DeFazio, A., Emmanuel, C., George, J., Dobrovic, A., Birrer, M. J., Webb, P. M., Stewart, C., Friedlander, M., Fox, S., Bowtell, D., & Mitchell, G. (2012b). Brca mutation frequency and patterns of treatment response in brca mutation-positive women with ovarian cancer: A report from the australian ovarian cancer study group. *Journal of Clinical Oncology*, *30*, 2654–2663. doi:10.1200/JCO.2011.39.8545.

Altaf, T., Anwar, S. M., Gul, N., Majeed, M. N., & Majid, M. (2018). Multi-class alzheimer's disease classification using image and clinical features. *Biomedical Signal Processing and Control*, *43*, 64–74. doi:10.1016/j.bspc.2018.02.019.

Altman, D. G., & Bland, J. M. (1994a). Statistics notes: Diagnostic test 2: predictive values. *BMJ*, *309*, 102. doi:10.1136/bmj.309.6947.102.

Altman, D. G., & Bland, J. M. (1994b). Statistics notes: Diagnostic tests 1: sensitivity and specificity. *BMJ*, *308*, 1552. doi:10.1136/bmj.308.6943.1552.

Amrane, M., Oukid, S., Gagaoua, I., & Ensarİ, T. (2018). Breast cancer classification using machine learning. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* (pp. 1–4). doi:10.1109/EBBT.2018.8391453.

An, H., Wang, Y., Wong, E. M. F., Lyu, S., Han, L., Perucho, J. A. U., Cao, P., & Lee, E. Y. P. (2021). Ct texture analysis in histological classification of epithelial ovarian carcinoma. *European Radiology*, *31*, 5050–5058. doi:10.1007/s00330-020-07565-3.

Angelakis, A., Soulioti, I., & Filippakis, M. (2023). Diagnosis of acute myeloid leukaemia on microarray gene expression data using categorical gradient boosted trees. *Heliyon*, *9*, e20530. doi:10.1016/j.heliyon.2023.e20530.

Anglesio, M. S., Kommoss, S., Tolcher, M. C., Clarke, B., Galletta, L., Porter, H., Damaraju, S., Fereday, S., Winterhoff, B. J., Kalloger, S. E., Senz, J., Yang, W., Steed, H., Allo, G., Ferguson, S., Shaw, P., Teoman, A., Garcia, J. J., Schoolmeester, J. K., Bakkum-Gamez, J., Tinker, A. V., Bowtell, D. D., Huntsman, D. G., Gilks, C. B., & McAlpine, J. N. (2013). Molecular characterization of mucinous ovarian tumours supports a stratified treatment approach with her2 targeting in 19% of carcinomas. *The Journal of Pathology*, *229*, 111–120. doi:https://doi.org/10.1002/path.4088.

Antunes, A., Andrade-Campos, A., Sardinha-Lourenço, A., & Oliveira, M. S. (2018). Short-term water demand forecasting using machine learning techniques. *Journal of Hydroinformatics*, *20*, 1343–1366. doi:10.2166/hydro.2018.163.

Aramendía-Vidaurreta, V., Cabeza, R., Villanueva, A., Navallas, J., & Alcázar, J. L. (2016). Ultrasound image discrimination between benign and malignant adnexal masses based on a neural network approach. *Ultrasound*

in Medicine & Biology, *42*, 742–752. doi:10.1016/j.ultrasmedbio.2015.11.014.

Arend, R. C., Londoño, A. I., Montgomery, A. M., Smith, H. J., Dobbin, Z. C., Katre, A. A., Martinez, A., Yang, E. S., Alvarez, R. D., Huh, W. K., Bevis, K. S., Straughn, J., J. Michael, Estes, J. M., Novak, L., Crossman, D. K., Cooper, S. J., Landen, C. N., & Leath, I., Charles A. (2018). Molecular Response to Neoadjuvant Chemotherapy in High-Grade Serous Ovarian Carcinoma. *Molecular Cancer Research*, *16*, 813–824. doi:10.1158/1541-7786.MCR-17-0594.

Arezzo, F., Cormio, G., La Forgia, D., Santarsiero, C. M., Mongelli, M., Lombardi, C., Cazzato, G., Cicinelli, E., & Loizzi, V. (2022). A machine learning approach applied to gynecological ultrasound to predict progression-free survival in ovarian cancer patients. *Archives of Gynecology and Obstetrics*, *306*, 2143–2154. doi:10.1007/s00404-022-06578-1.

Arya, M., Mittal, N., & Singh, G. (2018). Texture-based feature extraction of smear images for the detection of cervical cancer. *IET Computer Vision*, *12*, 1049–1059. doi:10.1049/iet-cvi.2018.5349.

Assem, H., Rambau, P. F., Lee, S., Ogilvie, T., Sienko, A., Kelemen, L. E., & Köbel, M. (2018). High-grade endometrioid carcinoma of the ovary. *The American Journal of Surgical Pathology*, *42*, 534–544. doi:10.1097/PAS.0000000000001016.

Ataseven, B., Grimm, C., Harter, P., Heitz, F., Traut, A., Prader, S., & du Bois, A. (2016). Prognostic impact of debulking surgery and residual tumor in patients with epithelial ovarian cancer figo stage iv. *Gynecologic Oncology*, *140*, 215–220. doi:10.1016/j.ygyno.2015.12.007.

Audeh, M. W., Carmichael, J., Penson, R. T., Friedlander, M., Powell, B., Bell-McGuinn, K. M., Scott, C., Weitzel, J. N., Oaknin, A., Loman, N., Lu, K., Schmutzler, R. K., Matulonis, U., Wickens, M., & Tutt, A. (2010). Oral poly(adp-ribose) polymerase inhibitor olaparib in patients with brca1 or brca2 mutations and recurrent ovarian cancer: a proof-of-concept trial. *The Lancet*, *376*, 245–251. doi:10.1016/S0140-6736(10)60893-8.

Awad, M., & Khanna, R. (2015). Chapter 3. support vector machines for classification. In *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 39–50). doi:10.1007/978-1-4302-5990-9.

Bamias, A., Psaltopoulou, T., Sotiropoulou, M., Haidopoulos, D., Lianos, E., Bournakis, E., Papadimitriou, C., Rodolakis, A., Vlahos, G., & Dimopoulos, M. A. (2010). Mucinous but not clear cell histology is associated with inferior survival in patients with advanced stage ovarian carcinoma treated with platinum-paclitaxel chemotherapy. *Cancer*, *116*, 1462–1468. doi:10.1002/cncr.24915.

Bao, M., Zhang, L., & Hu, Y. (2020). Novel gene signatures for prognosis prediction in ovarian cancer. *Journal of Cellular and Molecular Medicine*, *24*, 9972–9984. doi:10.1111/jcmm.15601.

Barnabas, G. D., Bahar-Shany, K., Sapoznik, S., Helpman, L., Kadan, Y., Beiner, M., Weitzner, O., Arbib, N., Korach, J., Perri, T., Katz, G., Blecher, A., Brandt, B., Friedman, E., Stockheim, D., Jakobson-Setton, A., Eitan, R., Armon, S., Brand, H., Zadok, O., Aviel-Ronen, S., Harel, M., Geiger, T., & Levanon, K. (2019). Microvesicle proteomic profiling of uterine liquid biopsy for ovarian cancer early detection *. *Molecular & Cellular Proteomics*, *18*, 865–875. doi:10.1074/mcp.RA119.001362.

Beck, M., Baranger, M., Moufok-Sadoun, A., Bersuder, E., Hinkel, I., Mellitzer, G., Martin, E., Marisa, L., Duluc, I., De Reynies, A., & et al. (2021). The atypical cadherin mucdhl antagonizes colon cancer formation and inhibits oncogenic signaling through multiple mechanisms. *Oncogene*, *40*, 522–535. doi:10.1038/s41388-020-01546-y.

Ben Azzouz, F., Michel, B., Lasla, H., Gouraud, W., François, A.-F., Girka, F., Lecointre, T., Guérin-Charbonnel, C., Juin, P. P., Campone, M., & Jézéquel, P. (2021). Development of an absolute assignment predictor for triple-negative breast cancer subtyping using machine learning approaches. *Computers in Biology and Medicine*, *129*, 104171. doi:10.1016/j.compbiomed.2020.104171.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., & Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLOS Computational Biology*, *4*, e1000173. doi:10.1371/journal.pcbi.1000173.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*, 289–300. URL: http://www.jstor.org/stable/2346101.

BenTaieb, A., Nosrati, M. S., Li-Chang, H., Huntsman, D., & Hamarneh, G. (2016). Clinically-inspired automatic classification of ovarian carcinoma subtypes. *Journal of Pathology Informatics*, *7*, 28. doi:10.4103/2153-3539.186899.

Berek, J. S., Renz, M., Kehoe, S., Kumar, L., & Friedlander, M. (2021). Cancer of the ovary, fallopian tube, and peritoneum: 2021 update. *International Journal of Gynecology & Obstetrics*, *155*, 61–85. doi:10.1002/ijgo.13878.

Berman, M., Triki, A. R., & Blaschko, M. B. (2018). The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4413–4421). doi:10.1109/CVPR.2018.00464.

Berrar, D. (2019). Cross-validation. In *Encyclopedia of Bioinformatics and Computational Biology.* (pp. 542–545). Oxford: Academic Press. volume 1. doi:10.1016/B978-0-12-809633-8.20349-X.

Berry, D. A., Cronin, K. A., Plevritis, S. K., Fryback, D. G., Clarke, L., Zelen, M., Mandelblatt, J. S., Yakovlev, A. Y., Habbema, J. D. F., & Feuer, E. J. (2005). Effect of screening and adjuvant therapy on mortality from breast cancer. *New England Journal of Medicine*, *353*, 1784–1792. doi:10.1056/NEJMoa050518.

Berthier, A., Seguin, S., Sasco, A. J., Bobin, J. Y., De Laroche, G., Datchary, J., Saez, S., Rodriguez-Lafrasse, C., Tolle, F., Fraichard, A., Boyer-Guittaut, M., Jouvenot, M., Delage-Mourroux, R., & Descotes, F. (2010). High expression of gabarapl1 is associated with a better outcome for patients with lymph node-positive breast cancer. *British Journal of Cancer*, *102*, 1024–1031. doi:10.1038/sj.bjc.6605568.

Bian, C., Yao, K., Li, L., Yi, T., & Zhao, X. (2016). Primary debulking surgery vs. neoadjuvant chemotherapy followed by interval debulking surgery for patients with advanced ovarian cancer. *Archives of Gynecology and Obstetrics*, *293*, 163–168. doi:10.1007/s00404-015-3813-z.

Bircan, H. A., Gurbuz, N., Pataer, A., Caner, A., Kahraman, N., Bayraktar, E., Bayraktar, R., Erdogan, M. A., Kabil, N., & Ozpolat, B. (2018). Elongation factor-2 kinase (eef-2k) expression is associated with poor patient survival and promotes proliferation, invasion and tumor growth of lung cancer. *Lung cancer*, *124*, 31–39. doi:10.1016/j.lungcan.2018.07.027.

Birjais, R., Mourya, A. K., Chauhan, R., & Kaur, H. (2019). Prediction and diagnosis of future diabetes risk: a machine learning approach. *SN Applied Sciences*, *1*, 1112. doi:10.1007/s42452-019-1117-9.

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms k-nearest-neighbor, support vector machine, random forest and neural network: A review. *Journal of Data Analysis and Information Processing*, *8*, 341–357. doi:10.4236/jdaip.2020.84020.

Bodurka, D. C., Deavers, M. T., Tian, C., Sun, C. C., Malpica, A., Coleman, R. L., Lu, K. H., Sood, A. K., Birrer, M. J., Ozols, R., Baergen, R., Emerson, R. E., Steinhoff, M., Behmaram, B., Rasty, G., & Gershenson, D. M. (2012). Reclassification of serous ovarian carcinoma by a 2-tier system. *Cancer*, *118*, 3087–3094. doi:10.1002/cncr.26618.

Bogani, G., Ditto, A., Lopez, S., Bertolina, F., Murgia, F., Pinelli, C., Chiappa, V., & Raspagliesi, F. (2020). Adjuvant chemotherapy vs. observation in stage i clear cell ovarian carcinoma: A systematic review and meta-analysis. *Gynecologic Oncology*, *157*, 293–298. doi:10.1016/j.ygyno.2019.12.045.

du Bois, A., Reuss, A., Pujade-Lauraine, E., Harter, P., Ray-Coquard, I., & Pfisterer, J. (2009). Role of surgical outcome as prognostic factor in advanced epithelial ovarian cancer: A combined exploratory analysis of 3 prospectively randomized phase 3 multicenter trials. *Cancer*, *115*, 1234–1244. doi:10.1002/cncr.24149.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, *19*, 185–193. doi:10.1093/bioinformatics/19.2.185.

Bolton, K. L., Chen, D., Corona de la Fuente, R., Fu, Z., Murali, R., Köbel, M., Tazi, Y., Cunningham, J. M., Chan, I. C. C., Wiley, B. J., Moukarzel, L. A., Winham, S. J., Armasu, S. M., Lester, J., Elishaev, E., Laslavic, A., Kennedy, C. J., Piskorz, A., Sekowska, M., Brand, A. H., Chiew, Y.-E., Pharoah, P., Elias, K. M., Drapkin, R., Churchman, M., Gourley, C., DeFazio, A., Karlan, B., Brenton, J. D., Weigelt, B., Anglesio, M. S., Huntsman, D., Gayther, S., Konner, J., Modugno, F., Lawrenson, K., Goode, E. L., & Papaemmanuil, E. (2022). Molecular Subclasses of Clear Cell Ovarian Carcinoma and Their Impact on Disease Behavior and Outcomes. *Clinical Cancer Research*, *28*, 4947–4956. doi:10.1158/1078-0432.CCR-21-3817.

Booth, A. L., Abels, E., & McCaffrey, P. (2021). Development of a prognostic model for mortality in covid-19 infection using machine learning. *Modern Pathology*, *34*, 522–531. doi:10.1038/s41379-020-00700-x.

Boylan, K. L. M., Geschwind, K., Koopmeiners, J. S., Geller, M. A., Starr, T. K., & Skubitz, A. P. N. (2017). A multiplex platform for the identification of ovarian cancer biomarkers. *Clinical Proteomics*, *14*, 34. doi:10.1186/s12014-017-9169-6.

Boyle, B. H. (2011). Support vector machines in medical classification tasks. In *Support vector machines: data analysis, machine learning and applications* (pp. 82–84). New York: Nova Science Publishers.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. doi:10.1023/a:1010933404324.

Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., & Cabitza, F. (2020). Detection of covid-19 infection from routine blood exams with machine learning: A feasibility study. *Journal of Medical Systems*, *44*, 135. doi:10.1007/s10916-020-01597-4.

Brown, J., & Frumovitz, M. (2014). Mucinous tumors of the ovary: Current thoughts on diagnosis and management. *Current Oncology Reports*, *16*, 389. doi:10.1007/s11912-014-0389-x.

Burger, R. A., Brady, M. F., Bookman, M. A., Fleming, G. F., Monk, B. J., Huang, H., Mannel, R. S., Homesley, H. D., Fowler, J., Greer, B. E., Boente, M., Birrer, M. J., & Liang, S. X. (2011). Incorporation of bevacizumab in the primary treatment of ovarian cancer. *New England Journal of Medicine*, *365*, 2473–2483. doi:10.1056/NEJMoa1104390.

Buttarelli, M., Ciucci, A., Palluzzi, F., Raspaglio, G., Marchetti, C., Perrone, E., Minucci, A., Giacò, L., Fagotti, A., Scambia, G., & Gallo, D. (2022). Identification of a novel gene signature predicting response to first-line chemotherapy in brca wild-type high-grade serous ovarian cancer patients. *Journal of Experimental & Clinical Cancer Research*, *41*, 50. doi:10.1186/s13046-022-02265-w.

Cabarcas-Petroski, S., Olshefsky, G., & Schramm, L. (2023). Bdp1 as a biomarker in serous ovarian cancer. *Cancer Medicine*, *12*, 6401–6418. doi:10.1002/cam4.5388.

Cabarcas-Petroski, S., & Schramm, L. (2022). Bdp1 expression correlates with clinical outcomes in activated b-cell diffuse large

b-cell lymphoma. *BioMedInformatics*, *2*, 169–183. doi:10.3390/biomedinformatics2010011.

Cai, J., Yang, F., Chen, X., Huang, H., & Miao, B. (2021). Signature panel of 11 methylated mrnas and 3 methylated lncrnas for prediction of recurrence-free survival in prostate cancer patients. *Pharmacogenomics and Personalized Medicine*, *14*, 797–811. doi:10.2147/PGPM.S312024.

Campbell, I. G., Russell, S. E., Choong, D. Y. H., Montgomery, K. G., Ciavarella, M. L., Hooi, C. S. F., Cristiano, B. E., Pearson, R. B., & Phillips, W. A. (2004). Mutation of the PIK3CA Gene in Ovarian and Breast Cancer. *Cancer Research*, *64*, 7678–7681. doi:10.1158/0008-5472.CAN-04-2933.

del Campo, J. M., Matulonis, U. A., Malander, S., Provencher, D., Mahner, S., Follana, P., Waters, J., Berek, J. S., Woie, K., Oza, A. M., Canzler, U., Gil-Martin, M., Lesoin, A., Monk, B. J., Lund, B., Gilbert, L., Wenham, R. M., Benigno, B., Arora, S., Hazard, S. J., & Mirza, M. R. (2019). Niraparib maintenance therapy in patients with recurrent ovarian cancer after a partial response to the last platinum-based chemotherapy in the engot-ov16/nova trial. *Journal of Clinical Oncology*, *37*, 2968–2973. doi:10.1200/JCO.18.02238.

Cancer Genome Atlas Research Network et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, *474*, 609–615. doi:10.1038/nature10166.

Cancer Research UK (2022). Survival and incidence by stage at diagnosis. https://crukcancerintelligence.shinyapps.io/EarlyDiagnosis/. Accessed: 07-02-23.

Cancer Research UK (n.d.a). Grades of ovarian cancer. https://www.cancerresearchuk.org/about-cancer/ovarian-cancer/stages-grades/about-stages-and-grades. Accessed: 07-10-22.

Cancer Research UK (n.d.b). Ovarian cancer statistics: incidence. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/ovarian-cancer#heading-Zero. Accessed: 07-10-22.

Cao, R., Ke, M., Wu, Q., Tian, Q., Liu, L., Dai, Z., Lu, S., & Liu, P. (2019). Azgp1 is androgen responsive and involved in ar-induced prostate cancer cell proliferation and metastasis. *Journal of cellular physiology*, *234*, 17444–17458. doi:10.1002/jcp.28366.

Chan, J. K., Brady, M. F., Penson, R. T., Huang, H., Birrer, M. J., Walker, J. L., DiSilvestro, P. A., Rubin, S. C., Martin, L. P., Davidson, S. A., Huh, W. K., O'Malley, D. M., Boente, M. P., Michael, H., & Monk, B. J. (2016). Weekly vs. every-3-week paclitaxel and carboplatin for ovarian cancer. *New England Journal of Medicine*, *374*, 738–748. doi:10.1056/NEJMoa1505067.

Chang, K.-L., Lee, M.-Y., Chao, W.-R., & Han, C.-P. (2016). The status of her2 amplification and kras mutations in mucinous ovarian carcinoma. *Human Genomics*, *10*, 40. doi:10.1186/s40246-016-0096-9.

Chang, S.-J., Bristow, R. E., & Ryu, H.-S. (2012). Impact of complete cytoreduction leaving no gross residual disease associated with radical cytoreductive surgical procedures on survival in advanced ovarian cancer. *Annals of Surgical Oncology*, *19*, 4059–4067. doi:10.1245/s10434-012-2446-8.

Chatfield, C., & Collins, A. J. (1980). Principal component analysis. In *Introduction to multivariate analysis* Science paperbacks (pp. 58–59). Chapman & Hall. (1st ed.).

Cheasley, D., Nigam, A., Zethoven, M., Hunter, S., Etemadmoghadam, D., Semple, T., Allan, P., Carey, M. S., Fernandez, M. L., Dawson, A., Martin, K., Huntsman, D. G., Le Page, C., Mes-Masson, A.-M., Provencher, D., Hacker, N., Gao, Y., Bowtell, D., deFazio, A., Gorringe, K. L., & Campbell, I. G. (2021). Genomic analysis of low-grade serous ovarian carcinoma to identify key drivers and therapeutic vulnerabilities. *The Journal of Pathology*, *253*, 41–54. doi:10.1002/path.5545.

Cheasley, D., Wakefield, M. J., Ryland, G. L., Allan, P. E., Alsop, K., Amarasinghe, K. C., Ananda, S., Anglesio, M. S., Au-Yeung, G., Böhm, M., Bowtell, D. D. L., Brand, A., Chenevix-Trench, G., Christie, M., Chiew, Y.-E., Churchman, M., DeFazio, A., Demeo, R., Dudley, R., Fairweather, N., Fedele, C. G., Fereday, S., Fox, S. B., Gilks, C. B., Gourley, C., Hacker, N. F., Hadley, A. M., Hendley, J., Ho, G.-Y., Hughes, S., Hunstman, D. G., Hunter, S. M., Jobling, T. W., Kalli, K. R., Kaufmann, S. H., Kennedy, C. J., Köbel, M., Le Page, C., Li, J., Lupat, R., McNally, O. M., McAlpine, J. N., Mes-Masson, A.-M., Mileshkin, L., Provencher, D. M., Pyman, J., Rahimi, K., Rowley, S. M., Salazar, C., Samimi, G., Saunders, H., Semple, T., Sharma, R., Sharpe, A. J., Stephens, A. N., Thio, N., Torres, M. C., Traficante, N., Xing, Z., Zethoven, M., Antill, Y. C., Scott, C. L., Campbell, I. G., & Gorringe, K. L. (2019). The molecular origin and taxonomy of mucinous ovarian carcinoma. *Nature Communications*, *10*, 3935. doi:10.1038/s41467-019-11862-x.

Chen, H., Li, M., & Huang, P. (2019). Lncrna snhg16 promotes hepatocellular carcinoma proliferation, migration and invasion by regulating mir-186 expression. *Journal of Cancer*, *10*, 3571–3581. doi:10.7150/jca.28428.

Chen, H.-z., Wang, X.-r., Zhao, F.-m., Chen, X.-j., Li, X.-s., Ning, G., & Guo, Y.-k. (2021a). A ct-based radiomics nomogram for predicting early recurrence in patients with high-grade serous ovarian cancer. *European Journal of Radiology*, *145*, 110018. doi:10.1016/j.ejrad.2021.110018.

Chen, J.-F., Wu, P., Xia, R., Yang, J., Huo, X.-Y., Gu, D.-Y., Tang, C.-J., De, W., & Yang, F. (2018). Stat3-induced lncrna haglros overexpression contributes to the malignant progression of gastric cancer cells via mtor signal-mediated inhibition of autophagy. *Molecular Cancer*, *17*, 6. doi:10.1186/s12943-017-0756-y.

Chen, S., Chen, Y., Wen, Y., Cai, W., Zhu, P., Yuan, W., Li, Y., Fan, X., Wan, Y., Li, F., & et al. (2021b). mir-590-5p targets rmnd5a and promotes migration in pancreatic adenocarcinoma cell lines. *Oncology Letters*, *22*, 532. doi:10.3892/ol.2021.12793.

Chen, V. W., Ruiz, B., Killeen, J. L., Coté, T. R., Wu, X. C., Correa, C. N., & Howe, H. L. (2003). Pathology and classification of ovarian tumors. *Cancer*, *97*, 2631–2642. doi:10.1002/cncr.11345.

Chen, W., Huang, F., Huang, J., Li, Y., Peng, J., Zhuang, Y., Huang, X., Lu, L., Zhu, Z., & Zhang, S. (2021c). Slc45a4 promotes glycolysis and prevents ampk/ulk1-induced autophagy in tp53 mutant pancreatic ductal adenocarcinoma. *Journal of Gene Medicine*, *23*, e3364. doi:10.1002/jgm.3364.

Chen, X., Lan, H., He, D., Wang, Z., Xu, R., Yuan, J., Xiao, M., Zhang, Y., Gong, L., Xiao, S. et al. (2021d). Analysis of autophagy-related signatures identified two distinct subtypes for evaluating the tumor immune microenvironment and predicting prognosis in ovarian cancer. *Frontiers in oncology*, *11*, 616133. doi:10.3389/fonc.2021.616133.

Chen, Y., Ji, S., Ying, J., Sun, Y., Liu, J., & Yin, G. (2022). Krt6a expedites bladder cancer progression, regulated by mir-31-5p. *Cell Cycle*, *21*, 1479–1490. doi:10.1080/15384101.2022.2054095.

Chinchor, N., & Sundheim, B. M. (1993). Muc-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993* (pp. 69–78). doi:10.3115/1072017.1072026.

Clamp, A. R., James, E. C., McNeish, I. A., Dean, A., Kim, J.-W., O'Donnell, D. M., Hook, J., Coyle, C., Blagden, S., Brenton, J. D., Naik, R., Perren, T., Sundar, S., Cook, A. D., Gopalakrishnan, G. S., Gabra, H., Lord, R., Dark, G., Earl, H. M., Hall, M., Banerjee, S., Glasspool, R. M., Jones, R., Williams, S., Swart, A. M., Stenning, S., Parmar, M., Kaplan, R., & Ledermann, J. A. (2019). Weekly dose-dense chemotherapy in first-line epithelial ovarian, fallopian tube, or primary peritoneal carcinoma treatment (icon8): primary progression free survival analysis results from a gcig phase 3 randomised controlled trial. *The Lancet*, *394*, 2084–2095. doi:10.1016/S0140-6736(19)32259-7.

Cole, A. J., Dwight, T., Gill, A. J., Dickson, K.-A., Zhu, Y., Clarkson, A., Gard, G. B., Maidens, J., Valmadre, S., Clifton-Bligh, R., & Marsh, D. J. (2016). Assessing mutant p53 in primary high-grade serous ovarian cancer using immunohistochemistry and massively parallel sequencing. *Scientific Reports*, *6*, 26191. doi:10.1038/srep26191.

Coleman, R. L., Monk, B. J., Sood, A. K., & Herzog, T. J. (2013). Latest research and treatment of advanced-stage epithelial ovarian cancer. *Nature Reviews Clinical Oncology*, *10*, 211–224. doi:10.1038/nrclinonc.2013.5.

Coleman, R. L., Oza, A. M., Lorusso, D., Aghajanian, C., Oaknin, A., Dean, A., Colombo, N., Weberpals, J. I., Clamp, A., Scambia, G., Leary, A., Holloway, R. W., Gancedo, M. A., Fong, P. C., Goh, J. C., O'Malley, D. M., Armstrong, D. K., Garcia-Donas, J., Swisher, E. M., Floquet, A., Konecny, G. E., McNeish, I. A., Scott, C. L., Cameron, T., Maloney, L., Isaacson, J., Goble, S., Grace, C., Harding, T. C., Raponi, M., Sun, J., Lin, K. K., Giordano, H., & Ledermann, J. A. (2017). Rucaparib maintenance treatment for recurrent ovarian carcinoma after response to platinum therapy (ariel3): a randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet*, *390*, 1949–1961. doi:10.1016/S0140-6736(17)32440-6.

Collinson, F., Qian, W., Fossati, R., Lissoni, A., Williams, C., Parmar, M., Ledermann, J., Colombo, N., & Swart, A. (2014). Optimal treatment of early-stage ovarian cancer. *Annals of Oncology*, *25*, 1165–1171. doi:10.1093/annonc/mdu116.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. doi:10.1007/BF00994018.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE*

transactions on information theory, *13*, 21–27. doi:10.1109/TIT.1967.1053964.

Croce, L., Coperchini, F., Magri, F., Chiovato, L., & Rotondi, M. (2019). The multifaceted anti-cancer effects of braf-inhibitors. *Oncotarget*, *10*, 6623–6640. doi:10.18632/oncotarget.27304.

Cruz-Rodriguez, N., Combita, A. L., Enciso, L. J., Quijano, S. M., Pinzon, P. L., Lozano, O. C., Castillo, J. S., Li, L., Bareño, J., Cardozo, C., Solano, J., Herrera, M. V., Cudris, J., & Zabaleta, J. (2016). High expression of id family and igj genes signature as predictor of low induction treatment response and worst survival in adult hispanic patients with b-acute lymphoblastic leukemia. *Journal of Experimental & Clinical Cancer Research*, *35*, 64. doi:10.1186/s13046-016-0333-z.

Cunningham, J. M., Winham, S. J., Wang, C., Weiglt, B., Fu, Z., Armasu, S. M., McCauley, B. M., Brand, A. H., Chiew, Y.-E., Elishaev, E., Gourley, C., Kennedy, C. J., Laslavic, A., Lester, J., Piskorz, A., Sekowska, M., Brenton, J. D., Churchman, M., DeFazio, A., Drapkin, R., Elias, K. M., Huntsman, D. G., Karlan, B. Y., Köbel, M., Konner, J., Lawrenson, K., Papaemmanuil, E., Bolton, K. L., Modugno, F., & Goode, E. L. (2022). DNA Methylation Profiles of Ovarian Clear Cell Carcinoma. *Cancer Epidemiology, Biomarkers & Prevention*, *31*, 132–141. doi:10.1158/1055-9965.EPI-21-0677.

Cybulska, P., Paula, A. D. C., Tseng, J., Leitao Jr, M. M., Bashashati, A., Huntsman, D. G., Nazeran, T. M., Aghajanian, C., Abu-Rustum, N. R., DeLair, D. F., Shah, S. P., & Weigelt, B. (2019). Molecular profiling and molecular classification of endometrioid ovarian carcinomas. *Gynecologic Oncology*, *154*, 516–523. doi:10.1016/j.ygyno.2019.07.012.

Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L., & Bellazzi, R. (2018). Machine learning methods to predict diabetes complications. *Journal of Diabetes Science and Technology*, *12*, 295–302. doi:10.1177/1932296817706375.

Dai, W., Xu, L., Yu, X., Zhang, G., Guo, H., Liu, H., Song, G., Weng, S., Dong, L., Zhu, J., Liu, T., Guo, C., & Shen, X. (2020). Ogdhl silencing promotes hepatocellular carcinoma by reprogramming glutamine metabolism. *Journal of Hepatology*, *72*, 909–923. doi:10.1016/j.jhep.2019.12.015.

Dai, Y., Yang, G., Yang, L., Jiang, L., Zheng, G., Pan, S., & Zhu, C. (2021). Expression of foxa1 gene regulates the proliferation and invasion of human

gastric cancer cells. *Cellular and Molecular Biology*, *67*, 161–165. doi:10.14715/cmb/2021.67.2.25.

Davidson, B., Bock, A. J., Holth, A., & Nymoen, D. A. (2020). Expression of palladin is associated with disease progression in metastatic high-grade serous carcinoma. *Cytopathology*, *31*, 572–578. doi:10.1111/cyt.12895.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233–240). doi:10.1145/1143844.1143874.

Davis, J., Martin, S. G., Patel, P. M., Green, A. R., Rakha, E. A., Ellis, I. O., & Storr, S. J. (2014). Low calpain-9 is associated with adverse disease-specific survival following endocrine therapy in breast cancer. *BMC Cancer*, *14*, 995. doi:10.1186/1471-2407-14-995.

de Lima, V. C. C., de Carvalho, A. F., Morato-Marques, M., Hashimoto, V. L., Spilborghs, G. M. G. T., Marques, S. M., Landman, G., Torres, C., Braga Ribeiro, K., Brentani, H., Reis, L. F., & Dias, A. A. M. (2013). Tnf-alpha and melphalan modulate a specific group of early expressed genes in a murine melanoma model. *Cytokine*, *62*, 217–225. doi:10.1016/j.cyto.2013.02.022.

Delgado, A., & Guddati, A. K. (2021). Clinical endpoints in oncology-a primer. *American journal of cancer research*, *11*, 1121. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8085844/.

Denard, B., Jiang, S., Peng, Y., & Ye, J. (2018). Creb3l1 as a potential biomarker predicting response of triple negative breast cancer to doxorubicin-based chemotherapy. *BMC Cancer*, *18*, 813. doi:10.1186/s12885-018-4724-8.

Deng, H., Guan, X., Gong, L., Zeng, J., Zhang, H., Chen, M. Y., & Li, G. (2019). Cbx6 is negatively regulated by ezh2 and plays a potential tumor suppressor role in breast cancer. *Scientific Reports*, *9*, 197. doi:10.1038/s41598-018-36560-4.

Deng, X., Xu, J., Hui, J., & Wang, C. (2009). Probability fold change: A robust computational approach for identifying differentially expressed gene lists. *Computer methods and programs in biomedicine*, *93*, 124–139. doi:10.1016/j.cmpb.2008.07.013.

Dieters-Castator, D. Z., Rambau, P. F., Kelemen, L. E., Siegers, G. M., Lajoie, G. A., Postovit, L.-M., & Köbel, M. (2019). Proteomics-Derived

Biomarker Panel Improves Diagnostic Precision to Classify Endometrioid and High-grade Serous Ovarian Carcinoma. *Clinical Cancer Research*, *25*, 4309–4319. doi:10.1158/1078-0432.CCR-18-3818.

Ding, X., Tian, X., Liu, W., & Li, Z. (2020). Cdhr5 inhibits proliferation of hepatocellular carcinoma and predicts clinical prognosis. *Irish Journal of Medical Science (1971 -)*, *189*, 439–447. doi:10.1007/s11845-019-02092-7.

Diniz, D. N., Rezende, M. T., Bianchi, A. G. C., Carneiro, C. M., Ushizima, D. M., de Medeiros, F. N. S., & Souza, M. J. F. (2021). A hierarchical feature-based methodology to perform cervical cancer classification. *Applied Sciences*, *11*, 4091. doi:10.3390/app11094091.

Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using dunn's test. *The Stata Journal: Promoting communications on statistics and Stata*, *15*, 292–300. doi:10.1177/1536867x1501500117.

Domchek, S. M., Aghajanian, C., Shapira-Frommer, R., Schmutzler, R. K., Audeh, M. W., Friedlander, M., na, J. B., Mitchell, G., Fried, G., Stemmer, S. M., Hubert, A., Rosengarten, O., Loman, N., Robertson, J. D., Mann, H., & Kaufman, B. (2016). Efficacy and safety of olaparib monotherapy in germline brca1/2 mutation carriers with advanced ovarian cancer and three or more lines of prior therapy. *Gynecologic Oncology*, *140*, 199–203. doi:10.1016/j.ygyno.2015.12.020.

Dong, C., Qiao, Y., Shang, C., Liao, X., Yuan, X., Cheng, Q., Li, Y., Zhang, J., Wang, Y., Chen, Y., Ge, Q., & Bao, Y. (2022). Non-contact screening system based for covid-19 on xgboost and logistic regression. *Computers in Biology and Medicine*, *141*, 105003. doi:10.1016/j.compbiomed.2021.105003.

Dong, J., & Xu, M. (2019). A 19-mirna support vector machine classifier and a 6-mirna risk score system designed for ovarian cancer patients. *Oncology reports*, *41*, 3233–3243. doi:10.3892/or.2019.7108.

Dorman, S. N., Baranova, K., Knoll, J. H. M., Urquhart, B. L., Mariani, G., Carcangiu, M. L., & Rogan, P. K. (2016). Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. *Molecular Oncology*, *10*, 85–100. doi:10.1016/j.molonc.2015.07.006.

Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. *Procedia Computer Science*, *165*, 631–641.

doi:10.1016/j.procs.2020.01.057. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

Drew, Y., Ledermann, J., Hall, G., Rea, D., Glasspool, R., Highley, M., Jayson, G., Sludden, J., Murray, J., Jamieson, D., Halford, S., Acton, G., Backholer, Z., Mangano, R., Boddy, A., Curtin, N., & Plummer, R. (2016). Phase 2 multicentre trial investigating intermittent and continuous dosing schedules of the poly(adp-ribose) polymerase inhibitor rucaparib in germline brca mutation carriers with advanced ovarian and breast cancer. *British Journal of Cancer*, *114*, 723–730. doi:10.1038/bjc.2016.41.

Du, W.-B., Huang, Z., Luo, L., Tong, S.-P., Li, H.-Q., Li, X., Tong, J.-H., Yao, Y.-L., Zhang, W.-B., & Meng, Y. (2020). Tcf19 aggravates the malignant progression of colorectal cancer by negatively regulating wwc1. *European Review for Medical & Pharmacological Sciences*, *24*, 655–663. doi:10.26355/eurrev_202001_20042.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*, 241–252. doi:10.1080/00401706.1964.10490181.

Dwivedi, A. K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural Computing and Applications*, *29*, 685–693. doi:10.1007/s00521-016-2604-1.

Dzik, C., Reis, S. T., Viana, N. I., Brito, G., Paloppi, I., Nahas, W., Srougi, M., & Leite, K. R. (2017). Gene expression profile of renal cell carcinomas after neoadjuvant treatment with sunitinib: New pathways revealed. *The International Journal of Biological Markers*, *32*, 210–217. doi:10.5301/jbm.5000234.

ElNaggar, A., Robins, D., Baca, Y., Arguello, D., Ulm, M., Arend, R., Mantia-Smaldone, G., Chu, C., Winer, I., Holloway, R., Krivak, T., Jones, N., Galvan-Turner, V., Herzog, T. J., & Brown, J. (2022). Genomic profiling in low grade serous ovarian cancer: Identification of novel markers for disease diagnosis and therapy. *Gynecologic Oncology*, *167*, 306–313. doi:10.1016/j.ygyno.2022.09.022.

Embaby, A., Kutzera, J., Geenen, J. J., Pluim, D., Hofland, I., Sanders, J., Lopez-Yurda, M., Beijnen, J. H., Huitema, A. D. R., Witteveen, P. O., Steeghs, N., van Haaften, G., van Vugt, M. A. T. M., de Ridder, J., & Opdam, F. L. (2023). Wee1 inhibitor adavosertib in combination with carboplatin in advanced tp53 mutated ovarian cancer: A

biomarker-enriched phase ii study. *Gynecologic Oncology*, *174*, 239–246. doi:10.1016/j.ygyno.2023.05.063.

Engqvist, H., Parris, T. Z., Kovács, A., Rönnerman, E. W., Sundfeldt, K., Karlsson, P., & Helou, K. (2020). Validation of novel prognostic biomarkers for early-stage clear-cell, endometrioid and mucinous ovarian carcinomas using immunohistochemistry. *Frontiers in Oncology*, *10*, 162. doi:10.3389/fonc.2020.00162.

Enroth, S., Berggrund, M., Lycke, M., Broberg, J., Lundberg, M., Assarsson, E., Olovsson, M., Stålberg, K., Sundfeldt, K., & Gyllensten, U. (2019). High throughput proteomics identifies a high-accuracy 11 plasma protein biomarker signature for ovarian cancer. *Communications Biology*, *2*, 221. doi:10.1038/s42003-019-0464-9.

Enshaei, A., Robson, C. N., & Edmondson, R. J. (2015). Artificial intelligence systems as prognostic and predictive tools in ovarian cancer. *Annals of Surgical Oncology*, *22*, 3970–3975. doi:10.1245/s10434-015-4475-6.

Erdogan, M. A., Ashour, A., Yuca, E., Gorgulu, K., & Ozpolat, B. (2021). Targeting eukaryotic elongation factor-2 kinase suppresses the growth and peritoneal metastasis of ovarian cancer. *Cellular Signalling*, *81*, 109938. doi:10.1016/j.cellsig.2021.109938.

Escobar, J., Klimowicz, A. C., Dean, M., Chu, P., Nation, J. G., Nelson, G. S., Ghatage, P., Kalloger, S. E., & Köbel, M. (2013). Quantification of er/pr expression in ovarian low-grade serous carcinoma. *Gynecologic Oncology*, *128*, 371–376. doi:10.1016/j.ygyno.2012.10.013.

Etemadmoghadam, D., Weir, B. A., Au-Yeung, G., Alsop, K., Mitchell, G., George, J., Australian Ovarian Cancer Study Group, Davis, S., D'Andrea, A. D., Simpson, K., Hahn, W. C., & Bowtell, D. D. L. (2013). Synthetic lethality between ccne1 amplification and loss of brca1. *Proceedings of the National Academy of Sciences*, *110*, 19489–19494. doi:10.1073/pnas.1314302110.

Everitt, B., & Dunn, G. (1991a). Chapter 4. reducing the dimensionality of multivariate data: principal and correspondence analysis. In *Applied multivariate data analysis* (pp. 46–48). London: Edward Arnold.

Everitt, B., & Dunn, G. (1991b). Chapter 4. reducing the dimensionality of multivariate data: principal and correspondence analysis. In *Applied multivariate data analysis* (p. 45). London: Edward Arnold.

Fader, A. N., Bergstrom, J., Jernigan, A., Tanner, E. J., Roche, K. L., Stone, R. L., Levinson, K. L., Ricci, S., Wethingon, S., Wang, T.-L., Shih, I.-M., Yang, B., Zhang, G., Armstrong, D. K., Gaillard, S., Michener, C., DeBernardo, R., & Rose, P. G. (2017). Primary cytoreductive surgery and adjuvant hormonal monotherapy in women with advanced low-grade serous ovarian carcinoma: Reducing overtreatment without compromising survival? *Gynecologic Oncology*, *147*, 85–91. doi:10.1016/j.ygyno.2017.07.127.

Fader, A. N., Java, J., Ueda, S., Bristow, R. E., Armstrong, D. K., Bookman, M. A., & Gershenson, D. M. (2013). Survival in women with grade 1 serous ovarian carcinoma. *Obstetrics and gynecology*, *122*, 225–232. doi:10.1097/AOG.0b013e31829ce7ec.

Farahani, H., Boschman, J., Farnell, D., Darbandsari, A., Zhang, A., Ahmadvand, P., Jones, S. J. M., Huntsman, D., Köbel, M., Gilks, C. B., Singh, N., & Bashashati, A. (2022). Deep learning-based histotype diagnosis of ovarian carcinoma whole-slide pathology images. *Modern Pathology*, *35*, 1983–1990. doi:https://doi.org/10.1038/s41379-022-01146-z.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*, 861–874. doi:10.1016/j.patrec.2005.10.010.

Fedorova, M. S., Kudryavtseva, A. V., Lakunina, V. A., Snezhkina, A. V., Volchenko, N. N., Slavnova, E. N., Danilova, T. V., Sadritdinova, A. F., Melnikova, N. V., Belova, A. A., Klimina, K. M., Sidorov, D. V., Alekseev, B. Y., Kaprin, A. D., Dmitriev, A. A., & Krasnov, G. S. (2015). Downregulation of ogdhl expression is associated with promoter hypermethylation in colorectal cancer. *Molecular biology*, *49*, 608–617. doi:10.1134/S0026893315040044.

Feng, Y., Gao, Y., Yu, J., Jiang, G., Zhang, X., Lin, X., Han, Q., Rong, X., Xu, H., Li, Q., Qiu, X., & Wang, E. (2019). Ccdc85b promotes non-small cell lung cancer cell proliferation and invasion. *Molecular carcinogenesis*, *58*, 126–134. doi:10.1002/mc.22914.

Fernández, A., López, V., Galar, M., del Jesus, M. J., & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems*, *42*, 97–110. doi:https://doi.org/10.1016/j.knosys.2013.01.018.

Ferroni, P., Zanzotto, F. M., Riondino, S., Scarpato, N., Guadagni, F., & Roselli, M. (2019). Breast cancer prognosis using a machine learning approach. *Cancers*, *11*, 328. doi:10.3390/cancers11030328.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, *78*, 553–569. doi:10.2307/2288117.

Freije, J. P., Fueyo, A., Uría, J., & López-Otin, C. (1991). Human zn-α2-glycoprotein cdna cloning and expression analysis in benign and malignant breast tissues. *FEBS letters*, *290*, 247–249. doi:10.1016/0014-5793(91)81271-9.

Friedlander, M. L., Russell, K., Millis, S., Gatalica, Z., Bender, R., & Voss, A. (2016). Molecular profiling of clear cell ovarian cancers: Identifying potential treatment targets for clinical trials. *International Journal of Gynecologic Cancer*, *26*, 648–654. doi:10.1097/IGC.0000000000000677.

Fu, T., Lin, Y., Lin, L., Yang, Y., Guo, Q., Long, Y., He, H., Bao, Y., Lin, T., Chen, J., Chen, Z., Du, L., Liao, G., Liao, B., & Huang, J. (2022). Network architecture of non-coding rnas provides insights into the pathogenesis of upper tract urothelial carcinoma. *Urologic Oncology: Seminars and Original Investigations*, *40*, 383.e11–383.e21. doi:10.1016/j.urolonc.2022.05.003.

Fu, X., Wang, D., Shu, T., Cui, D., & Fu, Q. (2020). Lncrna nr2f2-as1 positively regulates cdk4 to promote cancer cell proliferation in prostate carcinoma. *The Aging Male*, *23*, 1073–1079. doi:10.1080/13685538.2019.1670157.

Fu, Y., Biglia, N., Wang, Z., Shen, Y., Risch, H. A., Lu, L., Canuto, E. M., Jia, W., Katsaros, D., & Yu, H. (2016). Long non-coding rnas, asap1-it1, fam215a, and linc00472, in epithelial ovarian cancer. *Gynecologic Oncology*, *143*, 642–649. doi:10.1016/j.ygyno.2016.09.021.

Gadekallu, T. R., Khare, N., Bhattacharya, S., Singh, S., Maddikunta, P. K. R., & Srivastava, G. (2023). Deep neural networks to predict diabetic retinopathy. *Journal of Ambient Intelligence and Humanized Computing*, *14*, 5407–5420. doi:10.1007/s12652-020-01963-7.

Gao, J., Wang, M., Li, T., Liu, Q., You, L., & Liao, Q. (2020). Up-regulation of cdhr5 expression promotes malignant phenotype of pancreatic ductal adenocarcinoma. *Journal of Cellular and Molecular Medicine*, *24*, 12726–12735. doi:10.1111/jcmm.15856.

Gao, Y., Xu, D., Yu, G., & Liang, J. (2015). Overexpression of metabolic markers hk1 and pkm2 contributes to lymphatic metastasis and adverse prognosis in chinese gastric cancer. *International journal of clinical and experimental pathology*, *8*, 9264–9271. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4583907/.

Gao, Z.-W., Yang, L., Liu, C., Wang, X., Guo, W.-T., Zhang, H.-Z., & Dong, K. (2022). Distinct roles of adenosine deaminase isoenzymes ada1 and ada2: A pan-cancer analysis. *Frontiers in Immunology*, *13*, 903461. doi:10.3389/fimmu.2022.903461.

Gárate-Escamila, A. K., Hajjam El Hassani, A., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and pca. *Informatics in Medicine Unlocked*, *19*, 100330. doi:10.1016/j.imu.2020.100330.

García, S., & Herrera, F. (2008). Evolutionary training set selection to optimize c4.5 in imbalanced problems. In *2008 Eighth International Conference on Hybrid Intelligent Systems* (pp. 567–572). doi:10.1109/HIS.2008.67.

Geisler, J. P., Buller, E., & Manahan, K. J. (2008). Estrogen receptor α and β expression in a case matched series of serous and endometrioid adenocarcinomas of the ovary. *European journal of gynaecological oncology*, *29*, 126–128. URL: https://pubmed.ncbi.nlm.nih.gov/18459544/.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American statistical Association*, *70*, 320–328. doi:10.1080/01621459.1975.10479865.

Gelmon, K. A., Tischkowitz, M., Mackay, H., Swenerton, K., Robidoux, A., Tonkin, K., Hirte, H., Huntsman, D., Clemons, M., Gilks, B., Yerushalmi, R., Macpherson, E., Carmichael, J., & Oza, A. (2011). Olaparib in patients with recurrent high-grade serous or poorly differentiated ovarian carcinoma or triple-negative breast cancer: a phase 2, multicentre, open-label, non-randomised study. *The Lancet Oncology*, *12*, 852–861. doi:10.1016/S1470-2045(11)70214-5.

George, J., Alsop, K., Etemadmoghadam, D., Hondow, H., Mikeska, T., Dobrovic, A., deFazio, A., for the Australian Ovarian Cancer Study Group, Smyth, G. K., Levine, D. A., Mitchell, G., & Bowtell, D. D. (2013). Nonequivalent Gene Expression and Copy Number Alterations in High-Grade Serous Ovarian Cancers with BRCA1 and BRCA2 Mutations. *Clinical Cancer Research*, *19*, 3474–3484. doi:10.1158/1078-0432.CCR-13-0066.

Gershenson, D. M., Bodurka, D. C., Lu, K. H., Nathan, L. C., Milojevic, L., Wong, K. K., Malpica, A., & Sun, C. C. (2015). Impact of age and primary disease site on outcome in women with low-grade serous carcinoma of the ovary or peritoneum: results of a large single-institution registry of a rare tumor. *Journal of Clinical Oncology*, *33*, 2675–2682. doi:10.1200/JCO.2015.61.0873.

Gershenson, D. M., Sun, C. C., Westin, S. N., Eyada, M., Cobb, L. P., Nathan, L. C., Sood, A. K., Malpica, A., Hillman, R. T., & Wong, K. K. (2022). The genomic landscape of low-grade serous ovarian/peritoneal carcinoma and its impact on clinical outcomes. *Gynecologic Oncology*, *165*, 560–567. doi:10.1016/j.ygyno.2021.11.019.

Ghosh, S., Dasgupta, A., & Swetapadma, A. (2019). A study on support vector machine based linear and non-linear pattern classification. In *2019 International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 24–28). doi:10.1109/ISS1.2019.8908018.

Gilks, C. B., Ionescu, D. N., Kalloger, S. E., Köbel, M., Irving, J., Clarke, B., Santos, J., Le, N., Moravan, V., & Swenerton, K. (2008). Tumor cell type can be reproducibly diagnosed and is of independent prognostic significance in patients with maximally debulked ovarian carcinoma. *Human Pathology*, *39*, 1239–1251. doi:10.1016/j.humpath.2008.01.003.

Gluz, O., Liedtke, C., Gottschalk, N., Pusztai, L., Nitz, U., & Harbeck, N. (2009). Triple-negative breast cancerâĂŤcurrent status and future directions. *Annals of Oncology*, *20*, 1913–1927. doi:https://doi.org/10.1093/annonc/mdp492.

Goff, B. A., Mandel, L. S., Drescher, C. W., Urban, N., Gough, S., Schurman, K. M., Patras, J., Mahony, B. S., & Andersen, M. R. (2007). Development of an ovarian cancer symptom index. *Cancer*, *109*, 221–227. doi:10.1002/cncr.22371.

Goff, B. A., Mandel, L. S., Melancon, C. H., & Muntz, H. G. (2004). Frequency of Symptoms of Ovarian Cancer in Women Presenting to Primary Care Clinics. *JAMA*, *291*, 2705–2712. doi:10.1001/jama.291.22.2705.

González-Martín, A., Pothuri, B., Vergote, I., DePont Christensen, R., Graybill, W., Mirza, M. R., McCormick, C., Lorusso, D., Hoskins, P., Freyer, G., Baumann, K., Jardon, K., Redondo, A., Moore, R. G., Vulsteke, C., O'Cearbhaill, R. E., Lund, B., Backes, F., Barretina-Ginesta, P., Haggerty, A. F., Rubio-Pérez, M. J., Shahin, M. S., Mangili, G., Bradley,

W. H., Bruchim, I., Sun, K., Malinowska, I. A., Li, Y., Gupta, D., & Monk, B. J. (2019). Niraparib in patients with newly diagnosed advanced ovarian cancer. *New England Journal of Medicine*, *381*, 2391–2402. doi:10.1056/NEJMoa1910962.

Gorringe, K. L., Cheasley, D., Wakefield, M. J., Ryland, G. L., Allan, P. E., Alsop, K., Amarasinghe, K. C., Ananda, S., Bowtell, D. D., Christie, M., Chiew, Y.-E., Churchman, M., DeFazio, A., Fereday, S., Gilks, C. B., Gourley, C., Hadley, A. M., Hendley, J., Hunter, S. M., Kaufmann, S. H., Kennedy, C. J., Köbel, M., Le Page, C., Li, J., Lupat, R., McNally, O. M., McAlpine, J. N., Pyman, J., Rowley, S. M., Salazar, C., Saunders, H., Semple, T., Stephens, A. N., Thio, N., Torres, M. C., Traficante, N., Zethoven, M., Antill, Y. C., Campbell, I. G., & Scott, C. L. (2020). Therapeutic options for mucinous ovarian carcinoma. *Gynecologic Oncology*, *156*, 552–560. doi:10.1016/j.ygyno.2019.12.015.

Goundiam, O., Gestraud, P., Popova, T., De la Motte Rouge, T., Fourchotte, V., Gentien, D., Hupé, P., Becette, V., Houdayer, C., Roman-Roman, S., Stern, M.-H., & Sastre-Garau, X. (2015). Histo-genomic stratification reveals the frequent amplification/overexpression of ccne1 and brd4 genes in non-brcaness high grade ovarian carcinoma. *International Journal of Cancer*, *137*, 1890–1900. doi:10.1002/ijc.29568.

Gov, E. (2020). Co-expressed functional module-related genes in ovarian cancer stem cells represent novel prognostic biomarkers in ovarian cancer. *Systems Biology in Reproductive Medicine*, *66*, 255–266. doi:10.1080/19396368.2020.1759730.

Gov, E., & Arga, K. Y. (2017). Differential co-expression analysis reveals a novel prognostic gene module in ovarian cancer. *Scientific reports*, *7*, 4996. doi:10.1038/s41598-017-05298-w.

Grabowski, J. P., Harter, P., Heitz, F., Pujade-Lauraine, E., Reuss, A., Kristensen, G., Ray-Coquard, I., Heitz, J., Traut, A., Pfisterer, J., & du Bois, A. (2016). Operability and chemotherapy responsiveness in advanced low-grade serous ovarian cancer. an analysis of the ago study group meta-database. *Gynecologic Oncology*, *140*, 457–462. doi:10.1016/j.ygyno.2016.01.022.

Grisham, R. N., Iyer, G., Garg, K., DeLair, D., Hyman, D. M., Zhou, Q., Iasonos, A., Berger, M. F., Dao, F., Spriggs, D. R., Levine, D. A., Aghajanian, C., & Solit, D. B. (2013). Braf mutation is associated with early stage

disease and improved outcome in patients with low-grade serous ovarian cancer. *Cancer*, *119*, 548–554. doi:10.1002/cncr.27782.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Journal of the American Medical Association*, *316*, 2402–2410. doi:10.1001/jama.2016.17216.

Gunasegaran, T., & Cheah, Y.-N. (2017). Evolutionary cross validation. In *2017 8th International Conference on Information Technology (ICIT)* (pp. 89–95). doi:10.1109/ICITECH.2017.8079960.

Han, C., Bellone, S., Siegel, E. R., Altwerger, G., Menderes, G., Bonazzoli, E., Egawa-Takata, T., Pettinella, F., Bianchi, A., Riccio, F., Zammataro, L., Yadav, G., Marto, J. A., Penet, M.-F., Levine, D. A., Drapkin, R., Patel, A., Litkouhi, B., Ratner, E., Silasi, D.-A., Huang, G. S., Azodi, M., Schwartz, P. E., & Santin, A. D. (2018). A novel multiple biomarker panel for the early detection of high-grade serous ovarian carcinoma. *Gynecologic Oncology*, *149*, 585–591. doi:10.1016/j.ygyno.2018.03.050.

Han, J., Kamber, M., & Pei, J. (2011). Classification: Basic concepts. In *Data Mining* (pp. 365–366). United States: Elsevier Science & Technology.

Han, W., Du, X., Liu, M., Wang, J., Sun, L., & Li, Y. (2019). Increased expression of long non-coding rna snhg16 correlates with tumor progression and poor prognosis in non-small cell lung cancer. *International Journal of Biological Macromolecules*, *121*, 270–278. doi:10.1016/j.ijbiomac.2018.10.004.

Han, W. K., Alinani, A., Wu, C.-L., Michaelson, D., Loda, M., McGovern, F. J., Thadhani, R., & Bonventre, J. V. (2005). Human kidney injury molecule-1 is a tissue and urinary tumor marker of renal cell carcinoma. *Journal of the American Society of Nephrology: JASN*, *16*, 1126–1134. doi:10.1681/ASN.2004070530.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, *144*, 646–674. doi:10.1016/j.cell.2011.02.013.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009a). Chapter 14. unsupervised learning. In *The elements of statistical learning: data mining, inference, and prediction* (pp. 509–510). Springer volume 2.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009b). Chapter 14. unsupervised learning. In *The elements of statistical learning: data mining, inference, and prediction* (pp. 520–525). Springer volume 2.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009c). Chapter 16. ensemble learning. In *The elements of statistical learning: data mining, inference, and prediction* (pp. 605–606). Springer volume 2.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1263–1284. doi:10.1109/TKDE.2008.239.

He, X., Lin, X., Cai, M., Zheng, X., Lian, L., Fan, D., Wu, X., Lan, P., & Wang, J. (2016). Overexpression of hexokinase 1 as a poor prognosticator in human colorectal cancer. *Tumor Biology*, *37*, 3887–3895. doi:10.1007/s13277-015-4255-8.

He, Y., Jiang, Z., Chen, C., & Wang, X. (2018). Classification of triple-negative breast cancers based on immunogenomic profiling. *Journal of Experimental & Clinical Cancer Research*, *37*, 327. doi:10.1186/s13046-018-1002-1.

Hess, V., A'Hern, R., Nasiri, N., King, D. M., Blake, P. R., Barton, D. P. J., Shepherd, J. H., Ind, T., Bridges, J., Harrington, K., Kaye, S. B., & Gore, M. E. (2004). Mucinous epithelial ovarian cancer: A separate entity requiring specific treatment. *Journal of Clinical Oncology*, *22*, 1040–1044. doi:10.1200/JCO.2004.08.078.

Hewitson, P., Glasziou, P., Watson, E., Towler, B., & Irwig, L. (2008). Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): An update. *American Journal of Gastroenterology*, *103*, 1541–1549. doi:10.1111/j.1572-0241.2008.01875.x.

Hicklin, D. J., & Ellis, L. M. (2005). Role of the vascular endothelial growth factor pathway in tumor growth and angiogenesis. *Journal of Clinical Oncology*, *23*, 1011–1027. doi:10.1200/JCO.2005.06.081.

Hill, C. G., Matyunina, L. V., Walker, D., Benigno, B. B., & McDonald, J. F. (2014). Transcriptional override: a regulatory network model of indirect responses to modulations in microrna expression. *BMC Systems Biology*, *8*, 36. doi:10.1186/1752-0509-8-36.

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). The two-sample location problem. In *Nonparametric statistical methods* (pp. 115–118). John Wiley & Sons, Incorporated. (3rd ed.).

Hollis, R. L., Thomson, J. P., Stanley, B., Churchman, M., Meynert, A. M., Rye, T., Bartos, C., Iida, Y., Croy, I., Mackean, M., Nussey, F., Okamoto, A., Semple, C. A., Gourley, C., & Herrington, C. S. (2020). Molecular stratification of endometrioid ovarian carcinoma predicts clinical outcome. *Nature Communications*, *11*, 4995. doi:10.1038/s41467-020-18819-5.

Hoque, M. O., Kim, M. S., Ostrow, K. L., Liu, J., Wisman, G. B. A., Park, H. L., Poeta, M. L., Jeronimo, C., Henrique, R., Lendvai, A., Schuuring, E., Begum, S., Rosenbaum, E., Ongenaert, M., Yamashita, K., Califano, J., Westra, W., van der Zee, A. G., Criekinge, W. V., & Sidransky, D. (2008). Genome-Wide Promoter Analysis Uncovers Portions of the Cancer Methylome. *Cancer Research*, *68*, 2661–2670. doi:10.1158/0008-5472.CAN-07-5913.

Horowitz, N. S., Larry Maxwell, G., Miller, A., Hamilton, C. A., Rungruang, B., Rodriguez, N., Richard, S. D., Krivak, T. C., Fowler, J. M., Mutch, D. G., Van Le, L., Lee, R. B., Argenta, P., Bender, D., Tewari, K. S., Gershenson, D., Java, J. J., & Bookman, M. A. (2018). Predictive modeling for determination of microscopic residual disease at primary cytoreduction: An nrg oncology/gynecologic oncology group 182 study. *Gynecologic Oncology*, *148*, 49–55. doi:10.1016/j.ygyno.2017.10.011.

Hossain, M. E., Uddin, S., & Khan, A. (2021). Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. *Expert Systems with Applications*, *164*, 113918. doi:10.1016/j.eswa.2020.113918.

Hossin, M., Sulaiman, M. N., Mustapha, A., Mustapha, N., & Rahmat, R. W. (2011). A hybrid evaluation metric for optimizing classifier. In *2011 3rd Conference on Data Mining and Optimization (DMO)* (pp. 165–170). IEEE. doi:10.1109/DMO.2011.5976522.

Hu, C., Liu, Z., Jiang, Y., Shi, O., Zhang, X., Xu, K., Suo, C., Wang, Q., Song, Y., Yu, K., Mao, X., Wu, X., Wu, M., Shi, T., Jiang, W., Mu, L., Tully, D. C., Xu, L., Jin, L., Li, S., Tao, X., Zhang, T., & Chen, X. (2020). Early prediction of mortality risk among patients with severe COVID-19, using machine learning. *International Journal of Epidemiology*, *49*, 1918–1929. doi:10.1093/ije/dyaa171.

Hu, C., Xun, Q., Li, X., He, R., Lu, R., Zhang, S., Hu, X., & Feng, J. (2016). Glcci1 variation is associated with asthma susceptibility and inhaled corticosteroid response in a chinese han population. *Archives of Medical Research*, *47*, 118–125. doi:10.1016/j.arcmed.2016.04.005.

Hu, D., Zhou, M., & Zhu, X. (2019). Deciphering immune-associated genes to predict survival in clear cell renal cell cancer. *BioMed Research International*, *2019*, 2506843. doi:10.1155/2019/2506843.

Hu, K., Yao, L., Xu, Z., Yan, Y., & Li, J. (2022). Prognostic value and therapeutic potential of cbx family members in ovarian cancer. *Frontiers in Cell and Developmental Biology*, *10*, 832354. doi:10.3389/fcell.2022.832354.

Huang, C., Clayton, E. A., Matyunina, L. V., McDonald, L. D., Benigno, B. B., Vannberg, F., & McDonald, J. F. (2018). Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. *Scientific Reports*, *8*, 16444. doi:10.1038/s41598-018-34753-5.

Huang, G.-M., Huang, K.-Y., Lee, T.-Y., & Weng, J. T.-Y. (2015). An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. *BMC Bioinformatics*, *16*, S5. doi:10.1186/1471-2105-16-S1-S5.

Huang, Y., Pan, J., Chen, D., Zheng, J., Qiu, F., Li, F., Wu, Y., Wu, W., Huang, X., & Qian, J. (2017). Identification and functional analysis of differentially expressed genes in poorly differentiated hepatocellular carcinoma using rna-seq. *Oncotarget*, *8*, 35973–35983. doi:10.18632/oncotarget.16415.

Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, *36*, 1420–1438. doi:10.1016/j.ijforecast.2020.02.005.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., & Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*, 115–121. doi:doi.org/10.1038/nmeth.3252.

Hwang, J. R., Cho, Y.-J., Lee, Y., Park, Y., Han, H. D., Ahn, H. J., Lee, J.-H., & Lee, J.-W. (2016). The c-terminus of igfbp-5 suppresses tumor growth by inhibiting angiogenesis. *Scientific Reports*, *6*, 39334. doi:10.1038/srep39334.

Hwangbo, S., Kim, S. I., Kim, J.-H., Eoh, K. J., Lee, C., Kim, Y. T., Suh, D.-S., Park, T., & Song, Y. S. (2021). Development of machine learning models to predict platinum sensitivity of high-grade serous ovarian carcinoma. *Cancers*, *13*, 1875. doi:10.3390/cancers13081875.

Iida, Y., Okamoto, A., Hollis, R. L., Gourley, C., & Herrington, C. S. (2021). Clear cell carcinoma of the ovary: a clinical and molecular perspective. *International Journal of Gynecologic Cancer*, *31*, 605–616. doi:10.1136/ijgc-2020-001656.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003a). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, *31*, e15. doi:10.1093/nar/gng015.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, *4*, 249–264. doi:10.1093/biostatistics/4.2.249.

Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: A comparative study using machine learning techniques. *SN Computer Science*, *1*, 290. doi:10.1007/s42979-020-00305-w.

Itamochi, H., Kigawa, J., Sugiyama, T., Kikuchi, Y., Suzuki, M., & Terakawa, N. (2002). Low proliferation activity may be associated with chemoresistance in clear cell carcinoma of the ovary. *Obstetrics & Gynecology*, *100*, 281–287. doi:10.1016/S0029-7844(02)02040-9.

Itamochi, H., Kigawa, J., & Terakawa, N. (2008). Mechanisms of chemoresistance and poor prognosis in ovarian clear cell carcinoma. *Cancer Science*, *99*, 653–658. doi:10.1111/j.1349-7006.2008.00747.x.

Itamochi, H., Oishi, T., Oumi, N., Takeuchi, S., Yoshihara, K., Mikami, M., Yaegashi, N., Terao, Y., Takehara, K., Ushijima, K., Watari, H., Aoki, D., Kimura, T., Nakamura, T., Yokoyama, Y., Kigawa, J., & Sugiyama, T. (2017). Whole-genome sequencing revealed novel prognostic biomarkers and promising targets for therapy of ovarian clear cell carcinoma. *British Journal of Cancer*, *117*, 717–724. doi:10.1038/bjc.2017.228.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, *37*, 547–579. doi:10.5169/seals-266450.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). 9. support vector machines. In *An Introduction to Statistical Learning: With Applications in R* (pp. 337–349). Springer. (1st ed.).

Jeffery, I. B., Higgins, D. G., & Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC bioinformatics*, *7*, 359. doi:10.1186/1471-2105-7-359.

Jia, Q., Liao, X., Zhang, Y., Xu, B., Song, Y., Bian, G., & Fu, X. (2022). Anti-tumor role of camk2b in remodeling the stromal microenvironment and inhibiting proliferation in papillary renal cell carcinoma. *Frontiers in Oncology*, *12*, 740051. doi:10.3389/fonc.2022.740051.

Jiang, M., Qi, L., Jin, K., Li, L., Wu, Y., Song, D., Gan, J., Huang, M., Li, Y., & Song, C. (2021a). eef2k as a novel metastatic and prognostic biomarker in gastric cancer patients. *Pathology-Research and Practice*, *225*, 153568. doi:10.1016/j.prp.2021.153568.

Jiang, X., Zhang, W., Kayed, H., Zheng, P., Giese, N. A., Friess, H., & Kleeff, J. (2008). Loss of onecut1 expression in human pancreatic cancer cells. *Oncology reports*, *19*, 157–163. doi:10.3892/or.19.1.157.

Jiang, Y., Xun, Q., Wan, R., Deng, S., Hu, X., Luo, L., Li, X., & Feng, J. (2021b). Glcci1 gene body methylation in peripheral blood is associated with asthma and asthma severity. *Clinica Chimica Acta*, *523*, 97–105. doi:10.1016/j.cca.2021.09.006.

Jiao, Y., Li, Y., Fu, Z., Hou, L., Chen, Q., Cai, Y., Jiang, P., He, M., & Yang, Z. (2019). Ogdhl expression as a prognostic biomarker for liver cancer patients. *Disease Markers*, *2019*, 9037131. doi:10.1155/2019/9037131.

Jin, Q., Liu, G., Domeier, P. P., Ding, W., & Mulder, K. M. (2013). Decreased tumor progression and invasion by a novel anti-cell motility target for human colorectal carcinoma cells. *PloS one*, *8*, e66439. doi:10.1371/journal.pone.0066439.

John, C. R. (2020). *MLeval: Machine Learning Model Evaluation*. URL: https://CRAN.R-project.org/package=MLeval r package version 0.3.

Jones, S., Wang, T.-L., Kurman, R. J., Nakayama, K., Velculescu, V. E., Vogelstein, B., Kinzler, K. W., Papadopoulos, N., & Shih, I.-M. (2012). Low-grade serous carcinomas of the ovary contain very few point mutations. *The Journal of Pathology*, *226*, 413–420. doi:10.1002/path.3967.

Jones, S., Wang, T.-L., Shih, I.-M., Mao, T.-L., Nakayama, K., Roden, R., Glas, R., Slamon, D., Diaz, L. A., Vogelstein, B., Kinzler, K. W., Velculescu, V. E., & Papadopoulos, N. (2010). Frequent mutations of chromatin remodeling gene arid1a in ovarian clear cell carcinoma. *Science*, *330*, 228–231. doi:10.1126/science.1196333.

Jung, W. Y., Sung, C. O., Han, S. H., Kim, K., Kim, M., Ro, J. Y., Kang, M. J., Ahn, H., & Cho, Y. M. (2014). Azgp-1 immunohistochemical marker in prostate cancer: potential predictive marker of biochemical recurrence in post radical prostatectomy specimens. *Applied Immunohistochemistry & Molecular Morphology*, *22*, 652–657. doi:10.1097/PAI.0000000000000015.

Kang, E. Y., Cheasley, D., LePage, C., Wakefield, M. J., da Cunha Torres, M., Rowley, S., Salazar, C., Xing, Z., Allan, P., Bowtell, D. D. L., Mes-Masson, A.-M., Provencher, D. M., Rahimi, K., Kelemen, L. E., Fasching, P. A., Doherty, J. A., Goodman, M. T., Goode, E. L., Deen, S., Pharoah, P. D. P., Brenton, J. D., Sieh, W., Mateoiu, C., Sundfeldt, K., Cook, L. S., Le, N. D., Anglesio, M. S., Gilks, C. B., Huntsman, D. G., Kennedy, C. J., Traficante, N., Bowtell, D., Chenevix-Trench, G., Green, A., Webb, P., DeFazio, A., Gertig, D., Traficante, N., Fereday, S., Moore, S., Hung, J., Harrap, K., Sadkowsky, T., Pandeya, N., Malt, M., Mellon, A., Robertson, R., Bergh, T. V., Jones, M., Mackenzie, P., Maidens, J., Nattress, K., Chiew, Y. E., Stenlake, A., Sullivan, H., Alexander, B., Ashover, P., Brown, S., Corrish, T., Green, L., Jackman, L., Ferguson, K., Martin, K., Martyn, A., Ranieri, B., White, J., Jayde, V., Mamers, P., Bowes, L., Galletta, L., Giles, D., Hendley, J., Alsop, K., Schmidt, T., Shirley, H., Ball, C., Young, C., Viduka, S., Tran, H., Bilic, S., Glavinas, L., Brooks, J., Stuart-Harris, R., Kirsten, F., Rutovitz, J., Clingan, P., Glasgow, A., Proietto, A., Braye, S., Otton, G., Shannon, J., Bonaventura, T., Stewart, J., Begbie, S., Friedlander, M., Bell, D., Baron-Hay, S., Ferrier, A., Gard, G. et al. (2021a). Refined cut-off for tp53 immunohistochemistry improves prediction of tp53 mutation status in ovarian mucinous tumors: implications for outcome analyses. *Modern Pathology*, *34*, 194–206. doi:10.1038/s41379-020-0618-9.

Kang, H., Choi, M. C., Kim, S., Jeong, J.-Y., Kwon, A.-Y., Kim, T.-H., Kim,

G., Joo, W. D., Park, H., Lee, C., Song, S. H., Jung, S. G., Hwang, S., & An, H. J. (2021b). Usp19 and rpl23 as candidate prognostic markers for advanced-stage high-grade serous ovarian carcinoma. *Cancers*, *13*, 3976. doi:10.3390/cancers13163976.

Karabuk, E., Kose, M. F., Hizli, D., Taşkin, S., Karadağ, B., Turan, T., Boran, N., Ozfuttu, A., & Ortaç, U. F. (2013). Comparison of advanced stage mucinous epithelial ovarian cancer and serous epithelial ovarian cancer with regard to chemosensitivity and survival outcome: a matched case-control study. *J Gynecol Oncol*, *24*, 160–166. doi:10.3802/jgo.2013.24.2.160.

Karam, A., Ledermann, J. A., Kim, J.-W., Sehouli, J., Lu, K., Gourley, C., Katsumata, N., Burger, R. A., Nam, B.-H., Bacon, M., Ng, C., Pfisterer, J., Bekkers, R., Casado Herráez, A., Redondo, A., Fujiwara, H., Gleeson, N., Rosengarten, O., Scambia, G., Zhu, J., Okamoto, A., Stuart, G., & Ochiai, K. (2017). Fifth ovarian cancer consensus conference of the gynecologic cancer intergroup: first-line interventions. *Annals of Oncology*, *28*, 711–717. doi:10.1093/annonc/mdx011.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, *11*, 1–20. doi:10.18637/jss.v011.i09.

Katsumata, N., Yasuda, M., Isonishi, S., Takahashi, F., Michimae, H., Kimura, E., Aoki, D., Jobo, T., Kodama, S., Terauchi, F., Sugiyama, T., & Ochiai, K. (2013). Long-term results of dose-dense paclitaxel and carboplatin versus conventional paclitaxel and carboplatin for treatment of advanced epithelial ovarian, fallopian tube, or primary peritoneal cancer (jgog 3016): a randomised, controlled, open-label trial. *The Lancet Oncology*, *14*, 1020–1026. doi:10.1016/S1470-2045(13)70363-2.

Katsumata, N., Yasuda, M., Takahashi, F., Isonishi, S., Jobo, T., Aoki, D., Tsuda, H., Sugiyama, T., Kodama, S., Kimura, E., Ochiai, K., & Noda, K. (2009). Dose-dense paclitaxel once a week in combination with carboplatin every 3 weeks for advanced ovarian cancer: a phase 3, open-label, randomised controlled trial. *The Lancet*, *374*, 1331–1338. doi:10.1016/S0140-6736(09)61157-0.

Kawakami, E., Tabata, J., Yanaihara, N., Ishikawa, T., Koseki, K., Iida, Y., Saito, M., Komazaki, H., Shapiro, J. S., Goto, C., Akiyama, Y., Saito, R., Saito, M., Takano, H., Yamada, K., & Okamoto, A. (2019). Application

of Artificial Intelligence for Preoperative Diagnostic and Prognostic Prediction in Epithelial Ovarian Cancer Based on Blood Biomarkers. *Clinical Cancer Research*, *25*, 3006–3015. doi:10.1158/1078-0432.CCR-18-3378.

Kazemi, M., Moghimbeigi, A., Kiani, J., Mahjub, H., & Faradmal, J. (2016). Diabetic peripheral neuropathy class prediction by multicategory support vector machine model: a cross-sectional study. *Epidemiology and health*, *38*, e2016011. doi:10.4178/epih.e2016011.

Ke, C., Hou, Y., Zhang, H., Fan, L., Ge, T., Guo, B., Zhang, F., Yang, K., Wang, J., Lou, G., & Li, K. (2015). Large-scale profiling of metabolic dysregulation in ovarian cancer. *International journal of cancer*, *136*, 516–526. doi:10.1002/ijc.29010.

Kehoe, S., Hook, J., Nankivell, M., Jayson, G. C., Kitchener, H., Lopes, T., Luesley, D., Perren, T., Bannoo, S., Mascarenhas, M., Dobbs, S., Essapen, S., Twigg, J., Herod, J., McCluggage, G., Parmar, M., & Swart, A.-M. (2015). Primary chemotherapy versus primary surgery for newly diagnosed advanced ovarian cancer (chorus): an open-label, randomised, controlled, non-inferiority trial. *The Lancet*, *386*, 249–257. doi:10.1016/S0140-6736(14)62223-6.

Kessous, R., Laskov, I., Abitbol, J., Bitharas, J., Yasmeen, A., Salvador, S., Lau, S., & Gotlieb, W. H. (2017). Clinical outcome of neoadjuvant chemotherapy for advanced ovarian cancer. *Gynecologic Oncology*, *144*, 474–479. doi:10.1016/j.ygyno.2016.12.017.

Khalaj-Kondori, M., Hosseinnejad, M., Hosseinzadeh, A., Behroz Sharif, S., & Hashemzadeh, S. (2020). Aberrant hypermethylation of ogdhl gene promoter in sporadic colorectal cancer. *Current Problems in Cancer*, *44*, 100471. doi:10.1016/j.currproblcancer.2019.03.001.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*, 3735–3745. doi:10.1016/j.csda.2009.04.009.

Kim, J.-H., Kim, T.-W., & Kim, S.-J. (2011). Downregulation of arfgef1 and camk2b by promoter hypermethylation in breast cancer cells. *BMB reports*, *44*, 523–528. doi:10.5483/bmbrep.2011.44.8.523.

Kim, M.-K., Kim, K., Kim, S. M., Kim, J. W., Park, N.-H., Song, Y.-S., & Kang, S.-B. (2009). A hospital-based case-control study of identifying ovarian cancer using symptom index. *Journal of Gynecologic Oncology*, *20*, 238–242. doi:10.3802/jgo.2009.20.4.238.

Kim, S. I., Lee, J. W., Lee, M., Kim, H. S., Chung, H. H., Kim, J.-W., Park, N. H., Song, Y.-S., & Seo, J.-S. (2018). Genomic landscape of ovarian clear cell carcinoma via whole exome sequencing. *Gynecologic Oncology*, *148*, 375–382. doi:10.1016/j.ygyno.2017.12.005.

King, E. R., Tung, C. S., Tsang, Y. T. M., Zu, Z., Lok, G. T. M., Deavers, M. T., Malpica, A., Wolf, J. K., Lu, K. H., Birrer, M. J., Mok, S. C., Gershenson, D. M., & Wong, K.-K. (2011). The anterior gradient homolog 3 (agr3) gene is associated with differentiation and survival in ovarian cancer. *The American journal of surgical pathology*, *35*, 904–912. doi:10.1097/PAS.0b013e318212ae22.

Klein, O., Kanter, F., Kulbe, H., Jank, P., Denkert, C., Nebrich, G., Schmitt, W. D., Wu, Z., Kunze, C. A., Sehouli, J., Darb-Esfahani, S., Braicu, I., Lellmann, J., Thiele, H., & Taube, E. T. (2019). Maldi-imaging for classification of epithelial ovarian cancer histotypes from a tissue microarray using machine learning methods. *PROTEOMICS-Clinical Applications*, *13*, 1700181. doi:10.1002/prca.201700181.

Köbel, M., Kalloger, S. E., Boyd, N., McKinney, S., Mehl, E., Palmer, C., Leung, S., Bowen, N. J., Ionescu, D. N., Rajput, A., Prentice, L. M., Miller, D., Santos, J., Swenerton, K., Gilks, C. B., & Huntsman, D. (2008). Ovarian carcinoma subtypes are different diseases: Implications for biomarker studies. *PLOS Medicine*, *5*, e232. doi:10.1371/journal.pmed.0050232.

Köbel, M., Kalloger, S. E., Huntsman, D. G., Santos, J. L., Swenerton, K. D., Seidman, J. D., Gilks, C. B., & on behalf of the Cheryl Brown Ovarian Cancer Outcomes Unit of the British Columbia Cancer Agency, V. B. (2010). Differences in tumor type in low-stage versus high-stage ovarian carcinomas. *International Journal of Gynecological Pathology*, *29*, 203–211. doi:10.1097/PGP.0b013e3181c042b6.

Köbel, M., Rahimi, K., Rambau, P. F., Naugler, C., Le Page, C., Meunier, L., De Ladurantaye, M., Lee, S., Leung, S., Goode, E. L., Ramus, S. J., Carlson, J. W., Li, X., Ewanowich, C. A., Kelemen, L. E., Vanderhyden, B., Provencher, D., Huntsman, D., Lee, C.-H., Gilks, C. B., & Mes Masson, A.-M. (2016). An immunohistochemical algorithm for ovarian carcinoma typing. *International Journal of Gynecological Pathology*, *35*, 430–441. doi:10.1097/PGP.0000000000000274.

Koh, V., Kwan, H. Y., Tan, W. L., Mah, T. L., & Yong, W. P. (2016). Knockdown of pola2 increases gemcitabine resistance in lung cancer cells. *BMC Genomics*, *17*, 1029. doi:10.1186/s12864-016-3322-x.

Komatsu, S., Ichikawa, D., Hirajima, S., Nagata, H., Nishimura, Y., Kawaguchi, T., Miyamae, M., Okajima, W., Ohashi, T., Konishi, H., Shiozaki, A., Fujiwara, H., Okamoto, K., Tsuda, H., Imoto, I., Inazawa, J., & Otsuji, E. (2015). Overexpression of smyd2 contributes to malignant outcome in gastric cancer. *British Journal of Cancer*, *112*, 357–364. doi:10.1038/bjc.2014.543.

Konecny, G. E., & Kristeleit, R. S. (2016). Parp inhibitors for brca1/2-mutated and sporadic ovarian cancer: current practice and future directions. *British Journal of Cancer*, *115*, 1157–1173. doi:10.1038/bjc.2016.311.

Kristeleit, R., Shapiro, G. I., Burris, H. A., Oza, A. M., LoRusso, P., Patel, M. R., Domchek, S. M., Balmaña, J., Drew, Y., Chen, L.-m., Safra, T., Montes, A., Giordano, H., Maloney, L., Goble, S., Isaacson, J., Xiao, J., Borrow, J., Rolfe, L., & Shapira-Frommer, R. (2017). A Phase I-II Study of the Oral PARP Inhibitor Rucaparib in Patients with Germline BRCA1/2-Mutated Ovarian Carcinoma or Other Solid Tumors. *Clinical Cancer Research*, *23*, 4095–4106. doi:10.1158/1078-0432.CCR-16-2796.

Kuhn, E., Wang, T.-L., Doberstein, K., Bahadirli-Talbott, A., Ayhan, A., Sehdev, A. S., Drapkin, R., Kurman, R. J., & Shih, I.-M. (2016). Ccne1 amplification and centrosome number abnormality in serous tubal intraepithelial carcinoma: further evidence supporting its role as a precursor of ovarian high-grade serous carcinoma. *Modern Pathology*, *29*, 1254–1261. doi:10.1038/modpathol.2016.101.

Kukita, A., Sone, K., Oda, K., Hamamoto, R., Kaneko, S., Komatsu, M., Wada, M., Honjoh, H., Kawata, Y., Kojima, M., Oki, S., Sato, M., Asada, K., Taguchi, A., Miyasaka, A., Tanikawa, M., Nagasaka, K., Matsumoto, Y., Wada-Hiraike, O., Osuga, Y., & Fujii, T. (2019). Histone methyltransferase smyd2 selective inhibitor lly-507 in combination with poly adp ribose polymerase inhibitor has therapeutic potential against high-grade serous ovarian carcinomas. *Biochemical and Biophysical Research Communications*, *513*, 340–346. doi:10.1016/j.bbrc.2019.03.155.

Kumar, A., Sheedy, S., Kim, B., Suidan, R., Sarasohn, D. M., Nikolovski, I., Lakhman, Y., McGree, M. E., Weaver, A. L., Chi, D., & Cliby, W. A. (2019). Models to predict outcomes after primary debulking surgery: Independent validation of models to predict suboptimal cytoreduction and gross residual disease. *Gynecologic Oncology*, *154*, 72–76. doi:10.1016/j.ygyno.2019.04.011.

Kurman, R. J., & Shih, I.-M. (2016). The dualistic model of ovarian carcinogenesis: Revisited, revised, and expanded. *The American Journal of Pathology*, *186*, 733–747. doi:10.1016/j.ajpath.2015.11.011.

Laios, A., Kalampokis, E., Johnson, R., Thangavelu, A., Tarabanis, C., Nugent, D., & De Jong, D. (2022). Explainable artificial intelligence for prediction of complete surgical cytoreduction in advanced-stage epithelial ovarian cancer. *Journal of Personalized Medicine*, *12*, 607. doi:10.3390/jpm12040607.

Lan, A., & Yang, G. (2019). Clinicopathological parameters and survival of invasive epithelial ovarian cancer by histotype and disease stage. *Future Oncology*, *15*, 2029–2039. doi:10.2217/fon-2018-0886.

Lau, T. P., Roslani, A. C., Lian, L. H., Chai, H. C., Lee, P. C., Hilmi, I., Goh, K. L., & Chua, K. H. (2014). Pair-wise comparison analysis of differential expression of mrnas in early and advanced stage primary colorectal adenocarcinomas. *BMJ open*, *4*, e004930. doi:10.1136/bmjopen-2014-004930.

Lawrie, T. A., Winter-Roach, B. A., Heus, P., & Kitchener, H. C. (2015). Adjuvant (post-surgery) chemotherapy for early stage epithelial ovarian cancer. *Cochrane Database of Systematic Reviews*, . doi:10.1002/14651858.CD004706.pub5. Issue 12.

Ledermann, J., Harter, P., Gourley, C., Friedlander, M., Vergote, I., Rustin, G., Scott, C. L., Meier, W., Shapira-Frommer, R., Safra, T., Matei, D., Fielding, A., Spencer, S., Dougherty, B., Orr, M., Hodgson, D., Barrett, J. C., & Matulonis, U. (2014). Olaparib maintenance therapy in patients with platinum-sensitive relapsed serous ovarian cancer: a preplanned retrospective analysis of outcomes by brca status in a randomised phase 2 trial. *The Lancet Oncology*, *15*, 852–861. doi:10.1016/S1470-2045(14)70228-1.

Ledermann, J. A., Harter, P., Gourley, C., Friedlander, M., Vergote, I., Rustin, G., Scott, C., Meier, W., Shapira-Frommer, R., Safra, T., Matei, D., Fielding, A., Spencer, S., Rowe, P., Lowe, E., Hodgson, D., Sovak, M. A., & Matulonis, U. (2016). Overall survival in patients with platinum-sensitive recurrent serous ovarian cancer receiving olaparib maintenance monotherapy: an updated analysis from a randomised, placebo-controlled, double-blind, phase 2 trial. *The Lancet Oncology*, *17*, 1579–1589. doi:10.1016/S1470-2045(16)30376-X.

Lee, G. Y., Haverty, P. M., Li, L., Kljavin, N. M., Bourgon, R., Lee, J., Stern, H., Modrusan, Z., Seshagiri, S., Zhang, Z., Davis, D., Stokoe, D., Settleman, J., de Sauvage, F. J., & Neve, R. M. (2014). Comparative Oncogenomics Identifies PSMB4 and SHMT2 as Potential Cancer Driver Genes. *Cancer Research*, *74*, 3114–3126. doi:10.1158/0008-5472.CAN-13-2683.

Lee, J. K. (2014). 4. statistical testing and significance for large biological data analysis. In *Statistical Bioinformatics: For Biomedical and Life Science Researchers* (pp. 72–73). John Wiley & Sons, Incorporated.

Lee, P., Rosen, D. G., Zhu, C., Silva, E. G., & Liu, J. (2005). Expression of progesterone receptor is a favorable prognostic marker in ovarian cancer. *Gynecologic Oncology*, *96*, 671–677. doi:10.1016/j.ygyno.2004.11.010.

Lee, Y.-J., Lee, M.-Y., Ruan, A., Chen, C.-K., Liu, H.-P., Wang, C.-J., Chao, W.-R., & Han, C.-P. (2016). Multipoint kras oncogene mutations potentially indicate mucinous carcinoma on the entire spectrum of mucinous ovarian neoplasms. *Oncotarget*, *7*, 82097–82103. doi:10.18632\%2Foncotarget.13449.

Leong, H. S., Galletta, L., Etemadmoghadam, D., George, J., The Australian Ovarian Cancer Study, Köbel, M., Ramus, S. J., & Bowtell, D. (2015). Efficient molecular subtype classification of high-grade serous ovarian cancer. *The Journal of Pathology*, *236*, 272–277. doi:10.1002/path.4536.

Li, H. M., Gong, J., Li, R. M., Xiao, Z. B., Qiang, J. W., Peng, W. J., & Gu, Y. J. (2021a). Development of mri-based radiomics model to predict the risk of recurrence in patients with advanced high-grade serous ovarian carcinoma. *American Journal of Roentgenology*, *217*, 664–675. doi:10.2214/AJR.20.23195.

Li, J., Li, M.-h., Wang, T.-t., Liu, X.-n., Zhu, X.-t., Dai, Y.-z., Zhai, K.-c., Liu, Y.-d., Lin, J.-l., Ge, R.-l., Sun, S.-h., Wang, F., & Yuan, J.-h. (2021b). Slc38a4 functions as a tumour suppressor in hepatocellular carcinoma through modulating wnt/β-catenin/myc/hmgcs2 axis. *British Journal of Cancer*, *125*, 865–876. doi:10.1038/s41416-021-01490-y.

Li, J., Olson, L. M., Zhang, Z., Li, L., Bidder, M., Nguyen, L., Pfeifer, J., & Rader, J. S. (2008). Differential display identifies overexpression of the usp36 gene, encoding a deubiquitinating enzyme, in ovarian cancer. *International Journal of Medical Sciences*, *5*, 133–142. doi:10.7150/ijms.5.133.

Li, L. X., Zhou, J. X., Calvet, J. P., Godwin, A. K., Jensen, R. A., & Li, X. (2018). Lysine methyltransferase smyd2 promotes triple negative breast cancer progression. *Cell Death and Disease*, *9*, 326. doi:10.1038/s41419-018-0347-x.

Li, P., Zhang, K., Tang, S., & Tang, W. (2022). Knockdown of lncrna haglros inhibits metastasis and promotes apoptosis in nephroblastoma cells by inhibition of autophagy. *Bioengineered*, *13*, 7552–7562. doi:10.1080/21655979.2021.2023984.

Li, Q., Birkbak, N. J., Gyorffy, B., Szallasi, Z., & Eklund, A. C. (2011). Jetset: selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics*, *12*, 474. doi:10.1186/1471-2105-12-474.

Li, S., Lin, Y., Zhu, T., Fan, M., Xu, S., Qiu, W., Chen, C., Li, L., Wang, Y., Yan, J., Wong, J., Naing, L., & Xu, S. (2023). Development and external evaluation of predictions models for mortality of covid-19 patients using machine learning method. *Neural Computing and Applications*, *35*, 13037–13046. doi:10.1007/s00521-020-05592-1.

Li, X., Gou, J., Li, H., & Yang, X. (2020). Bioinformatic analysis of the expression and prognostic value of chromobox family proteins in human breast cancer. *Scientific reports*, *10*, 17739. doi:10.1038/s41598-020-74792-5.

Li, Y., Lu, Y., & Chen, Y. (2019). Long non-coding RNA SNHG16 affects cell proliferation and predicts a poor prognosis in patients with colorectal cancer via sponging miR-200a-3p. *Bioscience Reports*, *39*, BSR20182498. doi:10.1042/BSR20182498.

Liang, P.-I., Wang, Y.-H., Wu, T.-F., Wu, W.-R., Liao, A. C., Shen, K.-H., Hsing, C.-H., Shiue, Y.-L., Huang, H.-Y., Hsu, H.-P., Chen, L.-T., Lin, C.-Y., Tai, C., Wu, J.-Y., & Li, C.-F. (2013). Igfbp-5 overexpression as a poor prognostic factor in patients with urothelial carcinomas of upper urinary tracts and urinary bladder. *Journal of Clinical Pathology*, *66*, 573–582. doi:10.1136/jclinpath-2012-201278.

Liang, Y., Wu, X., Lee, J., Yu, D., Su, J., Guo, M., Meng, N., Qin, J., & Fan, X. (2022). lncrna nr2f2-as1 inhibits the methylation of mir-494 to regulate oral squamous cell carcinoma cell proliferation. *Archives of Oral Biology*, *134*, 105316. doi:10.1016/j.archoralbio.2021.105316.

Liang, Y.-K., Lin, H.-Y., Chen, C.-F., & Zeng, D. (2017). Prognostic values of distinct cbx family members in breast cancer. *Oncotarget*, *8*, 92375–92387. doi:10.18632/oncotarget.21325.

Lin, F., Zhang, P. L., Yang, X. J., Shi, J., Blasick, T., Han, W. K., Wang, H. L., Shen, S. S., Teh, B. T., & Bonventre, J. V. (2007). Human kidney injury molecule-1 (hkim-1): A useful immunohistochemical marker for diagnosing renal cell carcinoma and ovarian clear cell carcinoma. *The American Journal of Surgical Pathology*, *31*, 371–381. doi:10.1097/01.pas.0000213353.95508.67.

Lisowska, K. M., Olbryt, M., Student, S., Kujawa, K. A., Cortez, A. J., Simek, K., Dansonka-Mieszkowska, A., Rzepecka, I. K., Tudrej, P., & Kupryjańczyk, J. (2016). Unsupervised analysis reveals two molecular subgroups of serous ovarian cancer with distinct gene expression profiles and survival. *Journal of cancer research and clinical oncology*, *142*, 1239–1252. doi:10.1007/s00432-016-2147-y.

Liu, C., Xia, Y., Jiang, W., Liu, Y., & Yu, L. (2014a). Low expression of gabarapl1 is associated with a poor outcome for patients with hepatocellular carcinoma. *Oncology reports*, *31*, 2043–2048. doi:10.3892/or.2014.3096.

Liu, F., Liu, J., Zhang, J., Shi, J., Gui, L., & Xu, G. (2020a). Expression of stat1 is positively correlated with pd-l1 in human ovarian cancer. *Cancer biology & therapy*, *21*, 963–971. doi:10.1080/15384047.2020.1824479.

Liu, H., Xiang, L., Mei, Y. et al. (2022a). mir-877-5p inhibits epithelial mesenchymal transformation of breast cancer cells by targeting fgb. *Disease Markers*, *2022*, 4882375. doi:10.1155/2022/4882375.

Liu, J. F., Barry, W. T., Birrer, M., Lee, J.-M., Buckanovich, R. J., Fleming, G. F., Rimel, B. J., Buss, M. K., Nattam, S., Hurteau, J., Luo, W., Quy, P., Whalen, C., Obermayer, L., Lee, H., Winer, E. P., Kohn, E. C., Ivy, S. P., & Matulonis, U. A. (2014b). Combination cediranib and olaparib versus olaparib alone for women with recurrent platinum-sensitive ovarian cancer: a randomised phase 2 study. *The Lancet. Oncology*, *15*, 1207–1214. doi:10.1016/S1470-2045(14)70391-2.

Liu, Y., Meng, F., Wang, J., Liu, M., Yang, G., Song, R., Zheng, T., Liang, Y., Zhang, S., Yin, D., Wang, J., Yang, H., Pan, S., Sun, B., Han, J., Sun, J., Lan, Y., Wang, Y., Liu, X., Zhu, M., Cui, Y., Zhang, B., Wu, D., Liang, S., Liu, Y., Song, X., Lu, Z., Yang, J., Li, M., & Liu, L. (2019a). A

novel oxoglutarate dehydrogenase-like mediated mir-214/twist1 negative feedback loop inhibits pancreatic cancer growth and metastasis. *Clinical Cancer Research*, *25*, 5407–5421. doi:10.1158/1078-0432.ccr-18-4113.

Liu, Y., Wang, Y., Ni, Y., Cheung, C. K. Y., Lam, K. S. L., Wang, Y., Xia, Z., Ye, D., Guo, J., Tse, M. A., Panagiotou, G., & Xu, A. (2020b). Gut microbiome fermentation determines the efficacy of exercise for diabetes prevention. *Cell Metabolism*, *31*, 77–91.e5. doi:10.1016/j.cmet.2019.11.001.

Liu, Y., Yin, C., Deng, M.-M., Wang, Q., He, X.-Q., Li, M.-T., Li, C.-P., & Wu, H. (2019b). High expression of shmt2 is correlated with tumor progression and predicts poor prognosis in gastrointestinal tumors. *European Review for Medical & Pharmacological Sciences*, *23*, 9379–9392. doi:10.26355/eurrev_201911_19431.

Liu, Z., Wang, Y., Aizimuaji, Z., Ma, S., & Xiao, T. (2022b). Elevated foxa1 expression indicates poor prognosis in liver cancer due to its effects on cell proliferation and metastasis. *Disease Markers*, *2022*, 3317315. doi:10.1155/2022/3317315.

Llaurado Fernandez, M., Dawson, A., Kim, H., Lam, N., Russell, H., Bruce, M., Bittner, M., Hoenisch, J., Scott, S. A., Talhouk, A., Chiu, D., Provencher, D., Nourmoussavi, M., DiMattia, G., Lee, C.-H., Gilks, C. B., Köbel, M., & Carey, M. S. (2020). Hormone receptor expression and outcomes in low-grade serous ovarian carcinoma. *Gynecologic Oncology*, *157*, 12–20. doi:10.1016/j.ygyno.2019.11.029.

Long, X., Chen, L., Jiang, C., Zhang, L., & AlzheimerâĂŹs Disease Neuroimaging Initiative (2017). Prediction and classification of alzheimer disease based on quantification of mri deformation. *PLOS ONE*, *12*, e0173372. doi:10.1371/journal.pone.0173372.

Losi, L., Parenti, S., Ferrarini, F., Rivasi, F., Gavioli, M., Natalini, G., Ferrari, S., & Grande, A. (2011). Down-regulation of μ-protocadherin expression is a common event in colorectal carcinogenesis. *Human pathology*, *42*, 960–971. doi:10.1016/j.humpath.2010.10.009.

Lu, C., Fang, S., Weng, Q., Lv, X., Meng, M., Zhu, J., Zheng, L., Hu, Y., Gao, Y., Wu, X., Mao, J., Tang, B., Zhao, Z., Huang, L., & Ji, J. (2020a). Integrated analysis reveals critical glycolytic regulators in hepatocellular carcinoma. *Cell Communication and Signaling*, *18*, 97. doi:10.1186/s12964-020-00539-4.

Lu, J., Cai, S., Wang, F., Wu, P.-Y., Pan, X., Qiang, J., Li, H., & Zeng, M. (2023). Development of a prediction model for gross residual in high-grade serous ovarian cancer by combining preoperative assessments of abdominal and pelvic metastases and multiparametric mri. *Academic Radiology*, *30*, 1823–1831. doi:10.1016/j.acra.2022.12.019.

Lu, M., Fan, Z., Xu, B., Chen, L., Zheng, X., Li, J., Znati, T., Mi, Q., & Jiang, J. (2020b). Using machine learning to predict ovarian cancer. *International Journal of Medical Informatics*, *141*, 104195. doi:10.1016/j.ijmedinf.2020.104195.

Lu, T.-P., Kuo, K.-T., Chen, C.-H., Chang, M.-C., Lin, H.-P., Hu, Y.-H., Chiang, Y.-C., Cheng, W.-F., & Chen, C.-A. (2019). Developing a prognostic gene panel of epithelial ovarian cancer patients by a machine learning model. *Cancers*, *11*, 270. doi:10.3390/cancers11020270.

Luo, L., Zheng, Y., Lin, Z., Li, X., Li, X., Li, M., Cui, L., & Luo, H. (2021). Identification of shmt2 as a potential prognostic biomarker and correlating with immune infiltrates in lung adenocarcinoma. *Journal of Immunology Research*, *2021*, 6647122. doi:10.1155/2021/6647122.

Luo, Z., Wang, L., Shang, Z., Guo, Q., Liu, Q., Zhang, M., Li, T., Wang, Y., Zhang, Y., Zhang, Y., & Zhang, X. (2022). A panel of necroptosis-related genes predicts the prognosis of pancreatic adenocarcinoma. *Translational Oncology*, *22*, 101462. doi:10.1016/j.tranon.2022.101462.

Lurie, G., Thompson, P. J., McDuffie, K. E., Carney, M. E., & Goodman, M. T. (2009). Prediagnostic symptoms of ovarian carcinoma: a case-control study. *Gynecologic oncology*, *114*, 231–236. doi:10.1016/j.ygyno.2009.05.001.

Luyckx, M., Leblanc, E., Filleron, T., Morice, P., Darai, E., Classe, J.-M., Ferron, G., Stoeckle, E., Pomel, C., Vinet, B., Chereau, E., Bergzoll, C., & Querleu, D. (2012). Maximal cytoreduction in patients with figo stage iiic to stage iv ovarian, fallopian, and peritoneal cancer in day-to-day practice: A retrospective french multicentric study. *International Journal of Gynecologic Cancer*, *22*, 1337–1343. doi:10.1097/IGC.0b013e31826a3559.

Lv, G. Y., Yu, Y., An, L., Sun, X. D., & Sun, D. W. (2018). Preoperative plasma fibrinogen is associated with poor prognosis in esophageal carcinoma: a meta-analysis. *Clinical and Translational Oncology*, *20*, 853–861. doi:10.1007/s12094-017-1794-z.

Ma, C., Zhou, Y., Zhou, S., Zhao, K., Lu, B., & Sun, E. (2017). Preoperative peripheral plasma fibrinogen level is an independent prognostic marker in penile cancer. *Oncotarget*, *8*, 12355–12363. doi:10.18632/oncotarget.12563.

Ma, Y., Zhang, H., Li, X., & Liu, Y. (2022). Haglros promotes cell proliferation and angiogenesis and inhibits apoptosis by activating multiple signaling pathways in lscc cells. *Journal of Oral Pathology & Medicine*, *51*, 510–519. doi:10.1111/jop.13249.

Mackay, H. J., Brady, M. F., Oza, A. M., Reuss, A., Pujade-Lauraine, E., Swart, A. M., Siddiqui, N., Colombo, N., Bookman, M. A., Pfisterer, J., & du Bois, A. (2010). Prognostic relevance of uncommon ovarian histology in women with stage iii/iv epithelial ovarian cancer. *International Journal of Gynecologic Cancer*, *20*, 945–952. doi:10.1111/IGC.0b013e3181dd0110.

Mackenzie, R., Kommoss, S., Winterhoff, B. J., Kipp, B. R., Garcia, J. J., Voss, J., Halling, K., Karnezis, A., Senz, J., Yang, W., Prigge, E.-S., Reuschenbach, M., Doeberitz, M. V. K., Gilks, B. C., Huntsman, D. G., Bakkum-Gamez, J., McAlpine, J. N., & Anglesio, M. S. (2015). Targeted deep sequencing of mucinous ovarian tumors reveals multiple overlapping ras-pathway activating mutations in borderline and cancerous neoplasms. *BMC Cancer*, *15*, 415. doi:10.1186/s12885-015-1421-8.

Maeda, D., Mao, T.-L., Fukayama, M., Nakagawa, S., Yano, T., Taketani, Y., & Shih, I.-M. (2010). Clinicopathological significance of loss of arid1a immunoreactivity in ovarian clear cell carcinoma. *International Journal of Molecular Sciences*, *11*, 5120–5128. doi:10.3390/ijms11125120.

Manchana, T., Phoolcharoen, N., & Tantbirojn, P. (2019). Brca mutation in high grade epithelial ovarian cancers. *Gynecologic Oncology Reports*, *29*, 102–105. doi:10.1016/j.gore.2019.07.007.

Margolis, B., Dao, F., Licciardi, M., Misirlioglu, S., Olvera, N., Ramaswami, S., & Levine, D. A. (2021). Ccne1 amplification among metastatic sites in patients with gynecologic high-grade serous carcinoma. *Gynecologic Oncology Reports*, *37*, 100850. doi:https://doi.org/10.1016/j.gore.2021.100850.

Martin, K., Rahouti, M., Ayyash, M., & Alsmadi, I. (2022). Anomaly detection in blockchain using network representation and machine learning. *SECURITY AND PRIVACY*, *5*, e192. doi:https://doi.org/10.1002/spy2.192.

Mastropietro, A., Feldmann, C., & Bajorath, J. (2023). Calculation of exact shapley values for explaining support vector machine models using the radial basis function kernel. *Scientific Reports*, *13*, 19561. doi:10.1038/s41598-023-46930-2.

Mateescu, B., Batista, L., Cardon, M., Gruosso, T., De Feraudy, Y., Mariani, O., Nicolas, A., Meyniel, J.-P., Cottu, P., Sastre-Garau, X., & Mechta-Grigoriou, F. (2011). mir-141 and mir-200a act on ovarian tumorigenesis by controlling oxidative stress response. *Nature medicine*, *17*, 1627–1635. doi:10.1038/nm.2512.

Matulonis, U. A., Sood, A. K., Fallowfield, L., Howitt, B. E., Sehouli, J., & Karlan, B. Y. (2016). Ovarian cancer. *Nature reviews Disease primers*, *2*, 16061. doi:10.1038/nrdp.2016.61.

McCluggage, W. G. (2011). Morphological subtypes of ovarian carcinoma: a review with emphasis on new developments and pathogenesis. *Pathology*, *43*, 420–432. doi:10.1097/PAT.0b013e328348a6e7.

McConechy, M. K., Anglesio, M. S., Kalloger, S. E., Yang, W., Senz, J., Chow, C., Heravi-Moussavi, A., Morin, G. B., Mes-Masson, A.-M., Australian Ovarian Cancer Study Group, Carey, M. S., McAlpine, J. N., Kwon, J. S., Prentice, L. M., Boyd, N., Shah, S. P., Gilks, C. B., & Huntsman, D. G. (2011). Subtype-specific mutation of ppp2r1a in endometrial and ovarian carcinomas. *The Journal of Pathology*, *223*, 567–573. doi:https://doi.org/10.1002/path.2848.

McConechy, M. K., Ding, J., Senz, J., Yang, W., Melnyk, N., Tone, A. A., Prentice, L. M., Wiegand, K. C., McAlpine, J. N., Shah, S. P., Lee, C.-H., Goodfellow, P. J., Gilks, C. B., & Huntsman, D. G. (2014). Ovarian and endometrial endometrioid carcinomas have distinct ctnnb1 and pten mutation profiles. *Modern Pathology*, *27*, 128–134. doi:10.1038/modpathol.2013.107.

Mei, Y., Liu, H., Sun, X., Li, X., Zhao, S., & Ma, R. (2017). Plasma fibrinogen level may be a possible marker for the clinical response and prognosis of patients with breast cancer receiving neoadjuvant chemotherapy. *Tumor Biology*, *39*, 1–7. doi:10.1177/1010428317700002.

Menon, U., Gentry-Maharaj, A., Burnell, M., Singh, N., Ryan, A., Karpinskyj, C., Carlino, G., Taylor, J., Massingham, S. K., Raikou, M., Kalsi, J. K., Woolas, R., Manchanda, R., Arora, R., Casey, L., Dawnay, A., Dobbs, S., Leeson, S., Mould, T., Seif, M. W., Sharma, A., Williamson,

K., Liu, Y., Fallowfield, L., McGuire, A. J., Campbell, S., Skates, S. J., Jacobs, I. J., & Parmar, M. (2021). Ovarian cancer population screening and mortality after long-term follow-up in the uk collaborative trial of ovarian cancer screening (ukctocs): a randomised controlled trial. *The Lancet*, *397*, 2182–2193. doi:10.1016/S0140-6736(21)00731-5.

Meti, N., Saednia, K., Lagree, A., Tabbarah, S., Mohebpour, M., Kiss, A., Lu, F.-I., Slodkowska, E., Gandhi, S., Jerzak, K. J., Fleshner, L., Law, E., Sadeghi-Naini, A., & Tran, W. T. (2021). Machine learning frameworks to predict neoadjuvant chemotherapy response in breast cancer using clinical and pathological features. *JCO Clinical Cancer Informatics*, (pp. 66–80). doi:10.1200/CCI.20.00078.

Meyniel, J.-P., Cottu, P. H., Decraene, C., Stern, M.-H., Couturier, J., Lebigot, I., Nicolas, A., Weber, N., Fourchotte, V., Alran, S., Rapinat, A., Gentien, D., Roman-Roman, S., Mignot, L., & Sastre-Garau, X. (2010). A genomic and transcriptomic approach for a differential diagnosis between primary and secondary ovarian carcinomas in patients with a previous history of breast cancer. *BMC cancer*, *10*, 222. doi:10.1186/1471-2407-10-222.

Mir, S. E., De Witt Hamer, P. C., Krawczyk, P. M., Balaj, L., Claes, A., Niers, J. M., Van Tilborg, A. A. G., Zwinderman, A. H., Geerts, D., Kaspers, G. J. L., Vandertop, W. P., Cloos, J., Tannous, B. A., Wesseling, P., Aten, J. A., Noske, D. P., Van Noorden, C. J. F., & Würdinger, T. (2010). In silico analysis of kinase expression identifies wee1 as a gatekeeper against mitotic catastrophe in glioblastoma. *Cancer Cell*, *18*, 244–257. doi:10.1016/j.ccr.2010.08.011.

Mirza, M. R., Monk, B. J., Herrstedt, J., Oza, A. M., Mahner, S., Redondo, A., Fabbro, M., Ledermann, J. A., Lorusso, D., Vergote, I., Ben-Baruch, N. E., Marth, C., Mądry, R., Christensen, R. D., Berek, J. S., Dørum, A., Tinker, A. V., du Bois, A., González-Martín, A., Follana, P., Benigno, B., Rosenberg, P., Gilbert, L., Rimel, B. J., Buscema, J., Balser, J. P., Agarwal, S., & Matulonis, U. A. (2016). Niraparib maintenance therapy in platinum-sensitive, recurrent ovarian cancer. *New England Journal of Medicine*, *375*, 2154–2164. doi:10.1056/NEJMoa1611310.

Mittal, S., & Tyagi, S. (2019). Performance evaluation of machine learning algorithms for credit card fraud detection. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 320–324). doi:10.1109/CONFLUENCE.2019.8776925.

Mizuno, M., Kajiyama, H., Shibata, K., Mizuno, K., Kawai, M., Nagasaka, T., & Kikkawa, F. (2015). Prognostic value of histological type in stage iv ovarian carcinoma: a retrospective analysis of 223 patients. *British Journal of Cancer*, *112*, 1376–1383. doi:10.1038/bjc.2015.97.

Mizuno, M., Kajiyama, H., Shibata, K., Mizuno, K., Yamamuro, O., Kawai, M., Nakanishi, T., Nagasaka, T., & Kikkawa, F. (2012). Adjuvant chemotherapy for stage i ovarian clear cell carcinoma: Is it necessary for stage ia? *International Journal of Gynecological Cancer*, *22*, 1143–1149. doi:10.1097/IGC.0b013e31825c7cbe.

Mohseni, M., Cidado, J., Croessmann, S., Cravero, K., Cimino-Mathews, A., Wong, H. Y., Scharpf, R., Zabransky, D. J., Abukhdeir, A. M., Garay, J. P., Wang, G. M., Beaver, J. A., Cochran, R. L., Blair, B. G., M, R. D., Erlanger, B., Argani, P., Hurley, P. J., Lauring, J., & Park, B. H. (2014). Macrod2 overexpression mediates estrogen independent growth and tamoxifen resistance in breast cancers. *Proceedings of the National Academy of Sciences*, *111*, 17606–17611. doi:10.1073/pnas.1408650111.

Möller, C., Pijnenburg, Y. A. L., van der Flier, W. M., Versteeg, A., Tijms, B., de Munck, J. C., Hafkemeijer, A., Rombouts, S. A. R. B., van der Grond, J., van Swieten, J., Dopper, E., Scheltens, P., Barkhof, F., Vrenken, H., & Wink, A. M. (2016). Alzheimer disease and behavioral variant frontotemporal dementia: Automatic classification based on cortical atrophy for single-subject diagnosis. *Radiology*, *279*, 838–848. doi:10.1148/radiol.2015150220.

Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. *Technology and Health Care*, *24*, 31–42. doi:10.3233/THC-151071.

Montero, J. C., Chen, X., Ocaña, A., & Pandiella, A. (2012). Predominance of mtorc1 over mtorc2 in the regulation of proliferation of ovarian cancer cells: therapeutic implications. *Molecular cancer therapeutics*, *11*, 1342–1352. doi:10.1158/1535-7163.MCT-11-0723.

Moore, K., Colombo, N., Scambia, G., Kim, B.-G., Oaknin, A., Friedlander, M., Lisyanskaya, A., Floquet, A., Leary, A., Sonke, G. S., Gourley, C., Banerjee, S., Oza, A., González-Martín, A., Aghajanian, C., Bradley, W., Mathews, C., Liu, J., Lowe, E. S., Bloomfield, R., & DiSilvestro, P. (2018). Maintenance olaparib in patients with newly diagnosed advanced ovarian cancer. *New England Journal of Medicine*, *379*, 2495–2505. doi:10.1056/NEJMoa1810858.

Mueller, J. J., Schlappe, B. A., Kumar, R., Olvera, N., Dao, F., Abu-Rustum, N., Aghajanian, C., DeLair, D., Hussein, Y. R., Soslow, R. A., Levine, D. A., & Weigelt, B. (2018). Massively parallel sequencing analysis of mucinous ovarian carcinomas: genomic profiling and differential diagnoses. *Gynecologic Oncology*, *150*, 127–135. doi:10.1016/j.ygyno.2018.05.008.

Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, 292–299. doi:10.1016/j.procs.2020.01.047. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

Murakami, R., Matsumura, N., Brown, J., Higasa, K., Tsutsumi, T., Kamada, M., Abou-Taleb, H., Hosoe, Y., Kitamura, S., Yamaguchi, K., Abiko, K., Hamanishi, J., Baba, T., Koshiyama, M., Okuno, Y., Yamada, R., Matsuda, F., Konishi, I., & Mandai, M. (2017). Exome sequencing landscape analysis in ovarian clear cell carcinoma shed light on key chromosomal regions and mutation gene networks. *The American Journal of Pathology*, *187*, 2246–2258. doi:10.1016/j.ajpath.2017.06.012.

Nasioudis, D., Mastroyannis, S. A., Albright, B. B., Haggerty, A. F., Ko, E. M., & Latif, N. A. (2018). Adjuvant chemotherapy for stage i ovarian clear cell carcinoma: Patterns of use and outcomes. *Gynecologic Oncology*, *150*, 14–18. doi:10.1016/j.ygyno.2018.04.567.

National Cancer Institute (2017). Common terminology criteria for adverse events (ctcae) v5.0. https://ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/ctcae_v5_quick_ Accessed: 21-07-24.

National Cancer Institute (n.d.). Nci dictionary of cancer terms: Ojective response rate. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/objective-response-rate. Accessed: 21-07-24.

National Institute for Health and Care Excellence. (2011). Ovarian cancer: recognition and initial management [clinical guideline cg122]. https://www.nice.org.uk/guidance/cg122. Accessed: 03-3-23.

Nishiwada, S., Sho, M., Yasuda, S., Shimada, K., Yamato, I., Akahori, T., Kinoshita, S., Nagai, M., Konishi, N., & Nakajima, Y. (2015). Clinical significance of cd155 expression in human pancreatic cancer. *Anticancer research*, *35*, 2287–2297. URL: https://ar.iiarjournals.org/content/35/4/2287.

Niu, L., Qin, H.-Z., Xi, H.-Q., Wei, B., Xia, S.-Y., & Chen, L. (2015). Rnf43 inhibits cancer cell proliferation and could be a potential prognostic factor for human gastric carcinoma. *Cellular Physiology and Biochemistry*, *36*, 1835–1846. doi:10.1159/000430154.

Norquist, B. M., Harrell, M. I., Brady, M. F., Walsh, T., Lee, M. K., Gulsuner, S., Bernards, S. S., Casadei, S., Yi, Q., Burger, R. A., Chan, J. K., Davidson, S. A., Mannel, R. S., DiSilvestro, P. A., Lankes, H. A., Ramirez, N. C., King, M. C., Swisher, E. M., & Birrer, M. J. (2016). Inherited mutations in women with ovarian carcinoma. *JAMA Oncology*, *2*, 482–490. doi:10.1001/jamaoncol.2015.5495.

Noske, A., Henricksen, L. A., LaFleur, B., Zimmermann, A.-K., Tubbs, A., Singh, S., Storz, M., Fink, D., & Moch, H. (2015). Characterization of the 19q12 amplification including ccne1 and uri in different epithelial ovarian cancer subtypes. *Experimental and Molecular Pathology*, *98*, 47–54. doi:10.1016/j.yexmp.2014.12.004.

Oda, K., Hamanishi, J., Matsuo, K., & Hasegawa, K. (2018). Genomics to immunotherapy of ovarian clear cell carcinoma: Unique opportunities for management. *Gynecologic Oncology*, *151*, 381–389. doi:10.1016/j.ygyno.2018.09.001.

Okamoto, A., Glasspool, R. M., Mabuchi, S., Matsumura, N., Nomura, H., Itamochi, H., Takano, M., Takano, T., Susumu, N., Aoki, D., Konishi, I., Covens, A., Ledermann, J., Mezzazanica, D., Steer, C., Millan, D., Mcneish, I. A., Pfisterer, J., Kang, S., Gladieff, L., Bryce, J., & Oza, A. (2014). Gynecologic cancer intergroup (gcig) consensus review for clear cell carcinoma of the ovary. *International Journal of Gynecological Cancer*, *24*, S20–S25. doi:10.1097/IGC.0000000000000289.

Ore, R. M., Baldwin, L., Woolum, D., Elliott, E., Wijers, C., Chen, C.-Y., Miller, R. W., DeSimone, C. P., Ueland, F. R., Kryscio, R. J., van Nagell, J. R., & Pavlik, E. J. (2017). Symptoms relevant to surveillance for ovarian cancer. *Diagnostics*, *7*, 18. doi:10.3390/diagnostics7010018.

Oseledchyk, A., Leitao, M. M., Konner, J., O'Cearbhaill, R. E., Zamarin, D., Sonoda, Y., Gardner, G. J., Long Roche, K., Aghajanian, C. A., Grisham, R. N., Brown, C. L., Snyder, A., Chi, D. S., Soslow, R. A., Abu-Rustum, N. R., & Zivanovic, O. (2017). Adjuvant chemotherapy in patients with stage i endometrioid or clear cell ovarian cancer in the platinum era: a surveillance, epidemiology, and end results cohort study, 2000-2013. *Annals of Oncology*, *28*, 2985–2993. doi:10.1093/annonc/mdx525.

Otoom, A. F., Abdallah, E. E., Kilani, Y., Kefaye, A., & Ashour, M. (2015). Effective diagnosis and monitoring of heart disease. *International Journal of Software Engineering and Its Applications*, *9*, 143–156. URL: https://www.researchgate.net/publication/282747944_Effective_diagnosis_and_monitoring_of_heart_disease.

Oza, A. M., Cibula, D., Benzaquen, A. O., Poole, C., Mathijssen, R. H. J., Sonke, G. S., Colombo, N., Špaček, J., Vuylsteke, P., Hirte, H., Mahner, S., Plante, M., Schmalfeldt, B., Mackay, H., Rowbottom, J., Lowe, E. S., Dougherty, B., Barrett, J. C., & Friedlander, M. (2015). Olaparib combined with chemotherapy for recurrent platinum-sensitive ovarian cancer: a randomised phase 2 trial. *The Lancet Oncology*, *16*, 87–97. doi:10.1016/S1470-2045(14)71135-0.

Oza, A. M., Matulonis, U. A., Malander, S., Hudgens, S., Sehouli, J., del Campo, J. M., Berton-Rigaud, D., Banerjee, S., Scambia, G., Berek, J. S., Lund, B., Tinker, A. V., Hilpert, F., Vázquez, I. P., D'Hondt, V., Benigno, B., Provencher, D., Buscema, J., Agarwal, S., & Mirza, M. R. (2018). Quality of life in patients with recurrent ovarian cancer treated with niraparib versus placebo (engot-ov16/nova): results from a double-blind, phase 3, randomised controlled trial. *The Lancet Oncology*, *19*, 1117–1125. doi:10.1016/S1470-2045(18)30333-4.

Oza, A. M., Tinker, A. V., Oaknin, A., Shapira-Frommer, R., McNeish, I. A., Swisher, E. M., Ray-Coquard, I., Bell-McGuinn, K., Coleman, R. L., O'Malley, D. M., Leary, A., Chen, L.-m., Provencher, D., Ma, L., Brenton, J. D., Konecny, G. E., Castro, C. M., Giordano, H., Maloney, L., Goble, S., Lin, K. K., Sun, J., Raponi, M., Rolfe, L., & Kristeleit, R. S. (2017). Antitumor activity and safety of the parp inhibitor rucaparib in patients with high-grade ovarian carcinoma and a germline or somatic brca1 or brca2 mutation: Integrated analysis of data from study 10 and ariel2. *Gynecologic Oncology*, *147*, 267–275. doi:10.1016/j.ygyno.2017.08.022.

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Rajendra Acharya, U. (2020). Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in Biology and Medicine*, *121*, 103792. doi:10.1016/j.compbiomed.2020.103792.

Paik, E. S., Lee, J.-W., Park, J.-Y., Kim, J.-H., Kim, M., Kim, T.-J., Choi, C. H., Kim, B.-G., Bae, D.-S., & Seo, S. W. (2019). Prediction of survival outcomes in patients with epithelial ovarian cancer using machine learning methods. *Journal of Gynecologic Oncology*, *30*, e65. doi:10.3802/jgo.2019.30.e65.

Pal, T., Permuth-Wey, J., Betts, J. A., Krischer, J. P., Fiorica, J., Arango, H., LaPolla, J., Hoffman, M., Martino, M. A., Wakeley, K., Wilbanks, G., Nicosia, S., Cantor, A., & Sutphen, R. (2005). Brca1 and brca2 mutations account for a large proportion of ovarian carcinoma cases. *Cancer*, *104*, 2807–2816. doi:10.1002/cncr.21536.

Palacios, J., & Gamallo, C. (1998). Mutations in the β-Catenin Gene (CTNNB1) in Endometrioid Ovarian Carcinomas. *Cancer Research*, *58*, 1344–1347. URL: https://aacrjournals.org/cancerres/article/58/7/1344/505075/Mutations-in-the-Catenin-Gene-CTNNB1-in.

Pan, Z., Xu, T., Bao, L., Hu, X., Jin, T., Chen, J., Chen, J., Qian, Y., Lu, X., Li, L., Zheng, G., Zhang, Y., Zou, X., Song, F., Zheng, C., Jiang, L., Wang, J., Tan, Z., Huang, P., & Ge, M. (2022). Creb3l1 promotes tumor growth and metastasis of anaplastic thyroid carcinoma by remodeling the tumor microenvironment. *Molecular Cancer*, *21*, 190. doi:10.1186/s12943-022-01658-x.

Papanicolau-Sengos, A., Yang, Y., Pabla, S., Lenzo, F. L., Kato, S., Kurzrock, R., DePietro, P., Nesline, M., Conroy, J., Glenn, S., Chatta, G., & Morrison, C. (2019). Identification of targets for prostate cancer immunotherapy. *The Prostate*, *79*, 498–505. doi:10.1002/pros.23756.

Parra-Herran, C., Lerner-Ellis, J., Xu, B., Khalouei, S., Bassiouny, D., Cesari, M., Ismiil, N., & Nofech-Mozes, S. (2017). Molecular-based classification algorithm for endometrial carcinoma categorizes ovarian endometrioid carcinoma into prognostically significant groups. *Modern Pathology*, *30*, 1748–1759. doi:10.1038/modpathol.2017.81.

Parris, T. Z., Kovács, A., Aziz, L., Hajizadeh, S., Nemes, S., Semaan, M., Forssell-Aronsson, E., Karlsson, P., & Helou, K. (2014). Additive effect of the azgp1, pip, s100a8 and ube2c molecular biomarkers improves outcome prediction in breast carcinoma. *International Journal of Cancer*, *134*, 1617–1629. doi:10.1002/ijc.28497.

Pectasides, D., Fountzilas, G., Aravantinos, G., Kalofonos, H. P., Efstathiou, E., Salamalekis, E., Farmakis, D., Skarlos, D., Briasoulis, E., Economopoulos, T., & Dimopoulos, M. A. (2005). Advanced stage mucinous epithelial ovarian cancer: The hellenic cooperative oncology group experience. *Gynecologic Oncology*, *97*, 436–441. doi:10.1016/j.ygyno.2004.12.056.

Peng, P., Wu, W., Zhao, J., Song, S., Wang, X., Jia, D., Shao, M., Zhang, M., Li, L., Wang, L., Duan, F., Zhao, R., Yang, C., Wu, H., Zhang, J.,

Shen, Z., Ruan, Y., & Gu, J. (2016). Decreased expression of calpain-9 predicts unfavorable prognosis in patients with gastric cancer. *Scientific Reports*, *6*, 29604. doi:10.1038/srep29604.

Peres, L. C., Cushing-Haugen, K. L., Köbel, M., Harris, H. R., Berchuck, A., Rossing, M. A., Schildkraut, J. M., & Doherty, J. A. (2019). Invasive Epithelial Ovarian Cancer Survival by Histotype and Disease Stage. *JNCI: Journal of the National Cancer Institute*, *111*, 60–68. doi:10.1093/jnci/djy071.

Perren, T. J., Swart, A. M., Pfisterer, J., Ledermann, J. A., Pujade-Lauraine, E., Kristensen, G., Carey, M. S., Beale, P., Cervantes, A., Kurzeder, C., du Bois, A., Sehouli, J., Kimmig, R., Stähle, A., Collinson, F., Essapen, S., Gourley, C., Lortholary, A., Selle, F., Mirza, M. R., Leminen, A., Plante, M., Stark, D., Qian, W., Parmar, M. K. B., & Oza, A. M. (2011). A phase 3 trial of bevacizumab in ovarian cancer. *New England Journal of Medicine*, *365*, 2484–2496. doi:10.1056/NEJMoa1103799.

Peto, J., Gilham, C., Fletcher, O., & Matthews, F. E. (2004). The cervical cancer epidemic that screening has prevented in the uk. *Lancet*, *364*, 249–256. doi:10.1016/S0140-6736(04)16674-9.

Picot, N., Guerrette, R., Beauregard, A.-P., Jean, S., Michaud, P., Harquail, J., Benzina, S., & Robichaud, G. A. (2016). Mammaglobin 1 promotes breast cancer malignancy and confers sensitivity to anticancer drugs. *Molecular Carcinogenesis*, *55*, 1150–1162. doi:10.1002/mc.22358.

Pierson, W. E., Peters, P. N., Chang, M. T., Chen, L.-m., Quigley, D. A., Ashworth, A., & Chapman, J. S. (2020). An integrated molecular profile of endometrioid ovarian cancer. *Gynecologic Oncology*, *157*, 55–61. doi:10.1016/j.ygyno.2020.02.011.

Pisano, C., Greggi, S., Tambaro, R., Losito, S., Iodice, F., Di Maio, M., Ferrari, E., Falanga, M., Formato, R., Iaffaioli, V. R., & Pignata, S. (2005). Activity of chemotherapy in mucinous epithelial ovarian cancer: A retrospective study. *Anticancer Research*, *25*, 3501–3505. URL: https://ar.iiarjournals.org/content/25/5/3501.

Polterauer, S., Vergote, I., Concin, N., Braicu, I., Chekerov, R., Mahner, S., Woelber, L., Cadron, I., Gorp, T. V., Zeillinger, R., Castillo-Tong, D. C., & Sehouli, J. (2012). Prognostic value of residual tumor size in patients with epithelial ovarian cancer figo stages iia–iv: Analysis of the

ovcad data. *International Journal of Gynecologic Cancer*, *22*, 380–385. doi:10.1097/IGC.0b013e31823de6ae.

Porkka, K. P., Tammela, T. L. J., Vessella, R. L., & Visakorpi, T. (2004). Rad21 and kiaa0196 at 8q24 are amplified and overexpressed in prostate cancer. *Genes, Chromosomes and Cancer*, *39*, 1–10. doi:10.1002/gcc.10289.

Poropatich, K., Paunesku, T., Zander, A., Wray, B., Schipma, M., Dalal, P., Agulnik, M., Chen, S., Lai, B., Antipova, O., Maxey, E., Brown, K., Wanzer, M. B., Gursel, D., Fan, H., Rademaker, A., Woloschak, G. E., & Mittal, B. B. (2019). Elemental zn and its binding protein zinc-α2-glycoprotein are elevated in hpv-positive oropharyngeal squamous cell carcinoma. *Scientific Reports*, *9*, 16965. doi:10.1038/s41598-019-53268-1.

Porter, C. C., Kim, J., Fosmire, S., Gearheart, C. M., Van Linden, A., Baturin, D., Zaberezhnyy, V., Patel, P. R., Gao, D., Tan, A. C., & DeGregori, J. (2012). Integrated genomic analyses identify wee1 as a critical mediator of cell fate and a novel therapeutic target in acute myeloid leukemia. *Leukemia*, *26*, 1266–1276. doi:10.1038/leu.2011.392.

PosthumaDeBoer, J., Würdinger, T., Graat, H. C. A., Van Beusechem, V. W., Helder, M. N., Van Royen, B. J., & Kaspers, G. J. L. (2011). Wee1 inhibition sensitizes osteosarcoma to radiotherapy. *BMC Cancer*, *11*, 156. doi:10.1186/1471-2407-11-156.

Prajapati, G. L., & Patle, A. (2010). On performing classification using svm with radial basis and polynomial kernel functions. In *2010 3rd International Conference on Emerging Trends in Engineering and Technology* (pp. 512–515). doi:10.1109/ICETET.2010.134.

Previs, R. A., Kilgore, J., Craven, R., Broadwater, G., Bean, S., Wobker, S., DiFurio, M., Bae-Jump, V., Gehrig, P. A., & Secord, A. A. (2014). Obesity is associated with worse overall survival in women with low grade papillary serous epithelial ovarian cancer. *International Journal of Gynecological Cancer*, *24*, 670–675. doi:10.1097/IGC.0000000000000109.

Pujade-Lauraine, E., Hilpert, F., Weber, B., Reuss, A., Poveda, A., Kristensen, G., Sorio, R., Vergote, I., Witteveen, P., Bamias, A., Pereira, D., Wimberger, P., Oaknin, A., Mirza, M. R., Follana, P., Bollag, D., & Ray-Coquard, I. (2014). Bevacizumab combined with chemotherapy for platinum-resistant recurrent ovarian cancer: The aurelia open-label

randomized phase iii trial. *Journal of Clinical Oncology*, *32*, 1302–1308. doi:10.1200/JCO.2013.51.4489.

Pujade-Lauraine, E., Ledermann, J. A., Penson, R. T., Oza, A. M., Korach, J., Huzarski, T., Poveda, A., Pignata, S., Friedlander, M., & Colombo, N. (2017a). Treatment with olaparib monotherapy in the maintenance setting significantly improves progression-free survival in patients with platinum-sensitive relapsed ovarian cancer: results from the phase iii solo2 study. *Gynecologic Oncology*, *145*, 219–220. doi:10.1016/j.ygyno.2017.03.505.

Pujade-Lauraine, E., Ledermann, J. A., Selle, F., Gebski, V., Penson, R. T., Oza, A. M., Korach, J., Huzarski, T., Poveda, A., Pignata, S., Friedlander, M., Colombo, N., Harter, P., Fujiwara, K., Ray-Coquard, I., Banerjee, S., Liu, J., Lowe, E. S., Bloomfield, R., Pautier, P., Korach, J., Huzarski, T., Byrski, T., Pautier, P., & theSOLO2/ENGOT-Ov21 investigators. (2017b). Olaparib tablets as maintenance therapy in patients with platinum-sensitive, relapsed ovarian cancer and a brca1/2 mutation (solo2/engot-ov21): a double-blind, randomised, placebo-controlled, phase 3 trial. *The Lancet Oncology*, *18*, 1274–1284. doi:10.1016/S1470-2045(17)30469-2.

Qian, L., Li, L., Li, Y., Li, S., Zhang, B., Zhu, Y., & Yang, B. (2023). Lncrna hotair as a cerna is related to breast cancer risk and prognosis. *Breast Cancer Research and Treatment*, *200*, 375–390. doi:10.1007/s10549-023-06982-4.

Qian, L., Ren, J., Liu, A., Gao, Y., Hao, F., Zhao, L., Wu, H., & Niu, G. (2020). Mr imaging of epithelial ovarian cancer: a combined model to predict histologic subtypes. *European Radiology*, *30*, 5815–5825. doi:10.1007/s00330-020-06993-5.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics*, *32*, 496–501. doi:10.1038/ng1032.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

Rauh-Hain, J. A., Rodriguez, N., Growdon, W. B., Goodman, A. K., Boruta, D. M., Horowitz, N. S., del Carmen, M. G., & Schorge, J. O. (2012). Primary debulking surgery versus neoadjuvant chemotherapy in stage iv ovarian cancer. *Annals of Surgical Oncology*, *19*, 959–965. doi:10.1245/s10434-011-2100-x.

Rawal, S., Rawal, B., Shaheen, A., & Malik, S. (2017). Phishing detection in e-mails using machine learning. *International Journal of Applied Information Systems*, *12*, 21–24. URL: https://www.ijais.org/archives/volume12/number7/rawal-2017-ijais-451713.pdf.

Ray-Coquard, I., Pautier, P., Pignata, S., Pérol, D., González-Martín, A., Berger, R., Fujiwara, K., Vergote, I., Colombo, N., Mäenpää, J., Selle, F., Sehouli, J., Lorusso, D., Guerra Alía, E. M., Reinthaller, A., Nagao, S., Lefeuvre-Plesse, C., Canzler, U., Scambia, G., Lortholary, A., Marmé, F., Combe, P., de Gregorio, N., Rodrigues, M., Buderath, P., Dubot, C., Burges, A., You, B., Pujade-Lauraine, E., & Harter, P. (2019). Olaparib plus bevacizumab as first-line maintenance in ovarian cancer. *New England Journal of Medicine*, *381*, 2416–2428. doi:10.1056/NEJMoa1911361.

Riester, M., Wei, W., Waldron, L., Culhane, A. C., Trippa, L., Oliva, E., Kim, S.-h., Michor, F., Huttenhower, C., Parmigiani, G., & Birrer, M. J. (2014). Risk Prediction for Late-Stage Ovarian Cancer by Meta-analysis of 1525 Patient Samples. *JNCI: Journal of the National Cancer Institute*, *106*, dju048. doi:10.1093/jnci/dju048.

Risch, H. A., McLaughlin, J. R., Cole, D. E. C., Rosen, B., Bradley, L., Fan, I., Tang, J., Li, S., Zhang, S., Shaw, P. A., & Narod, S. A. (2006). Population BRCA1 and BRCA2 Mutation Frequencies and Cancer Penetrances: A Kin-Cohort Study in Ontario, Canada. *JNCI: Journal of the National Cancer Institute*, *98*, 1694–1706. doi:10.1093/jnci/djj465.

Risch, H. A., Mclaughlin, J. R., Cole, D. E. C., Rosen, B., Bradley, L., Kwan, E., Jack, E., Vesprini, D. J., Kuperstein, G., Abrahamson, J. L. A., Fan, I., Wong, B., & Narod, S. A. (2001). Prevalence and penetrance of germline brca1 and brca2 mutations in a population series of 649 women with ovarian cancer. *The American Journal of Human Genetics*, *68*, 700–710. doi:10.1086/318787.

Rodriguez-Romero, V., Bergstrom, R. F., Decker, B. S., Lahu, G., Vakilynejad, M., & Bies, R. R. (2019). Prediction of nephropathy in type 2 diabetes: An analysis of the accord trial applying machine learning techniques. *Clinical and Translational Science*, *12*, 519–528. doi:https://doi.org/10.1111/cts.12647.

Rosenthal, A. N., Fraser, L., Manchanda, R., Badman, P., Philpott, S., Mozersky, J., Hadwin, R., Cafferty, F. H., Benjamin, E., Singh, N., Evans, D. G., Eccles, D. M., Skates, S. J., Mackay, J., Menon, U., & Jacobs,

I. J. (2013). Results of annual screening in phase i of the united kingdom familial ovarian cancer screening study highlight the need for strict adherence to screening schedule. *Journal of Clinical Oncology*, *31*, 49–57. doi:10.1200/JCO.2011.39.7638.

Rosenthal, A. N., Fraser, L. S. M., Philpott, S., Manchanda, R., Burnell, M., Badman, P., Hadwin, R., Rizzuto, I., Benjamin, E., Singh, N., Evans, D. G., Eccles, D. M., Ryan, A., Liston, R., Dawnay, A., Ford, J., Gunu, R., Mackay, J., Skates, S. J., Menon, U., & Jacobs, I. J. (2017). Evidence of stage shift in women diagnosed with ovarian cancer during phase ii of the united kingdom familial ovarian cancer screening study. *Journal of Clinical Oncology*, *35*, 1411–1420. doi:10.1200/JCO.2016.69.9330.

Ryland, G. L., Hunter, S. M., Doyle, M. A., Rowley, S. M., Christie, M., Allan, P. E., Bowtell, D. D. L., Australian Ovarian Cancer Study Group, Gorringe, K. L., & Campbell, I. G. (2013). Rnf43 is a tumour suppressor gene mutated in mucinous tumours of the ovary. *The Journal of pathology*, *229*, 469–476. doi:10.1002/path.4134.

S K, S., & P, A. (2017). A machine learning ensemble classifier for early prediction of diabetic retinopathy. *Journal of Medical Systems*, *41*, 201. doi:10.1007/s10916-017-0853-x.

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from urls. *Expert Systems with Applications*, *117*, 345–357. doi:10.1016/j.eswa.2018.09.029.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*, e0118432. doi:10.1371/journal.pone.0118432.

Sakamoto, L. H. T., Andrade, R. V. d., Felipe, M. S. S., Motoyama, A. B., & Pittella Silva, F. (2014). Smyd2 is highly expressed in pediatric acute lymphoblastic leukemia and constitutes a bad prognostic factor. *Leukemia Research*, *38*, 496–502. doi:10.1016/j.leukres.2014.01.013.

Salamini-Montemurri, M., Lamas-Maceiras, M., Lorenzo-Catoira, L., Vizoso-Vázquez, A., Barreiro-Alonso, A., Rodríguez-Belmonte, E., Quindós-Varela, M., & Cerdán, M. E. (2023). Identification of lncrnas deregulated in epithelial ovarian cancer based on a gene expression profiling meta-analysis. *International Journal of Molecular Sciences*, *24*, 10798. doi:10.3390/ijms241310798.

Sans, M., Zhang, J., Lin, J. Q., Feider, C. L., Giese, N., Breen, M. T., Sebastian, K., Liu, J., Sood, A. K., & Eberlin, L. S. (2019). Performance of the masspec pen for rapid diagnosis of ovarian cancer. *Clinical Chemistry*, *65*, 674–683. doi:10.1373/clinchem.2018.299289.

Saravanan, R., & Sujatha, P. (2018). A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 945–949). doi:10.1109/ICCONS.2018.8663155.

Sato, D., Tsuchikawa, T., Mitsuhashi, T., Hatanaka, Y., Marukawa, K., Morooka, A., Nakamura, T., Shichinohe, T., Matsuno, Y., & Hirano, S. (2016). Stromal palladin expression is an independent prognostic factor in pancreatic ductal adenocarcinoma. *PLOS ONE*, *11*, e0152523. doi:10.1371/journal.pone.0152523.

Scelo, G., Muller, D. C., Riboli, E., Johansson, M., Cross, A. J., Vineis, P., Tsilidis, K. K., Brennan, P., Boeing, H., Peeters, P. H. M., Vermeulen, R. C. H., Overvad, K., Bueno-de Mesquita, H. B., Severi, G., Perduca, V., Kvaskoff, M., Trichopoulou, A., La Vecchia, C., Karakatsani, A., Palli, D., Sieri, S., Panico, S., Weiderpass, E., Sandanger, T. M., Nøst, T. H., Agudo, A., Quirós, J. R., Rodríguez-Barranco, M., Chirlaque, M.-D., Key, T. J., Khanna, P., Bonventre, J. V., Sabbisetti, V. S., & Bhatt, R. S. (2018). Kim-1 as a blood-based marker for early detection of kidney cancer: a prospective nested case–control study. *Clinical Cancer Research*, *24*, 5594–5601. doi:10.1158/1078-0432.CCR-18-1496.

Schilling, V., Beyerlein, P., & Chien, J. (2023). A bioinformatics analysis of ovarian cancer data using machine learning. *Algorithms*, *16*, 330. doi:10.3390/a16070330.

Schmeler, K. M., Sun, C. C., Bodurka, D. C., T. Deavers, M., Malpica, A., Coleman, R. L., Ramirez, P. T., & Gershenson, D. M. (2008). Neoadjuvant chemotherapy for low-grade serous carcinoma of the ovary or peritoneum. *Gynecologic Oncology*, *108*, 510–514. doi:10.1016/j.ygyno.2007.11.013.

Schwartz, D. R., Kardia, S. L. R., Shedden, K. A., Kuick, R., Michailidis, G., Taylor, J. M. G., Misek, D. E., Wu, R., Zhai, Y., Darrah, D. M., Reed, H., Ellenson, L. H., Giordano, T. J., Fearon, E. R., Hanash, S. M., & Cho, K. R. (2002). Gene expression in ovarian cancer reflects both morphology and biological behavior, distinguishing clear cell from

other poor-prognosis ovarian carcinomas. *Cancer research*, *62*, 4722–4729. URL: https://aacrjournals.org/cancerres/article/62/16/4722/509043/Gene-Expression-in-Ovarian-Cancer-Reflects-Both.

Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*, 345. doi:10.1007/s42979-020-00365-y.

Shan, J., Alam, S. K., Garra, B., Zhang, Y., & Ahmed, T. (2016). Computer-aided diagnosis for breast ultrasound using computerized bi-rads features and machine learning methods. *Ultrasound in Medicine and Biology*, *42*, 980–988. doi:10.1016/j.ultrasmedbio.2015.11.016.

Shankar, V. G., Sisodia, D. S., & Chandrakar, P. (2022). A novel discriminant feature selection-based mutual information extraction from mr brain images for alzheimer's stages detection and prediction. *International Journal of Imaging Systems and Technology*, *32*, 1172–1191. doi:10.1002/ima.22685.

Shattuck, D. W., Prasad, G., Mirza, M., Narr, K. L., & Toga, A. W. (2009). Online resource for validation of brain segmentation methods. *NeuroImage*, *45*, 431–439. doi:10.1016/j.neuroimage.2008.10.066.

Shen, Y., Katsaros, D., Loo, L. W. M., Hernandez, B. Y., Chong, C., Canuto, E. M., Biglia, N., Lu, L., Risch, H., Chu, W.-M., & Yu, H. (2015). Prognostic and predictive values of long non-coding rna linc00472 in breast cancer. *Oncotarget*, *6*, 8579–8592. doi:10.18632/oncotarget.3287.

Sheta, H., Abd El Hafez, A., Saif, M., Elsergany, A. R., Al Emam, D., & Abdelrazik, M. M. (2021). High foxa1 immunohistochemical expression level associates with mucinous histology, favorable clinico-pathological prognostic parameters and survival advantage in epithelial ovarian cancer. *Pathologica*, *113*, 102–114. doi:10.32074/1591-951x-217.

Shi, H., Hood, K. A., Hayes, M. T., & Stubbs, R. S. (2011). Proteomic analysis of advanced colorectal cancer by laser capture microdissection and two-dimensional difference gel electrophoresis. *Journal of Proteomics*, *75*, 339–351. doi:10.1016/j.jprot.2011.07.025.

Shi, X., Li, D., Wang, Y., Liu, S., Qin, J., Wang, J., Ran, J., Zhang, Y., Huang, Q., Liu, X., Zhou, J., & Liu, M. (2017). Discovery of centrosomal protein 70 as an important player in the development and progression of breast cancer. *The American Journal of Pathology*, *187*, 679–688. doi:10.1016/j.ajpath.2016.11.005.

Shibuya, Y., Tokunaga, H., Saito, S., Shimokawa, K., Katsuoka, F., Bin, L., Kojima, K., Nagasaki, M., Yamamoto, M., Yaegashi, N., & Yasuda, J. (2018). Identification of somatic genetic alterations in ovarian clear cell carcinoma with next generation sequencing. *Genes, Chromosomes and Cancer*, *57*, 51–60. doi:10.1002/gcc.22507.

Shih, I.-M., & Kurman, R. J. (2004). Ovarian tumorigenesis: A proposed model based on morphological and molecular genetic analysis. *American Journal of Pathology*, *164*, 1511–1518. doi:10.1016/S0002-9440(10)63708-X.

Shimada, M., Kigawa, J., Ohishi, Y., Yasuda, M., Suzuki, M., Hiura, M., Nishimura, R., Tabata, T., Sugiyama, T., & Kaku, T. (2009). Clinicopathological characteristics of mucinous adenocarcinoma of the ovary. *Gynecologic Oncology*, *113*, 331–334. doi:https://doi.org/10.1016/j.ygyno.2009.02.010.

Shu, C. A., Zhou, Q., Jotwani, A. R., Iasonos, A., Leitao, M. M., Konner, J. A., & Aghajanian, C. A. (2015). Ovarian clear cell carcinoma, outcomes by stage: The msk experience. *Gynecologic Oncology*, *139*, 236–241. doi:10.1016/j.ygyno.2015.09.016.

Shu, L., Guo, K., Lin, Z.-H., & Liu, H. (2022). Long non-coding rna haglros promotes the development of diffuse large b-cell lymphoma via suppressing mir-100. *Journal of Clinical Laboratory Analysis*, *36*, e24168. doi:10.1002/jcla.24168.

Sieh, W., Köbel, M., Longacre, T. A., Bowtell, D. D., deFazio, A., Goodman, M. T., Høgdall, E., Deen, S., Wentzensen, N., Moysich, K. B., Brenton, J. D., Clarke, B. A., Menon, U., Gilks, C. B., Kim, A., Madore, J., Fereday, S., George, J., Galletta, L., Lurie, G., Wilkens, L. R., Carney, M. E., Thompson, P. J., Matsuno, R. K., Kjær, S. K., Jensen, A., Høgdall, C., Kalli, K. R., Fridley, B. L., Keeney, G. L., Vierkant, R. A., Cunningham, J. M., Brinton, L. A., Yang, H. P., Sherman, M. E., García-Closas, M., Lissowska, J., Odunsi, K., Morrison, C., Lele, S., Bshara, W., Sucheston, L., Jimenez-Linan, M., Driver, K., Alsop, J., Mack, M., McGuire, V., Rothstein, J. H., Rosen, B. P., Bernardini, M. Q., Mackay, H., Oza, A., Wozniak, E. L., Benjamin, E., Gentry-Maharaj, A., Gayther, S. A., Tinker, A. V., Prentice, L. M., Chow, C., Anglesio, M. S., Johnatty, S. E., Chenevix-Trench, G., Whittemore, A. S., Pharoah, P. D. P., Goode, E. L., Huntsman, D. G., & Ramus, S. J. (2013). Hormone-receptor expression and ovarian cancer survival: an ovarian tumor tissue analysis consortium study. *The Lancet Oncology*, *14*, 853–862. doi:10.1016/S1470-2045(13)70253-5.

Simons, M., Ezendam, N., Bulten, J., Nagtegaal, I., & Massuger, L. (2015). Survival of patients with mucinous ovarian carcinoma and ovarian metastases: A population-based cancer registry study. *International Journal of Gynecologic Cancer*, *25*, 1208–1215. doi:10.1097/IGC.0000000000000473.

Simons, M., Massuger, L., Bruls, J., Bulten, J., Teerenstra, S., & Nagtegaal, I. (2017). Relatively poor survival of mucinous ovarian carcinoma in advanced stage: A systematic review and meta-analysis. *International Journal of Gynecologic Cancer*, *27*, 651–658. doi:10.1097/IGC.0000000000000932.

Singer, G., Kurman, R. J., Chang, H.-W., Cho, S. K. R., & Shih, I.-M. (2002). Diverse tumorigenic pathways in ovarian serous carcinoma. *The American Journal of Pathology*, *160*, 1223–1228. doi:10.1016/S0002-9440(10)62549-7.

Singer, G., Oldt III, R., Cohen, Y., Wang, B. G., Sidransky, D., Kurman, R. J., & Shih, I.-M. (2003). Mutations in braf and kras characterize the development of low-grade ovarian serous carcinoma. *Journal of the National Cancer Institute*, *95*, 484–486. doi:10.1093/jnci/95.6.484.

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science*, *132*, 1578–1585. doi:10.1016/j.procs.2018.05.122. International Conference on Computational Intelligence and Data Science.

Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*, *6*, 13. doi:10.1186/s40537-019-0175-6.

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*, 565–577. doi:10.1111/2041-210X.13140.

Sohn, S.-H., Sul, H. J., Kim, B., Kim, H. S., Kim, B. J., Lim, H., Kang, H. S., Soh, J. S., Kim, K. C., Cho, J. W., Seo, J., Koh, Y., & Zang, D. Y. (2021). Rnf43 and pwwp2b inhibit cancer cell proliferation and are predictive or prognostic biomarker for fda-approved drugs in patients with advanced gastric cancer. *Journal of Cancer*, *12*, 4616–4625. doi:10.7150/jca.56014.

Song, M. S., Salmena, L., & Pandolfi, P. P. (2012). The functions and regulation of the pten tumour suppressor. *Nature Reviews Molecular Cell Biology*, *13*, 283–296. doi:10.1038/nrm3330.

Song, Z., He, C.-D., Sun, C., Xu, Y., Jin, X., Zhang, Y., Xiao, T., Wang, Y., Lu, P., Jiang, Y., Wei, H., & Chen, H.-D. (2010). Increased expression of map2 inhibits melanoma cell proliferation, invasion and tumor growth in vitro and in vivo. *Experimental dermatology*, *19*, 958–964. doi:10.1111/j.1600-0625.2009.01020.x.

Sorayaie Azar, A., Babaei Rikan, S., Naemi, A., Bagherzadeh Mohasefi, J., Pirnejad, H., Bagherzadeh Mohasefi, M., & Wiil, U. K. (2022). Application of machine learning techniques for predicting survival in ovarian cancer. *BMC Medical Informatics and Decision Making*, *22*, 345. doi:10.1186/s12911-022-02087-y.

Sørensen, L., Igel, C., Liv Hansen, N., Osler, M., Lauritzen, M., Rostrup, E., Nielsen, M., & for the Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle Flagship Study of Ageing (2016). Early detection of alzheimer's disease using mri hippocampal texture. *Human Brain Mapping*, *37*, 1148–1161. doi:10.1002/hbm.23091.

Stavnes, H. T., Nymoen, D. A., Langerød, A., Holth, A., Børresen Dale, A.-L., & Davidson, B. (2013). Azgp1 and spdef mrna expression differentiates breast carcinoma from ovarian serous carcinoma. *Virchows Archiv*, *462*, 163–173. doi:10.1007/s00428-012-1347-3.

Stronach, E. A., Alfraidi, A., Rama, N., Datler, C., Studd, J. B., Agarwal, R., Guney, T. G., Gourley, C., Hennessy, B. T., Mills, G. B., Mai, A., Brown, R., Dina, R., & Gabra, H. (2011). HDAC4-Regulated STAT1 Activation Mediates Platinum Resistance in Ovarian Cancer. *Cancer Research*, *71*, 4412–4422. doi:10.1158/0008-5472.CAN-10-4111.

Stronach, E. A., Paul, J., Timms, K. M., Hughes, E., Brown, K., Neff, C., Perry, M., Gutin, A., El-Bahrawy, M., Steel, J. H., Liu, X., Lewsley, L.-A., Siddiqui, N., Gabra, H., Lanchbury, J. S., & Brown, R. (2018). Biomarker Assessment of HR Deficiency, Tumor BRCA1/2 Mutations, and CCNE1 Copy Number in Ovarian Cancer: Associations with Clinical Outcome Following Platinum Monotherapy. *Molecular Cancer Research*, *16*, 1103–1111. doi:10.1158/1541-7786.MCR-18-0034.

Su, W., Li, S., Chen, X., Yin, L., Ma, P., Ma, Y., & Su, B. (2017). Gabarapl1 suppresses metastasis by counteracting pi3k/akt pathway in prostate cancer. *Oncotarget*, *8*, 4449–4459. doi:10.18632/oncotarget.13879.

Sugiyama, T., Kamura, T., Kigawa, J., Terakawa, N., Kikuchi, Y., Kita, T., Suzuki, M., Sato, I., & Taguchi, K. (2000). Clinical characteristics of clear cell carcinoma of the ovary. *Cancer*, *88*, 2584–2589. doi:10.1002/1097-0142(20000601)88:11<2584::AID-CNCR22>3.0.CO;2-5.

Sujamol, S., Vimina, E. R., & Krishnakumar, U. (2021). Improving recurrence prediction accuracy of ovarian cancer using multi-phase feature selection methodology. *Applied Artificial Intelligence*, *35*, 206–226. doi:10.1080/08839514.2020.1854988.

Sun, C.-Y., Su, T.-F., Li, N., Zhou, B., Guo, E.-S., Yang, Z.-Y., Liao, J., Ding, D., Xu, Q., Lu, H., Meng, L., Wang, S.-X., Zhou, J.-F., Xing, H., Weng, D.-H., Ma, D., & Chen, G. (2016). A chemotherapy response classifier based on support vector machines for high-grade serous ovarian carcinoma. *Oncotarget*, *7*, 3245–3254. doi:10.18632/oncotarget.6569.

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, *23*, 687–719. doi:10.1142/S0218001409007326.

Suryawanshi, S., Vlad, A. M., Lin, H.-M., Mantia-Smaldone, G., Laskey, R., Lee, M., Lin, Y., Donnellan, N., Klein-Patel, M., Lee, T., Mansuria, S., Elishaev, E., Budiu, R., Edwards, R. P., & Huang, X. (2013). Plasma MicroRNAs as Novel Biomarkers for Endometriosis and Endometriosis-Associated Ovarian Cancer. *Clinical Cancer Research*, *19*, 1213–1224. doi:10.1158/1078-0432.CCR-12-2726.

Swisher, E. M., Lin, K. K., Oza, A. M., Scott, C. L., Giordano, H., Sun, J., Konecny, G. E., Coleman, R. L., Tinker, A. V., O'Malley, D. M., Kristeleit, R. S., Ma, L., Bell-McGuinn, K. M., Brenton, J. D., Cragun, J. M., Oaknin, A., Ray-Coquard, I., Harrell, M. I., Mann, E., Kaufmann, S. H., Floquet, A., Leary, A., Harding, T. C., Goble, S., Maloney, L., Isaacson, J., Allen, A. R., Rolfe, L., Yelensky, R., Raponi, M., & McNeish, I. A. (2017). Rucaparib in relapsed, platinum-sensitive high-grade ovarian carcinoma (ariel2 part 1): an international, multicentre, open-label, phase 2 trial. *The Lancet Oncology*, *18*, 75–87. doi:10.1016/S1470-2045(16)30559-9.

Takano, M., Kikuchi, Y., Yaegashi, N., Kuzuya, K., Ueki, M., Tsuda, H., Suzuki, M., Kigawa, J., Takeuchi, S., Moriya, T., & Sugiyama, T. (2006). Clear cell carcinoma of the ovary: a retrospective multicentre experience of 254 patients with complete surgical staging. *British Journal of Cancer*, *94*, 1369–1374. doi:10.1038/sj.bjc.6603116.

Tamir, A., Gangadharan, A., Balwani, S., Tanaka, T., Patel, U., Hassan, A., Benke, S., Agas, A., D'Agostino, J., Shin, D., Yoon, S., Goy, A., Pecora, A., & Suh, K. S. (2016). The serine protease prostasin (prss8) is a potential biomarker for early detection of ovarian cancer. *Journal of Ovarian Research*, *9*, 20. doi:10.1186/s13048-016-0228-9.

Tan, T. Z., Ye, J., Yee, C. V., Lim, D., Ngoi, N. Y. L., Tan, D. S. P., & Huang, R. Y.-J. (2019). Analysis of gene expression signatures identifies prognostic and functionally distinct ovarian clear cell carcinoma subtypes. *EBioMedicine*, *50*, 203–210. doi:10.1016/j.ebiom.2019.11.017.

Tang, L., Wei, R., Chen, R., Fan, G., Zhou, J., Qi, Z., Wang, K., Wei, Q., Wei, X., & Xu, X. (2022). Establishment and validation of a cholesterol metabolism-related prognostic signature for hepatocellular carcinoma. *Computational and Structural Biotechnology Journal*, *20*, 4402–4414. doi:10.1016/j.csbj.2022.07.030.

Tewari, K. S., Burger, R. A., Enserro, D., Norquist, B. M., Swisher, E. M., Brady, M. F., Bookman, M. A., Fleming, G. F., Huang, H., Homesley, H. D., Fowler, J. M., Greer, B. E., Boente, M., Liang, S. X., Ye, C., Bais, C., Randall, L. M., Chan, J. K., Ferriss, J. S., Coleman, R. L., Aghajanian, C., Herzog, T. J., DiSaia, P. J., Copeland, L. J., Mannel, R. S., Birrer, M. J., & Monk, B. J. (2019). Final overall survival of a randomized trial of bevacizumab for primary treatment of ovarian cancer. *Journal of Clinical Oncology*, *37*, 2317–2328. doi:10.1200/JCO.19.01009.

Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S., & Kuruwitaarachchi, N. (2019). Real-time credit card fraud detection using machine learning. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 488–493). doi:10.1109/CONFLUENCE.2019.8776942.

Tian, Z., Liu, Z., Fang, X., Cao, K., Zhang, B., Wu, R., Wen, X., Wen, Q., Shi, H., & Wang, R. (2020). ANP32A promotes the proliferation, migration and invasion of hepatocellular carcinoma by modulating the HMGA1/STAT3 pathway. *Carcinogenesis*, *42*, 493–506. doi:10.1093/carcin/bgaa138.

Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, *167*, 706–716. doi:10.1016/j.procs.2020.03.336. International Conference on Computational Intelligence and Data Science.

Timmermans, M., van der Aa, M., Lalisang, R. I., Witteveen, P. O., Van de Vijver, K. K., Kruitwagen, R. F., & Sonke, G. S. (2018). Interval between debulking surgery and adjuvant chemotherapy is associated with overall survival in patients with advanced ovarian cancer. *Gynecologic Oncology*, *150*, 446–450. doi:10.1016/j.ygyno.2018.07.004.

Tomasi Cont, N. (2015). Medical treatment of early stage and rare histological variants of epithelial ovarian cancer. *ecancermedicalscience*, *9*, 584. doi:10.3332/ecancer.2015.584.

Torheim, T., Malinen, E., Kvaal, K., Lyng, H., Indahl, U. G., Andersen, E. K. F., & Futsæther, C. M. (2014). Classification of dynamic contrast enhanced mr images of cervical cancers using texture analysis and support vector machines. *IEEE Transactions on Medical Imaging*, *33*, 1648–1656. doi:10.1109/TMI.2014.2321024.

Torre, L. A., Trabert, B., DeSantis, C. E., Miller, K. D., Samimi, G., Runowicz, C. D., Gaudet, M. M., Jemal, A., & Siegel, R. L. (2018). Ovarian cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, *68*, 284–296. doi:10.3322/caac.21456.

Tothill, R. W., Tinker, A. V., George, J., Brown, R., Fox, S. B., Lade, S., Johnson, D. S., Trivett, M. K., Etemadmoghadam, D., Locandro, B., Traficante, N., Fereday, S., Hung, J. A., Chiew, Y.-E., Haviv, I., Australian Ovarian Cancer Study Group, Gertig, D., deFazio, A., & Bowtell, D. D. L. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical cancer research*, *14*, 5198–5208. doi:10.1158/1078-0432.CCR-08-0196.

Tsang, Y. T., Deavers, M. T., Sun, C. C., Kwan, S.-Y., Kuo, E., Malpica, A., Mok, S. C., Gershenson, D. M., & Wong, K.-K. (2013). Kras (but not braf) mutations in ovarian serous borderline tumour are associated with recurrent low-grade serous carcinoma. *The Journal of Pathology*, *231*, 449–456. doi:10.1002/path.4252.

Tsao, H.-Y., Chan, P.-Y., & Su, E. C.-Y. (2018). Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. *BMC Bioinformatics*, *19*, 283. doi:10.1186/s12859-018-2277-0.

Tschoellitsch, T., Dünser, M., Böck, C., Schwarzbauer, K., & Meier, J. (2020). Machine Learning Prediction of SARS-CoV-2 Polymerase Chain

Reaction Results with Routine Blood Tests. *Laboratory Medicine*, *52*, 146–149. doi:10.1093/labmed/lmaa111.

Türke, C., Horn, S., Petto, C., Labudde, D., Lauer, G., & Wittenburg, G. (2017). Loss of heterozygosity in fancg, fancf and brip1 from head and neck squamous cell carcinoma of the oral cavity. *International Journal of Oncology*, *50*, 2207–2220. doi:10.3892/ijo.2017.3974.

Uehara, Y., Oda, K., Ikeda, Y., Koso, T., Tsuji, S., Yamamoto, S., Asada, K., Sone, K., Kurikawa, R., Makii, C., Hagiwara, O., Tanikawa, M., Maeda, D., Hasegawa, K., Nakagawa, S., Wada-Hiraike, O., Kawana, K., Fukayama, M., Fujiwara, K., Yano, T., Osuga, Y., Fujii, T., & Aburatani, H. (2015). Integrated copy number and expression analysis identifies profiles of whole-arm chromosomal alterations and subgroups with favorable outcome in ovarian clear cell carcinomas. *PLOS ONE*, *10*, e0128066. doi:10.1371/journal.pone.0128066.

Vanderstichele, A., Busschaert, P., Smeets, D., Landolfo, C., Van Nieuwenhuysen, E., Leunen, K., Neven, P., Amant, F., Mahner, S., Braicu, E. I., Zeilinger, R., Coosemans, A., Timmerman, D., Lambrechts, D., & Vergote, I. (2017). Chromosomal Instability in Cell-Free DNA as a Highly Specific Biomarker for Detection of Ovarian Cancer in Women with Adnexal Masses. *Clinical Cancer Research*, *23*, 2223–2231. doi:10.1158/1078-0432.CCR-16-1078.

Vapnik, V. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*, 988–999. doi:10.1109/72.788640.

Vapnik, V. N. (1998). Chapter 5. constructing learning algorithms. In *The nature of statistical learning theory* (p. 140). New York: Springer. (1st ed.).

Velmurugan, B. K., Yeh, K.-T., Lee, C.-H., Lin, S.-H., Chin, M.-C., Chiang, S.-L., Wang, Z.-H., Hua, C.-H., Tsai, M.-H., Chang, J.-G., & Ko, Y.-C. (2016). Acidic leucine-rich nuclear phosphoprotein-32a (anp32a) association with lymph node metastasis predicts poor survival in oral squamous cell carcinoma patients. *Oncotarget*, *7*, 10879–10890. doi:10.18632/oncotarget.7681.

Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using naive bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, *2*, 441–444. URL: https://www.ijiset.com/vol2/v2s9/IJISET_V2_I9_54.pdf.

Vergote, I., Tropé, C. G., Amant, F., Kristensen, G. B., Ehlen, T., Johnson, N., Verheijen, R. H. M., van der Burg, M. E. L., Lacave, A. J., Panici, P. B., Kenter, G. G., Casado, A., Mendiola, C., Coens, C., Verleye, L., Stuart, G. C. E., Pecorelli, S., & Reed, N. S. (2010). Neoadjuvant chemotherapy or primary surgery in stage iiic or iv ovarian cancer. *New England Journal of Medicine*, *363*, 943–953. doi:10.1056/NEJMoa0908806.

Wan, S., Xi, M., Zhao, H.-B., Hua, W., Liu, Y.-L., Zhou, Y.-L., Zhuo, Y.-J., Liu, Z.-Z., Cai, Z.-D., Wan, Y.-P., & Zhong, W.-D. (2019). Hmgcs2 functions as a tumor suppressor and has a prognostic impact in prostate cancer. *Pathology-Research and Practice*, *215*, 152464. doi:10.1016/j.prp.2019.152464.

Wang, J., Sahengbieke, S., Xu, X., Zhang, L., Xu, X., Sun, L., Deng, Q., Wang, D., Chen, D., Pan, Y. et al. (2018). Gene expression analyses identify a relationship between stanniocalcin 2 and the malignant behavior of colorectal cancer. *OncoTargets and therapy*, *11*, 7155–7168. doi:10.2147/OTT.S167780.

Wang, L., Londono, L. M., Cowell, J., Saatci, O., Aras, M., Ersan, P. G., Serra, S., Pei, H., Clift, R., Zhao, Q., Phan, K. B., Huang, L., LaBarre, M. J., Li, X., Shepard, H. M., Deaglio, S., Linden, J., Thanos, C. D., Sahin, O., & Cekic, C. (2021). Targeting adenosine with adenosine deaminase 2 to inhibit growth of solid tumors. *Cancer Research*, *81*, 3319–3332. doi:10.1158/0008-5472.can-21-0340.

Wang, M., Perucho, J. A. U., Hu, Y., Choi, M. H., Han, L., Wong, E. M. F., Ho, G., Zhang, X., Ip, P., & Lee, E. Y. P. (2022a). Computed Tomographic Radiomics in Differentiating Histologic Subtypes of Epithelial Ovarian Carcinoma. *JAMA Network Open*, *5*, e2245141–e2245141. doi:10.1001/jamanetworkopen.2022.45141.

Wang, N., Zhu, D., Liu, Y., Wu, J., Wang, M., Jin, S., Fu, F., Li, B., Ji, H., Du, C., & Zheng, Z. (2023). Nploc4 is a potential target and a poor prognostic signature in lung squamous cell carcinoma. *Scientific Reports*, *13*, 20430. doi:10.1038/s41598-023-47782-6.

Wang, Q., Li, X., Wang, Y., Qiu, J., Wu, J., He, Y., Li, J., Kong, Q., Han, J., & Jiang, Y. (2022b). Development and validation of a three-gene prognostic signature based on tumor microenvironment for gastric cancer. *Frontiers in Genetics*, *12*. doi:10.3389/fgene.2021.801240.

Wang, S., Hong, Q., Geng, X., Chi, K., Cai, G., & Wu, D. (2019a). Insulin-like growth factor binding protein 5-a probable target of kidney renal papillary renal cell carcinoma. *BioMed Research International*, *2019*, 3210324. doi:10.1155/2019/3210324.

Wang, W., Yu, D., & Zhong, M. (2019b). Lncrna haglros accelerates the progression of lung carcinoma via sponging microrna-152. *European Review for Medical & Pharmacological Sciences*, *23*, 6531–6538. doi:10.26355/eurrev_201908_18538.

Wang, Y., Martin, T. A., & Jiang, W. G. (2013). Havcr-1 expression in human colorectal cancer and its effects on colorectal cancer cells in vitro. *Anticancer research*, *33*, 207–214. URL: https://ar.iiarjournals.org/content/33/1/207.

Ward, A. K., Mellor, P., Smith, S. E., Kendall, S., Just, N. A., Vizeacoumar, F. S., Sarker, S., Phillips, Z., Alvi, R., Saxena, A., Vizeacoumar, F. J., Carlsen, S. A., & Anderson, D. H. (2016). Epigenetic silencing of creb3l1 by dna methylation is associated with high-grade metastatic breast cancers with poor prognosis and is prevalent in triple negative breast cancers. *Breast Cancer Research*, *18*, 12. doi:10.1186/s13058-016-0672-x.

Wei, J.-R., Dong, J., & Li, L. (2020). Cancer-associated fibroblasts-derived gamma-glutamyltransferase 5 promotes tumor growth and drug resistance in lung adenocarcinoma. *Aging*, *12*, 13220–13233. doi:10.18632/aging.103429.

Wiegand, K. C., Shah, S. P., Al-Agha, O. M., Zhao, Y., Tse, K., Zeng, T., Senz, J., McConechy, M. K., Anglesio, M. S., Kalloger, S. E., Yang, W., Heravi-Moussavi, A., Giuliany, R., Chow, C., Fee, J., Zayed, A., Prentice, L., Melnyk, N., Turashvili, G., Delaney, A. D., Madore, J., Yip, S., McPherson, A. W., Ha, G., Bell, L., Fereday, S., Tam, A., Galletta, L., Tonin, P. N., Provencher, D., Miller, D., Jones, S. J., Moore, R. A., Morin, G. B., Oloumi, A., Boyd, N., Aparicio, S. A., Shih, I.-M., Mes-Masson, A.-M., Bowtell, D. D., Hirst, M., Gilks, B., Marra, M. A., & Huntsman, D. G. (2010). Arid1a mutations in endometriosis-associated ovarian carcinomas. *New England Journal of Medicine*, *363*, 1532–1543. doi:10.1056/NEJMoa1008433.

Willis, S., Villalobos, V. M., Gevaert, O., Abramovitz, M., Williams, C., Sikic, B. I., & Leyland-Jones, B. (2016). Single gene prognostic biomarkers in ovarian cancer: A meta-analysis. *PLOS ONE*, *11*, e0149183. doi:10.1371/journal.pone.0149183.

Wimberger, P., Wehling, M., Lehmann, N., Kimmig, R., Schmalfeldt, B., Burges, A., Harter, P., Pfisterer, J., & du Bois, A. (2010). Influence of residual tumor on outcome in ovarian cancer patients with figo stage iv disease. *Annals of Surgical Oncology*, *17*, 1642–1648. doi:10.1245/s10434-010-0964-9.

Winter, W. E., Maxwell, G. L., Tian, C., Carlson, J. W., Ozols, R. F., Rose, P. G., Markman, M., Armstrong, D. K., Muggia, F., & McGuire, W. P. (2007). Prognostic factors for stage iii epithelial ovarian cancer: A gynecologic oncology group study. *Journal of Clinical Oncology*, *25*, 3621–3627. doi:10.1200/JCO.2006.10.2517.

Winterhoff, B., Hamidi, H., Wang, C., Kalli, K. R., Fridley, B. L., Dering, J., Chen, H.-W., Cliby, W. A., Wang, H.-J., Dowdy, S., Gostout, B. S., Keeney, G. L., Goode, E. L., & Konecny, G. E. (2016). Molecular classification of high grade endometrioid and clear cell ovarian cancer using tcga gene expression signatures. *Gynecologic Oncology*, *141*, 95–100. doi:10.1016/j.ygyno.2016.02.023.

Wong, K.-K., Tsang, Y. T., Deavers, M. T., Mok, S. C., Zu, Z., Sun, C., Malpica, A., Wolf, J. K., Lu, K. H., & Gershenson, D. M. (2010). Braf mutation is rare in advanced-stage low-grade ovarian serous carcinomas. *The American Journal of Pathology*, *177*, 1611–1617. doi:10.2353/ajpath.2010.100212.

Woodbeck, R., Kelemen, L. E., & Köbel, M. (2019). Ovarian endometrioid carcinoma misdiagnosed as mucinous carcinoma: an underrecognized problem. *International Journal of Gynecological Pathology*, *38*, 568–575. doi:10.1097/PGP.0000000000000564.

Wu, Y. H., Chang, T. H., Huang, Y. F., Huang, H. D., & Chou, C. Y. (2014). Col11a1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer. *Oncogene*, *33*, 3432–3440. doi:10.1038/onc.2013.307.

Xia, G., Wu, S., & Cui, X. (2023). An immune infiltration-related prognostic model of kidney renal clear cell carcinoma with two valuable markers: Capn12 and msc. *Frontiers in Oncology*, *13*. doi:10.3389/fonc.2023.1161666.

Xie, J., Ruan, S., Zhu, Z., Wang, M., Cao, Y., Ou, M., Yu, P., & Shi, J. (2021). Database mining analysis revealed the role of the putative h+/sugar transporter solute carrier family 45 in skin cutaneous melanoma. *Channels*, *15*, 496–506. doi:10.1080/19336950.2021.1956226.

Xie, M., Ji, Z., Bao, Y., Zhu, Y., Xu, Y., Wang, L., Gao, S., Liu, Z., Tian, Z., Meng, Q., Shi, H., & Yu, R. (2018). Phap1 promotes glioma cell proliferation by regulating the akt/p27/stathmin pathway. *Journal of Cellular and Molecular Medicine*, *22*, 3595–3604. doi:10.1111/jcmm.13639.

Xie, S., Qin, J., Liu, S., Zhang, Y., Wang, J., Shi, X., Li, D., Zhou, J., & Liu, M. (2016). Cep70 overexpression stimulates pancreatic cancer by inducing centrosome abnormality and microtubule disorganization. *Scientific Reports*, *6*, 21263. doi:10.1038/srep21263.

Xing, L., Mi, W., Zhang, Y., Tian, S., Zhang, Y., Qi, R., Zhang, C., & Lou, G. (2020). The identification of six risk genes for ovarian cancer platinum response based on global network algorithm and verification analysis. *Journal of Cellular and Molecular Medicine*, *24*, 9839–9852. doi:10.1111/jcmm.15567.

Xu, W., Chen, Z., Liu, G., Dai, Y., Xu, X., Ma, D., & Liu, L. (2021). Identification of a potential ppar-related multigene signature predicting prognosis of patients with hepatocellular carcinoma. *PPAR Research*, *2021*, 6642939. doi:10.1155/2021/6642939.

Xun, Q., Hu, C., Li, X., Hu, X., Qin, L., He, R., Lu, R., & Feng, J. (2019). Glcci1 rs37973 is associated with the response of adrenal hormone to inhaled corticosteroids in asthma. *World Allergy Organization Journal*, *12*, 100017. doi:10.1016/j.waojou.2019.100017.

Yan, W., Bai, Z., Wang, J., Li, X., Chi, B., & Chen, X. (2017). Anp32a modulates cell growth by regulating p38 and akt activity in colorectal cancer. *Oncology Reports*, *38*, 1605–1612. doi:10.3892/or.2017.5845.

Yang, B., Zhang, W., Zhang, M., Wang, X., Peng, S., & Zhang, R. (2020). Krt6a promotes emt and cancer stem cell transformation in lung adenocarcinoma. *Technology in Cancer Research & Treatment*, *19*. doi:10.1177/1533033820921248.

Yang, B. Y., Meng, Q., Sun, Y., Gao, L., & Yang, J. X. (2018a). Long non-coding rna snhg16 contributes to glioma malignancy by competitively binding mir-20a-5p with e2f1. *Journal of biological regulators and homeostatic agents*, *32*, 251–261. URL: https://pubmed.ncbi.nlm.nih.gov/29685003/.

Yang, J., Zhang, A., Luo, H., & Ma, C. (2022). Construction and validation of a novel gene signature for predicting the prognosis of osteosarcoma. *Scientific Reports*, *12*, 1279. doi:10.1038/s41598-022-05341-5.

Yang, M., Zhai, Z., Zhang, Y., & Wang, Y. (2019). Clinical significance and oncogene function of long noncoding rna haglros overexpression in ovarian cancer. *Archives of Gynecology and Obstetrics*, *300*, 703–710. doi:10.1007/s00404-019-05218-5.

Yang, S. Y. C., Lheureux, S., Karakasis, K., Burnier, J. V., Bruce, J. P., Clouthier, D. L., Danesh, A., Quevedo, R., Dowar, M., Hanna, Y., Li, T., Lu, L., Xu, W., Clarke, B. A., Ohashi, P. S., Shaw, P. A., Pugh, T. J., & Oza, A. M. (2018b). Landscape of genomic alterations in high-grade serous ovarian cancer from exceptional long- and short-term survivors. *Genome Medicine*, *10*, 81. doi:10.1186/s13073-018-0590-x.

Yang, X.-S., Wang, G.-X., & Luo, L. (2018c). Long non-coding rna snhg16 promotes cell growth and metastasis in ovarian cancer. *European Review for Medical & Pharmacological Sciences*, *22*, 616–622. URL: https://www.europeanreview.org/wp/wp-content/uploads/616-622.pdf.

Ye, J., He, H., Chen, S., Ren, Y., Guo, W., & Jin, Z. (2022a). Long non-coding rna nr2f2-as1 regulates human osteosarcoma growth and metastasis through mir-425-5p-mediated hmgb2. *International Journal of Clinical Oncology*, *27*, 1891–1903. doi:10.1007/s10147-022-02245-2.

Ye, Y., Yang, S., Han, Y., Sun, J., Xv, L., Wu, L., Wang, Y., & Ming, L. (2018). Linc00472 suppresses proliferation and promotes apoptosis through elevating pdcd4 expression by sponging mir-196a in colorectal cancer. *Aging*, *10*, 1523–1533. doi:10.18632/aging.101488.

Ye, Z., Zhang, Y., Liang, Y., Lang, J., Zhang, X., Zang, G., Yuan, D., Tian, G., Xiao, M., & Yang, J. (2022b). Cervical cancer metastasis and recurrence risk prediction based on deep convolutional neural network. *Current Bioinformatics*, *17*, 164–173. doi:10.2174/1574893616666210708143556.

Yeung, T.-L., Leung, C. S., Wong, K.-K., Gutierrez-Hartmann, A., Kwong, J., Gershenson, D. M., & Mok, S. C. (2017). Elf3 is a negative regulator of epithelial-mesenchymal transition in ovarian cancer cells. *Oncotarget*, *8*, 16951–16963. doi:10.18632/oncotarget.15208.

Yi, R., Feng, J., Yang, S., Huang, X., Liao, Y., Hu, Z., & Luo, M. (2018). mir-484/map2/c-myc-positive regulatory loop in glioma promotes tumor-initiating properties through erk1/2 signaling. *Journal of Molecular Histology*, *49*, 209–218. doi:10.1007/s10735-018-9760-9.

Yi, X., Liu, Y., Zhou, B., Xiang, W., Deng, A., Fu, Y., Zhao, Y., Ouyang, Q., Liu, Y., Sun, Z., Zhang, K., Li, X., Zeng, F., Zhou, H., & Chen,

B. T. (2021). Incorporating sulf1 polymorphisms in a pretreatment ct-based radiomic model for predicting platinum resistance in ovarian cancer treatment. *Biomedicine and Pharmacotherapy*, *133*, 111013. doi:10.1016/j.biopha.2020.111013.

Yoshino, H., Yamada, Y., Enokida, H., Osako, Y., Tsuruda, M., Kuroshima, K., Sakaguchi, T., Sugita, S., Tatarano, S., & Nakagawa, M. (2020). Targeting npl4 via drug repositioning using disulfiram for the treatment of clear cell renal cell carcinoma. *PLoS One*, *15*, e0236119. doi:10.1371/journal.pone.0236119.

Yu, K.-H., Hu, V., Wang, F., Matulonis, U. A., Mutter, G. L., Golden, J. A., & Kohane, I. S. (2020). Deciphering serous ovarian carcinoma histopathology and platinum response by convolutional neural networks. *BMC Medicine*, *18*, 236. doi:10.1186/s12916-020-01684-w.

Zaino, R. J., Brady, M. F., Lele, S. M., Michael, H., Greer, B., & Bookman, M. A. (2011). Advanced stage mucinous adenocarcinoma of the ovary is both rare and highly lethal. *Cancer*, *117*, 554–562. doi:10.1002/cncr.25460.

Zang, D., Zhang, C., Li, C., Fan, Y., Li, Z., Hou, K., Che, X., Liu, Y., & Qu, X. (2020). Lppr4 promotes peritoneal metastasis via sp1/integrin α/fak signaling in gastric cancer. *American journal of cancer research*, *10*, 1026–1044. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7136906/.

Zeng, C. X., Fu, S. B., Feng, W. S., Zhao, J. Y., Li, F. X., & Gao, P. (2019). Tcf19 enhances cell proliferation in hepatocellular carcinoma by activating the atk/foxo1 signaling pathway. *Neoplasma*, *66*, 46–53. doi:10.4149/neo_2018_171227N845.

Zeng, H., Chen, L., Zhang, M., Luo, Y., & Ma, X. (2021). Integration of histopathological images and multi-dimensional omics analyses predicts molecular features and prognosis in high-grade serous ovarian cancer. *Gynecologic Oncology*, *163*, 171–180. doi:10.1016/j.ygyno.2021.07.015.

Zeng, L.-J., Xiang, C.-L., Gong, Y.-Z., Kuang, Y., Lu, F.-F., Yi, S.-Y., Zhang, Y., & Liao, M. (2016). Neoadjuvant chemotherapy for patients with advanced epithelial ovarian cancer: A meta-analysis. *Scientific Reports*, *6*, 35914. doi:10.1038/srep35914.

Zhang, A. Y., Grogan, J. S., Mahon, K. L., Rasiah, K., Sved, P., Eisinger, D. R., Boulas, J., Vasilaris, A., Henshall, S. M., Stricker, P. D., Kench,

J. G., & Horvath, L. G. (2017a). A prospective multicentre phase iii validation study of azgp1 as a biomarker in localized prostate cancer. *Annals of Oncology*, *28*, 1903–1909. doi:10.1093/annonc/mdx247.

Zhang, F., Zhang, Y., Ke, C., Li, A., Wang, W., Yang, K., Liu, H., Xie, H., Deng, K., Zhao, W., Yang, C., Lou, G., Hou, Y., & Li, K. (2018). Predicting ovarian cancer recurrence by plasma metabolic profiles before and after surgery. *Metabolomics*, *14*, 65. doi:10.1007/s11306-018-1354-8.

Zhang, H., Mao, Y., Chen, X., Wu, G., Liu, X., Zhang, P., Bai, Y., Lu, P., Yao, W., Wang, Y., Yu, J., & Zhang, G. (2019). Magnetic resonance imaging radiomics in categorizing ovarian masses and predicting clinical outcome: a preliminary study. *European Radiology*, *29*, 3358–3371. doi:10.1007/s00330-019-06124-9.

Zhang, L., Lu, L., Nogues, I., Summers, R. M., Liu, S., & Yao, J. (2017b). Deeppap: Deep convolutional networks for cervical cell classification. *IEEE Journal of Biomedical and Health Informatics*, *21*, 1633–1643. doi:10.1109/JBHI.2017.2705583.

Zhang, Y., Li, L., Liu, R., & Zeng, C. (2020). Dna primase subunit 1 expression in hepatocellular carcinoma and its clinical implication. *BioMed Research International*, *2020*, 9689312. doi:10.1155/2020/9689312.

Zhao, H., Sun, Q., Li, L., Zhou, J., Zhang, C., Hu, T., Zhou, X., Zhang, L., Wang, B., Li, B., Zhu, T., & Li, H. (2019). High expression levels of aggf1 and mfap4 predict primary platinum-based chemoresistance and are associated with adverse prognosis in patients with serous ovarian cancer. *Journal of Cancer*, *10*, 397–407. doi:10.7150/jca.28127.

Zhao, J., Zheng, W., Wang, L., Jiang, H., Wang, X., Hou, J., Xu, A., & Cong, J. (2023). Human papillomavirus (hpv) integration signature in cervical lesions: identification of macrod2 gene as hpv hot spot integration site. *Archives of Gynecology and Obstetrics*, *307*, 1115–1123. doi:10.1007/s00404-022-06748-1.

Zheng, H., Jiang, W.-h., Tian, T., Tan, H.-s., Chen, Y., Qiao, G.-l., Han, J., Huang, S.-y., Yang, Y., Li, S., Wang, Z.-g., Gao, R., Ren, H., Xing, H., Ni, J.-s., Wang, L.-H., Ma, L.-j., & Zhou, W.-p. (2017). Cbx6 overexpression contributes to tumor progression and is predictive of a poor prognosis in hepatocellular carcinoma. *Oncotarget*, *8*, 18872–18884. doi:10.18632/oncotarget.14770.

Zhou, J., Li, L., Wang, L., Li, X., Xing, H., & Cheng, L. (2018a). Establishment of a svm classifier to predict recurrence of ovarian cancer. *Molecular Medicine Reports*, *18*, 3589–3598. doi:10.3892/mmr.2018.9362.

Zhou, J., Wu, S.-G., Wang, J., Sun, J.-Y., He, Z.-Y., Jin, X., & Zhang, W.-W. (2018b). The effect of histological subtypes on outcomes of stage iv epithelial ovarian cancer. *Frontiers in Oncology*, *8*. doi:10.3389/fonc.2018.00577.

Zhu, C., Zhu, J., Qian, L., Liu, H., Shen, Z., Wu, D., Zhao, W., Xiao, W., & Zhou, Y. (2021). Clinical characteristics and prognosis of ovarian clear cell carcinoma: a 10-year retrospective study. *BMC Cancer*, *21*, 322. doi:10.1186/s12885-021-08061-7.

Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of covid-19 diagnosis based on symptoms. *npj Digital Medicine*, *4*, 3. doi:10.1038/s41746-020-00372-6.

Zou, J., Li, Y., Liao, N., Liu, J., Zhang, Q., Luo, M., Xiao, J., Chen, Y., Wang, M., Chen, K., Zeng, J., & Mo, Z. (2022). Identification of key genes associated with polycystic ovary syndrome (pcos) and ovarian cancer using an integrated bioinformatics analysis. *Journal of Ovarian Research*, *15*, 30. doi:10.1186/s13048-022-00962-w.

Zou, K., Hu, Y., Li, M., Wang, H., Zhang, Y., Huang, L., Xie, Y., Li, S., Dai, X., Xu, W., Ke, Z., Gong, S., & Wang, Y. (2019). Potential role of hmgcs2 in tumor angiogenesis in colorectal cancer and its potential use as a diagnostic marker. *Canadian Journal of Gastroenterology and Hepatology*, *2019*, 8348967. doi:10.1155/2019/8348967.

Zou, Y., Wang, F., Liu, F.-Y., Huang, M.-Z., Li, W., Yuan, X.-Q., Huang, O.-P., & He, M. (2013). Rnf43 mutations are recurrent in chinese patients with mucinous ovarian carcinoma but absent in other subtypes of ovarian cancer. *Gene*, *531*, 112–116. doi:10.1016/j.gene.2013.08.054.

Appendix A

Tables

Table A.1: Grading System for Ovarian Cancer (Cancer Research UK, n.d.a).

Grade	Differentation	Cell Abnormality
I	Well differentiated	Cell looks like a normal cell
II	Moderately Differentiated	Cell looks less like a normal cell
III	Poorly Differentiated/Undifferentiated	Cell looks underdeveloped/not like a normal cell

Table A.2: FIGO staging of Ovarian Cancer. Adapted from Berek et al. (2021).

Stage I: Tumour confined to ovaries.	
IA	Tumour limited to 1 ovary (capsule intact); no tumour on ovarian surface; no malignant cells in the ascites or peritoneal washings.
IB	Tumour limited to both ovaries (capsules intact); no tumour on ovarian surface; no malignant cells in the ascites or the peritoneal washings.
IC	Tumour limited to 1 or more ovaries with any of following: IC1 | surgical spill IC2 | capsule ruptured before surgery or tumour on ovary surface IC3 | malignant cells in the ascites or peritoneal washings
Stage II: Tumour involves 1 or both ovaries with pelvic extension or peritoneal cancer.	
IIA	Extension and/or implants on uterus and/or fallopian tubes.
IIB	Extension to other pelvic intraperitoneal tissues.
Stage III: Tumour involves 1 or both ovaries or peritoneal cancer, with cytologically or histologically confirmed spread to the peritoneum outside the pelvis and/or metastasis to the retroperitoneal lymph nodes.	
IIIA1	Positive retroperitoneal lymph nodes only (cytologically or histologically proven): IIIA1(i) | metastasis up to 10mm in greatest dimension IIIA1(ii) | metastasis more than 10mm in greatest dimension
IIIA2	Microscopic extrapelvic peritoneal involvement with or without positive retroperitoneal lymph nodes.
IIIB	Macroscopic peritoneal metastasis beyond the pelvic up to 2cm in greatest dimension, with or without metastasis to the retroperitoneal lymph nodes.
IIIC	Macroscopic peritoneal metastasis beyond the pelvis more than 2cm in greatest dimension, with or without metastasis to the retroperitoneal lymph nodes (includes extension of tumour to capsule of liver and spleen without parenchymal involvement of either organ.)
Stage IV: Distant metastasis excluding peritoneal metastases.	
IVA	Pleural effusion with positive cytology.
IVB	Parenchymal metastases and metastases to extra-abdominal organs (including inguinal lymph nodes and lymph nodes outside of the abdominal cavity).

Appendix B

Copyright Statement