M.SC. BIOINFORMATICS

UNIVERSITY OF CRETE, MEDICINE DEPARTMENT

Master's Thesis

"Optimization of γ-OMP algorithm for feature selection in gene expression data"

Authors: Tzagkaraki Sofia

Supervisor: Tsagris Michail

Examiners: Nicolaou Christoforos, Pantazis Yannis

Table of Contents:

List of Figures

List of Tables

Acknowledgments

I would like to express my sincere thanks to Dr. Tsagris Michail, my thesis supervisor who guided me throughout the Master of Science program. I express my thanks to all the professors of the Bioinformatics graduate program of the University of Crete for their valuable help. I am grateful to the University of Crete for giving me the opportunity to pursue this master's degree. Finally, I would like to thank my family for their unwavering support throughout this journey.

Περίληψη

Με την ανακάλυψη της τεχνολογίας νέας γενιάς, ο σύγχρονος κόσμος της βιοτεχνολογίας πλέον μπορεί να αλληλουχήσει εκατοντάδες χιλιάδες γονίδια ή και ολόκληρο το γονιδίωμα σε σύντομο χρονικό διάστημα και με χαμηλό κόστος. Επίσης, μπορεί να εστιάσει στην χαρτογράφηση παραλλαγών/μεταλλάξεων που ενδεχομένως να έχουν κύρια βιολογική σημασία για την φυσιολογική λειτουργία του κυττάρου. Τέτοιου είδους αποτελέσματα, λοιπόν, έχουν μεγάλη ισχύ για τη διάγνωση της νόσου, την πρόγνωση, τη θεραπευτική απόφαση και την παρακολούθηση ασθενών. Αυτή η επαναστατική τεχνολογική δυνατότητα της μαζικής παράλληλης αλληλουχίας προσφέρει νέες ευκαιρίες για εξατομικευμένη ιατρική ακριβείας, ωστόσο δημιουργεί και μια νέα μεγάλη πρόκληση που αφορά την υπερπληθώρα δεδομένων.

Αναλυτικότερα, οι γονιδιωματικές μελέτες δημιουργούν τεράστιες ποσότητες δεδομένων, που συχνά περιλαμβάνουν χιλιάδες γονίδια σε πολλαπλά δείγματα. Ο εντοπισμός του κατάλληλου υποσυνόλου των γονιδίων που είναι πιο σχετικοί με την βιολογική καταστασή που μελέταται και την δίακριση της είναι απαραίτητος. Στα πλαίσια αυτής της μελέτης εστιάσαμε στην κατάλληλη επιλογή χαρακτηριστικών σε δεδομένα γονιδιακής έκφρασης που σχετίζονται με τον καρκίνο. Υλοποιήθηκε η δημιουργία ενός αλγορίθμου επιλογής χαρακτηριστικών (dcor-OMP) που αποτελεί μια παραλλαγή του αλγορίθμου γ-OMP. Επίσης, αξιολογήθηκε η απόδοσή του σε σχέση με τον προκάτοχό του, γ-OMP. Οι δύο αυτοί αλγόριθμοι είναι μια παραλλαγή του Orthogonal Matching Pursuit (OMP) που χρησιμοποιείται ευρέως στην επεξεργασία σήματος (Chen, Billings, & Luo, 1989; Davis, Mallat, & Zhang, 1994) και έχουν προσαρμοστεί ειδικά για ανάλυση δεδομένων γονιδιακής έκφρασης.

***Λέξεις-κλειδιά**: Συσχέτιση απόστασης, Επιλογή χαρακτηριστικών, Δεδομένα έκφρασης γονιδίου, αλγόριθμος γ-OMP, αλγόριθμος dcor-OMP, Τεχνολογία μικροσυστοιχιών, Αλληλουχία επόμενης γενιάς (NGS), Μηχανική μάθηση, Καμπύλη ROC, AUC (Περιοχή κάτω από την καμπύλη).*

Abstract

The advent of next-generation technologies revolutionized the field of biotechnology, allowing scientists to sequence hundreds of thousands of genes or an entire gene in a short amount of time. It can also focus on mapping variants/mutations that may have biological signaling implications for normal cell function. As a result, these findings have significant implications for illness diagnosis, prognosis, therapy decisions, and patient monitoring. This groundbreaking technological possibility of massively parallel sequencing opens up new opportunities for precision medicine, but it also introduces a new challenge: data overload.

First and foremost, genomics examines massive volumes of data, frequently including hundreds of genes in numerous samples. Identifying the proper selection of genes for identifying biological states is critical to understanding the underlying biological mechanisms. In this study, we looked at the most appropriate feature selection in cancer-related gene expression data. We developed a feature selection method (dcor-OMP), a version of the γ-OMP algorithm. The performance compared to the γ-OMP precursor was also assessed. Both of these methods are variations on the Orthogonal Matching Pursuit (OMP), which is widely used in signal processing (Chen, Billings, & Luo, 1989; Davis, Mallat, & Zhang, 1994), and have been specially customized for gene expression analysis.

Keywords: Distance Correlation, Feature Selection, Gene Expression Data, γ-OMP Algorithm, dcor-OMP Algorithm, Microarray Technology, Next-generation Sequencing (NGS), Machine Learning, ROC Curve, AUC (Area Under the Curve).

List of abbreviations

AUC	Area under the curve
CDFs	Cumulative distribution functions
cDNA	Complementary DNA
LR	Logistic regression
CSCs	Cancer stem cells
dCor	Distance correlation
dCov	Distance covariance
dVar	Distance variance
EFS	Ensemble feature selection
FPR	False positive rate
FS	Feature selection
GEO	Gene expression omnibus
GSE	Gene expression data series
HTS	High-throughput Sequencing
iid	Independent and identically distributed
k-NN	k-Nearest Neighbors
LOOCV	Leave-one-out cross validation
MCCD	Matched case control design
ML	Machine learning
MPFS	Matched-pairs feature selection
MRMD	Max-relevance-max-distance
mRMR	Minimum-Redundancy-Maximum-Relevancy
MSE	Mean squared error
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
NHGRI	National human genome research institute
OMP	Orthogonal matching pursuit
RNAseq	RNA sequencing
ROC	Receiver operating characteristic
RT	Reverse transcriptase
SSE	Sum of squares of errors
TMAs	Tissue microarrays
TPR	True positive rate
WGS	Whole genome sequencing

Chapter 1 Introduction

1.1. Understanding Cancer

Cancer has been understood since ancient times, as evidenced by fossilized bone tumors, Egyptian mummies, and old texts. The oldest description, from approximately 3000 BC in Egypt, references breast cancers that were treated with cauterization. Hippocrates later developed the name "cancer" in reference to tumors' crab-like appearance. During the Renaissance, scientific approaches enhanced cancer research, resulting in important advances in oncology. The nineteenth century witnessed more advancement with microscopes, allowing extensive analysis and assisting cancer surgery (American Cancer Society, 2021).

The word "cancer" or "(malignant) neoplasms" refers to a category of illnesses that originate at the cellular level. Cancer is a hereditary condition characterized by uncontrolled cell growth. Under normal conditions, somatic cells proliferate and multiply in a controlled way; nonetheless, cells can become aberrant and continue to expand. These aberrant cells can aggregate to create a tumor. If this uncontrolled growth is not immediately halted, it might spread, a situation known as metastasis, which can lead to serious clinical problems or even death (NCI, 2021). Cancer may afflict practically every bodily component and take several forms (Martel et al., 2020). Cancer is classified into five primary types: carcinoma, sarcoma, leukemia, lymphoma, and myeloma. Each cancer kind is unique, having its own genesis, symptoms, and treatment options (NCI, 2017).

Today, cancer is a major global health issue and one of the leading causes of death worldwide. In 2022, nearly 20 million new cancer cases were diagnosed, and there were approximately 9.7 million deaths attributed to cancer. Projections indicate that by 2040, these numbers will significantly increase, with new cancer cases expected to reach 29.9 million annually and cancer-related deaths anticipated to rise to 15.3 million per year. Generally, the highest cancer rates are observed in countries with high life expectancy, education levels, and standards of living. However, certain cancers, such as cervical cancer, exhibit higher incidence rates in countries where these measures are comparatively lower (Cancer Statistics - NCI, 2022). Lung cancer is the most prevalent cancer globally, with 2.5 million new cases, representing 12.4% of all new cancer cases. Following lung cancer, female breast cancer is the second most common, with 2.3 million cases, accounting for 11.6% of the total. Colorectal cancer ranks third with 1.9 million cases (9.6%), followed by prostate cancer with 1.5 million cases (7.3%), and stomach cancer with 970,000 cases (4.9%) (IARC, 2022).

1.2. The Genetic Background of Cancer

Cancer is caused by DNA alterations that develop as a result of either random mistakes during genome replication or carcinogens. Individual mutations are usually insufficient to induce cancer; but, the accumulation of genetic mutations in genes that govern cell growth and division can convert healthy cells into malignant ones. These mutations can be hereditary (genetic mutations inherited from parents) or acquired (mutations acquired during a person's lifetime due to environmental influences such as radiation or carcinogenic substances). The overwhelming majority of cancers occur randomly as a result of this process over time (The Genetics of Cancer, 2015). A variety of genetic changes have been related to the development of cancer. A DNA mutation, also known as a genetic variant, is a sort of modification that causes alterations to the DNA sequence. Some mutations only affect one nucleotide, hence a single nucleotide in DNA may be missing or replaced by another. These are referred to as point mutations. For example, roughly 5% of cancer patients have a point mutation in the KRAS gene, altering the nucleotide G to A, resulting in the development of an aberrant KRAS protein that causes uncontrolled cell proliferation. (Consortium et al., 2017; Guyon and Elisseeff, 2003). Genetic changes that cause cancer can also occur through rearrangements, deletions, or duplications of large segments of DNA, known as chromosomal rearrangements. For instance, chronic myeloid leukemia is often caused by a chromosomal rearrangement that joins the BCR gene with the ABL gene, creating the BCR-ABL protein that leads to uncontrolled cell growth (Hasty and Montagna, 2014). Some carcinogenic DNA changes occur outside of genes, in areas that act like "on" or "off" switches for nearby genes. Additionally, epigenetic changes, which are reversible and do not alter the DNA code but affect how DNA is packaged, can also cause cancer. Environmental factors such as cigarette smoke and the Epstein-Barr virus can induce both genetic and epigenetic changes (Abumsimir, Al-Qaisi and Kasmi, 2022).

It is generally accepted that driver gene mutations initiate the development of cancer. The most characteristic cases of such genes belong to two main categories: proto-oncogenes, which mutate into oncogenes leading to the development of cancer cells, and tumor suppressor genes, which, when inactivated, cause uncontrolled cell proliferation and consequently cancer development (Dressler *et al.*, 2022). Oncogenes are mutated forms of genes that cause normal cells to grow excessively and transform into cancer cells. These mutations involve specific genes within the cell known as proto-oncogenes (proto-oncogenes are denoted with the prefix “c”). Proto-oncogenes normally control how often a cell will divide and the degree to which it will differentiate. Proto-oncogenes, such as Ras, transcribe into products like growth factors, receptors, transcription factors, and signaling enzymes for cellular proliferation. Gain-of-function mutations in proto-oncogenes, resulting in dominant oncogenes that differ from their proto-oncogenes or are overexpressed, occur through point mutations, localized duplication, or chromosomal translocation. Consequently, when an oncogene mutates, it becomes permanently “stimulated,” causing the cell to divide very quickly, leading to cancer. This disrupts the normal activity of a cell and can lead to uncontrolled cell division and ultimately to cancer cells (Nourbakhsh *et al.*, 2024).

Tumor suppressor genes are normal genes whose primary function is to inhibit cell division. This inhibition helps in repairing DNA errors and regulating cell apoptosis. When these genes are lost or under expressed, it leads to uncontrolled cell division, which can eventually result in cancer. To date, around 30 tumor suppressor genes have been identified, including well-known examples such as p53, BRCA1, BRCA2, APC, and RB1. These genes play crucial roles in maintaining cellular integrity and preventing the formation of malignant tumors, thereby acting as essential safeguards against cancer development. The loss or under expression of these genes leads to uncontrolled cell division, which can ultimately result in the development of cancer (Tang *et al.*, 2021). A major difference between oncogenes and tumor suppressor genes is that oncogenes result from the activation of proto-oncogenes, whereas tumor suppressor genes lead to cancer when they are inactivated. Another key distinction is that the majority of oncogenes originate from mutations in normal cells (proto-oncogenes) acquired during an individual's lifetime, known as acquired mutations, while abnormalities or disruptions in tumor suppressor genes can be inherited from one's parents (Dressler *et al.*, 2022).

1.3. Oncogenes

Over the past forty years, scientific investigations have unequivocally shown the significance of oncogenes in human cancer. Numerous attempts have been made to comprehend the causal function of activated oncogenes in the formation of cancer since it was discovered that these genes are present in human tumors (Der et al., 1982; Goldfarb et al., 1982; Parada et al., 1982; Pulciani et al., 1982; Santos et al., 1982; Shih and Weinberg, 1982). Nevertheless, all of this research has demonstrated that oncogene expression is necessary for both the onset and progression of cancer, maintaining oncogenes as the primary targets for anti-cancer therapy. In genetically engineered mouse models, oncogenic expression is driven by tissue-specific promoters, resulting in high frequency tumors that regress when the inducing stimulus is turned off (Chin et al, 1999; Huettner et al, 2000; Boxer et al, 2004). This suggests that oncogenes are the Achilles' heel of cancers (Weinstein, 2002). This current concept of cancer is consistent with the fact that, in human tumors, all cancerous cells, regardless of cellular heterogeneity within the tumor, carry the same beginning oncogenic genetic abnormalities. Since the temporary silencing of the several separate tumor inducing oncogenes can induce cancer remission in these model systems, these data appear to suggest a homogeneous method of action for oncogenes inside cancer cells overall. Sadly, though, the treatments based on this cancer model are unable to completely eradicate human tumors. These clinical observations imply that oncogene-induced carcinogenesis in humans may not be reversible by means of the specific inactivation of the gene defect(s) that cause cancer to occur. What are the processes of tumor relapse, nevertheless, by which tumors change to become independent of oncogenes? (Vicente-Dueñas *et al.*, 2013)

These therapeutic failures cannot be explained just by the presence of cancer stem cells (CSCs) or the recognized cellular plasticity of tumors. Indeed, both factors imply that a genetically

homogeneous tumoral population can look phenotypically heterogeneous due to the presence of cells in various stages of development (Hanahan and Weinberg, 2011). The failure of targeted therapies in humans may indicate that oncogenes behave differently across cancer cells. This could explain why cancer cells respond differently to anti-oncogene therapy based on their stage. Recent in vivo genetic data has indicated that human oncogenes can transform early stem/precursor cells into specific differentiated tumor cell fates, although they are not necessary within malignant cells. These findings not only highlight a previously unknown role for human oncogenes, but also support a previously unmodeled carcinogenesis process in which the malignant phenotype is already programmed at the stem cell stage (Vicente-Dueñas *et al.*, 2013).

1.4. Microarray technology

Microarray is a recent advance in cancer research (Kim, Kang and Park, 2004) aiding pharmacological treatments for various diseases, including oral lesions. It allows the analysis of large sets of new and previously recorded samples and can detect specific tumor markers. Microarray technology enables simultaneous genotyping of thousands of loci, facilitating association and linkage studies to identify chromosomal regions associated with disease. It can detect chromosomal abnormalities in cancer, such as segments of allelic imbalance defined by loss of heterozygosity (Gilad *et al.*, 2000). Comparative genome-wide DNA hybridization helps identify amplified or deleted chromosomal regions, useful in cases such as oral cancer (Lilenbaum *et al.*, 2001). Gene microarray technology allows tens of thousands of DNA sequences to be deposited onto a small surface, usually a glass slide, known as a 'chip'. These DNA fragments are systematically arranged so that the identity of each fragment is determined by its position. The chip is then bathed with DNA or RNA isolated from a study sample (such as cells or tissue). Complementary base pairing between the sample and the chip-immobilized fragments produces light via fluorescence that can be detected using a specialized machine. Microarray technology can be used for various purposes in research and clinical studies, such as measuring gene expression and detecting specific DNA sequences ((National Human Genome Research Institute, , 2024). There are two main types of microarrays: gene expression microarrays and tissue microarrays (TMAs).

In contrast to techniques such as Northern blot and RT-PCR, which test only a few genes per experiment, microarray technology allows the simultaneous examination of many genes without bias from gene preselection (The Genetics of Cancer, 2015). In recent years, the scientific community has heavily utilized microarray technology, generating vast amounts of gene expression data. This data is scattered and not easily accessible. To address this, the National Center for Biotechnology Information (NCBI) created the Gene Expression Omnibus (GEO), a data storage facility that compiles gene expression data from various sources. This repository will serve as the input for our algorithm (PREMIER Biosoft, 2024).

The flowchart below (Figure 1) explains the process of comparing gene expression between normal and cancer cells using cDNA microarray technology. It starts with collecting cell samples (a), growing them in culture (b), and isolating mRNA from both normal and cancer cells (c). The

mRNA is then converted to cDNA, with normal cell cDNA labeled with green fluorescent probes and cancer cell cDNA with red fluorescent probes (d). These labeled cDNAs are hybridized to a microarray with oligo sequences and scanned to detect fluorescence (e). The colors indicate gene expression: gray means not present, yellow indicates presence in both cell types, green shows expression in normal cells only, and red indicates expression in cancer cells only (f).

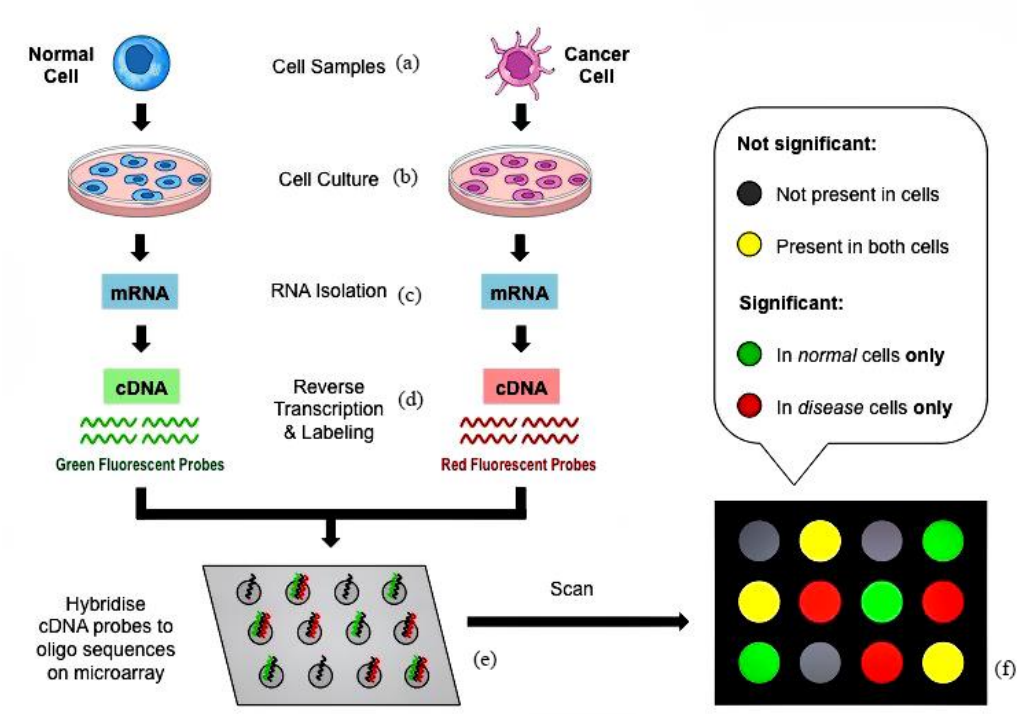

Figure 1. *Microarray Technique flow chart*

1.5. RNA-Seq (RNA Sequencing)

RNA sequencing (RNA-Seq) is among the most advanced and powerful technologies applied to study gene expression and genome function. RNA-Seq allows in-depth analysis of the RNA content within a biological sample with high-throughput NGS methods, which permits exact quantification and sequencing of RNA molecules. Major areas of improvement for RNA-Seq over classical approaches are the level at which it details a high-resolution snapshot of the transcriptome; it can describe the set of active genes and their levels in different conditions and time points (Wang, Gerstein, and Snyder, 2009). This technology can also identify previously unknown genes and transcripts, expanding our understanding of the genome (Mortazavi et al., 2008). In addition, RNA-Seq can detect alternative splicing events and post-transcriptional

modifications that are very important in understanding gene regulation and protein diversity. This kind of technology identifies gene fusions, mutations, and single nucleotide polymorphisms often coupled with diseases, reflecting further the versatility and importance of this technology. Mortazavi et al., 2008. That way, researchers can further compare gene expression profiles across different conditions, treatments, or time points, providing insight into the dynamic nature of the transcriptome. The other steps in the workflow of RNA-Seq are few but are of high importance. First, it begins with the isolation of RNA from biological samples, ensuring that it is of good quality and integrity. The isolated RNA is reverse-transcribed to complementary DNA (cDNA) by reverse transcriptase (RT) (Stark, Grzelak, and Hadfield, 2019). The cDNA is subsequently sequenced with high-throughput NGS technology, which produces massive amounts of sequence data (Ozsolak and Milos, 2011). Finally, bioinformatics tools are employed to analyze the sequencing results, such as mapping reads to a reference genome and predicting gene expression levels. RNA-Seq has changed genomics research by introducing a powerful and versatile approach for investigating the transcriptome. It enables researchers to analyze gene expression profiles, identify new transcripts, and decipher complex regulatory networks in unprecedented detail (Conesa et al., 2016). RNA-Seq, one of genomics' most advanced and effective tools, is still pushing research into gene expression and genome function. Its ability to provide comprehensive and precise data makes it an indispensable tool in modern biological research.

1.6. Next-generation Sequencing (NGS)

The introduction of Next-generation Sequencing (NGS), also known as High-throughput Sequencing (HTS), and Whole Genome Sequencing (WGS) technologies in genomics has revolutionized biological research by enabling the collection of massive amounts of genomic data in an unprecedented manner (Xuan et al., 2013). These technologies have transformed the field of genomics, providing rapid, scalable, and high-throughput sequencing capabilities. This data, derived from the genome and transcriptome, offers insights into the genetic makeup and gene expression patterns of organisms (Grafiati, 2021).

NGS has significantly advanced various fields, including personalized medicine, genetic diseases, clinical diagnostics, and microbiology. It has supplanted traditional sequencing methods by enabling the analysis of millions or even billions of sequences, offering a comprehensive view of the genetic landscape (Grafiati, 2021). The analysis of NGS data involves identifying genomic variations, studying gene expression, and understanding spatio-temporal dependencies within the data. Additionally, the development of computational methods for NGS data analysis is crucial to ensure accurate interpretation and extraction of meaningful biological information from the vast datasets (Eisele and Kappelmann-Fenzl, 2021). Integrating NGS data into research and clinical practice has opened new possibilities for understanding complex biological systems and developing innovative diagnostic and therapeutic strategies. These technological breakthroughs have allowed scientists to explore the molecular landscape of species, gaining vital insights into

their genetic makeup, disease causes, and evolutionary linkages (Mohammadi et al., 2023). However, the exponential expansion of NGS data poses significant challenges in extracting valuable knowledge from these enormous datasets (Dash et al., 2019). Machine learning, a subset of artificial intelligence, has emerged as a potent tool to address this difficulty by providing algorithms and methodologies for rapidly analyzing, interpreting, and extracting important information from biological NGS data (Yang et al., 2020).

In 2001, the completion of the human genome sequencing project marked a transformative milestone in scientific achievement, propelled by researchers using the Sanger DNA sequencing method (Lander *et al.*, 2001). Despite this accomplishment, the high costs and throughput limitations hindered the widespread application of DNA sequencing, especially in sequencing individuals' genomes. The initial estimate for sequencing the first human genome ranged from half to one billion dollars. Following the public release of the "completed" human genome (Nature, 2005), the National Human Genome Research Institute (NHGRI) invested $70 million in a DNA sequencing technology initiative aiming for a $1,000 human genome within a decade (Schloss, 2008). This commitment spurred significant advancements in HTS technologies, with platforms like Illumina, Ion Torrent, and PacBio emerging as pioneers, revolutionizing the landscape of molecular biology by facilitating rapid, cost-effective, and efficient sequencing of large volumes of genetic material.

HTS has overcome the limitations of the traditional Sanger sequencing method, which, despite reducing per-base costs by the end of the Human Genome Project, required a five-order magnitude reduction to reach the ambitious $1,000 genome threshold. Today, the cost of sequencing a genome (without interpretation) has dipped below $2,000, significantly closing this gap (Reuter, Spacek, and Snyder, 2015). HTS's innovative advantage lies not only in its affordability but also in its practicality for gene expression analysis and the detection of genetic diseases. This technology has enabled researchers to generate extensive datasets, fostering a more comprehensive understanding of genomic and transcriptional signatures across various diseases and developmental stages.

Within HTS technologies, whole exome sequencing has become instrumental in detecting new variants and mutations. RNA-Seq, an abbreviation for RNA sequencing, is a prominent technique utilizing NGS to provide profound insights into the presence and quantity of RNA in biological samples, thereby unraveling the continuously changing cellular transcriptome. The synergy of HTS technologies and bioinformatics tools has ushered in a new era of scientific inquiry, enabling researchers to delve into the intricate mechanisms behind gene expression profiles in healthy and diseased states. RNA-Seq extends beyond merely measuring gene expression levels to include the identification of novel transcripts, splice variants, and other non-coding RNAs. By leveraging NGS, RNA-Seq facilitates the generation of comprehensive snapshots of gene expression, regulatory networks, and diverse cellular activities within a given sample. Its high-throughput nature allows for the comparison of expression levels between different samples, facilitating the identification of differentially expressed genes or transcripts. RNA-Seq emerges as an

indispensable tool, not only for studying gene expression but also for unraveling the complexities of gene regulation and identifying novel transcripts and splice variants with profound implications for scientific and medical research (International Human Genome Sequencing Consortium, 2004; Reuter, Spacek, and Snyder, 2015)

1.7. Gene expression data

The rapid advancement in genomics has flooded biology with extensive information, ranging from genome sequences to high-throughput functional data, prompting a shift in focus towards understanding global biological mechanisms. Platforms like gene expression profiling and proteomics, especially utilizing high-density microarrays, have become pivotal in unraveling complex genetic puzzles (Bowtell, 1999). Despite the benefits, managing and extracting meaningful insights from the vast data pose significant challenges. Global efforts, such as the European Bioinformatics Institute's microarray database, aim to address these challenges by establishing common standards and large, publicly accessible gene expression databases. Simultaneously, the field is evolving towards more statistically robust methods, including predicting disease membership and modeling biological variables through gene expression data analysis (Dopazo *et al.*, 2001).

Gene expression data is often structured as a gene expression matrix, with columns representing samples or experiments and rows representing expression vectors for the genes queried by the microarray (Ravindran and Gunavathi, 2023).

Measure the level of mRNA transcription of protein-coding genes in a cell. The mRNA mix utilized in gene expression assays comes from biomaterials (samples) such tissues and cell lines. A microarray is often constructed to detect hundreds of unique target sequences linked with these genes using hybridization. The given measurements are only significant if something is known about the samples, target sequences, and genes. The primary purpose of gene expression data management is to combine expression data with sample and gene annotations, allowing users to explore, analyze, and understand expression data. A gene expression data management system usually integrates data from three different data sets: sample annotations, gene annotations, and gene expression measurements (Markowitz *et al.*, 2003). Below (Figure 2), is provided an overview which include the most widely used gene expression detection techniques for cancer tissue samples, beginning with tumor tissue.

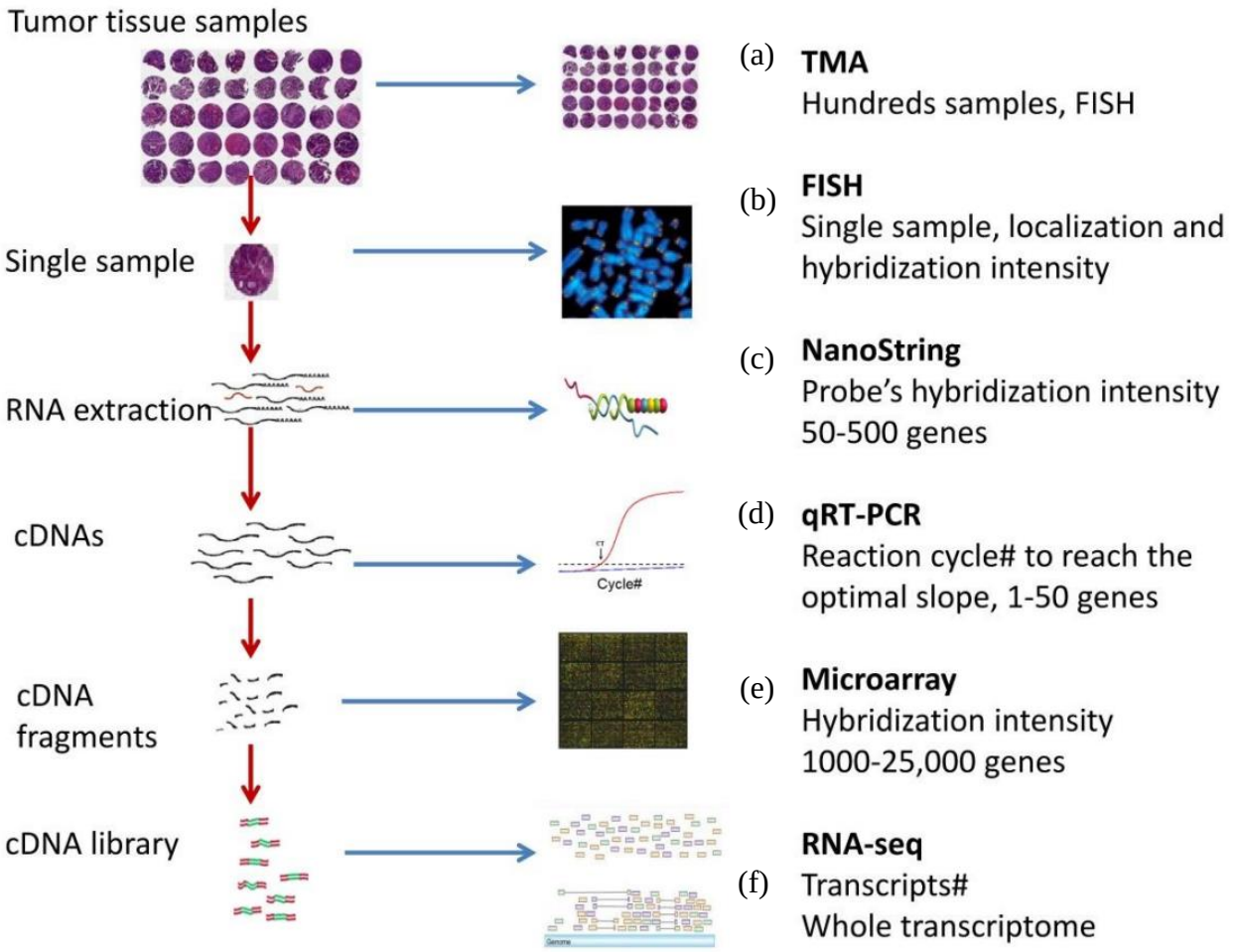

Figure 2. *Overview of the gene expression detection assays for cancer tissue sample by (Narrandes and Xu, 2018). (a) TMA (Tissue Microarray) can be used to evaluate hundreds of samples utilizing FISH. (b) FISH can be used to localize and quantify hybridization intensity in a single sample. (c) RNA is taken from the sample and examined using NanoString, which measures hybridization in 50-500 genes. (d) The RNA is subsequently transformed to cDNA for qRT-PCR, which measures 1-50 genes using reaction cycles. Furthermore, (e) cDNA fragments are employed in microarray analysis to detect 1,000-25,000 genes. Finally, (f) a cDNA library is generated for RNA-seq, allowing for a comprehensive investigation of the entire transcriptome through transcript quantification.*

.

Chapter 2 Machine Learning and Feature Selection

2.1. Machine learning (ML)

Machine learning technologies enable researchers to reveal hidden associations, identify biomarkers, classify samples, and generate predictions with high accuracy and efficiency by harnessing the inherent patterns and structures within the data (Marcos-Zambrano *et al.*, 2021). Machine learning is a branch of artificial intelligence that enables computers to learn from data and improve their performance without explicit programming. Machine learning can be used for various purposes, such as data analysis, pattern recognition, natural language processing, computer vision, robotics, and more. There are different types of machines learning techniques, depending on how the data is used and what kind of output is expected (Mahesh, 2019). On the first hand, Supervised learning involves training a model with labeled data, that is, data that has a known output or target variable. The model learns to map the input data to the output data and can then make predictions for new data. Examples of supervised learning algorithms are linear regression, classification, decision trees, and neural networks. Unsupervised learning on the other hand involves finding hidden patterns or structures in unlabeled data, that is, data that has no predefined output or target variable. The model learns to group or cluster the data based on some similarity or difference criteria and can then discover new insights or features from the data. Examples of unsupervised learning algorithms are clustering, dimensionality reduction, and word embedding. Moreover, Reinforcement learning involves learning from trial and error by interacting with an environment and receiving feedback or rewards for the actions taken. The model learns to optimize its behavior based on a goal or policy and can then adapt to changing situations or challenges. Examples of reinforcement learning algorithms are Q-learning, deep Q-networks, and policy gradients. Finally, Deep learning uses multiple layers of artificial neural networks to learn complex and nonlinear relationships between input and output data.

Deep learning can be applied to any type of machine learning technique, such as supervised, unsupervised, or reinforcement learning. Examples of deep learning applications are image recognition, natural language generation, speech synthesis, and generative adversarial networks (Mahesh, 2019; Ray, 2019). Machine learning is a rapidly evolving field that has many applications and challenges in various domains. Nevertheless, we will focus on the combination of machine learning techniques and biological NGS data that provide several significant benefits. First, it enables the discovery of complicated genomic variants and structural rearrangements that conventional approaches may overlook. Furthermore, machine learning allows for the detection of minor expression patterns and regulatory processes, which aids in the comprehension of gene regulation networks and disease pathways (Kalinin *et al.*, 2018). In addition, machine learning

algorithms may incorporate data from a variety of sources, including genomics, transcriptomics, and proteomics, allowing for a more comprehensive and holistic perspective of biological systems (Zitnik *et al.*, 2019). However, the appropriate application of machine learning in the processing of biological NGS data raises several issues and concerns (Yang *et al.*, 2020). These include the preprocessing and normalization of raw sequencing data, feature selection, handling high-dimensional datasets, addressing issues of class imbalance, ensuring model interpretability, and dealing with the ethical aspects of using sensitive genetic data. Thus, a preprocessing step is critical, since it plays a key role in improving analysis through the use of a technique known as Feature selection (FS).

2.2. Feature selection (FS)

2.2.1. What is FS?

FS refers to the process of selecting relevant features and at the same time rejecting irrelevant features from a dataset. It aids in the reduction of data complexity and dimensionality by reducing redundant and unimportant information, enabling the model to focus on the most important features (Ang et al., 2016). Combining pattern recognition algorithms with FS techniques has proven essential in many applications, as many of them were not initially intended to handle vast amounts of irrelevant information (Liu and Motoda, 2012; Guyon and Elisseeff, 2003).

There are several reasons that make FS an important tool, but the main ones are as follows, the first reason is that FS prevents overfitting and enhances model performance by increasing its accuracy, that is, prediction performance in supervised classification and improved cluster detection in clustering. The second reason why FS is chosen is because it provides faster and more affordable models by saving training time and reducing overfitting. The last reason is that FS obtains a better understanding of the underlying processes and trends that generated the data by identifying irrelevant features (Saeys, Inza and Larrañaga, 2007). For bioinformatics applications, FS methods must scale up to tens or even hundreds of thousands of features while maintaining good quality. Moreover, they must possess sufficient generality to manage continuous, censored time-to-event, multi-class, and binary outcomes. Furthermore, even if features are frequently simply continuous, an FS algorithm should include discrete characteristics as well to take clinical or other genetic factors into account (Tsagris et al., 2022).

2.2.2. FS methods

In the realm of ML, selecting the most relevant features for a classification model is crucial for enhancing performance and efficiency. For this reason, three basic strategies have been developed. FS strategies in the context of classification can be divided into three groups based on how they integrate the feature selection search with the model construction: filter methods, wrapper methods, and embedding methods (Yang *et al.*, 2020).

Filter techniques, which rely on heuristic scores and statistics, operate independently of ML algorithms. By figuring out the association between each feature and the target variable, features are chosen using this strategy. Integrated techniques: The learning algorithms that are employed to train the model serve as the foundation for these techniques. By being familiar with the feature weights, it is utilized to choose features throughout the training phase (Danasingh, Balamurugan and Epiphany, 2016). The main criterion for FS by ordering in filter methods is the application of variable ranking approaches. Because ranking systems are straightforward and have demonstrated strong results in real-world applications, they are widely employed. The variables are scored using a suitable ranking criterion, and variables that fall below the threshold are eliminated. Since ranking methods are used to exclude less important factors before categorization, they are filter methods (Chandrashekar and Sahin, 2014). Features are chosen using statistical measurements in the Filter technique. One benefit of this approach is that it requires less computational time because it selects the features as a pre-processing step, independent of the learning algorithm. A few of the most used statistical metrics for determining the significance of the features are variance threshold, information gain, chi-square test, Fisher score, and correlation coefficient. The Filter approach removes unnecessary columns from the models and identifies unimportant properties using the chosen measure. It provides the choice to isolate particular measures that improve a model. After the feature scores are calculated, the columns are ranked. However, the Filter approach has a few drawbacks, such as the potential for bias towards specific classes and the potential inability to detect non-linear connections between features. In addition, it takes a lot of time and might be challenging to use with big datasets (Saeys, Inza and Larrañaga, 2007; Venkatesh and Anuradha, 2019).

Conversely, the best subset of features is chosen utilizing **Wrapper techniques**, which employ various search tactics. This approach is predicated on assessing the correctness of the model using various feature combinations (Danasingh, Balamurugan and Epiphany, 2016). In addition, Wrapper technique evaluates the variable subset using the predictor's performance as the objective function and the predictor as a black box. Since evaluating subsets becomes an NP-hard job, suboptimal subsets are identified by utilizing search algorithms which choose a subset heuristically. Finding a subset of variables that maximizes the objective function the classification performance can be done using a variety of search strategies (Chandrashekar and Sahin, 2014). The Wrapper approach views feature set selection as a search problem, where many combinations are created, assessed, and contrasted with one another. Iteratively employing the subset of features,

the algorithm is trained. To assess a set of features and provide model performance ratings, a predictive model is utilized. The classifier determines how well the Wrapper technique performs. The classifier's output determines which subset of features is optimal. Because of its repeated procedure, the Wrapper technique is computationally intensive and therefore unsuitable for huge datasets. It can also take a lot of time and is prone to overfitting (Chandrashekar and Sahin, 2014; Venkatesh and Anuradha, 2019).

Lastly, the FS algorithm is incorporated into the learning algorithm in **Embedded techniques**, incorporating the benefits of both Filter and Wrapper methods by taking feature interaction and low computational cost into account. These techniques handle data quickly, much like the filter method, but they also have higher accuracy. The decision tree algorithm is the most widely used Embedded approach. This algorithm divides the data using a tree structure after beginning with all features. The resulting class is located at the leaf node of the tree. The most significant feature is chosen first, and then it is divided into two or more sub-trees according to that feature. Next, each feature is assessed using the Gini index or information gain. The best feature is then chosen for the following split after each feature is assessed using the information gain or Gini index (Saeys, Inza and Larrañaga, 2007; Venkatesh and Anuradha, 2019). Therefore, the goal of embedded techniques is to minimize the amount of time needed to compute the reclassification of various subsets, which is accomplished by wrapper methods. The primary strategy is to include feature selection in the training procedure (Pudil *et al.*, 1995; Alsberg *et al.*, 1998; Chuang *et al.*, 2008).

2.2.3. FS evaluation

To evaluate how efficient an FS method and algorithm are, data must be run through it and the performance measured. This can be accomplished by comparing the algorithm's output to baseline performance, such as a classification accuracy score or a regression score (Tsagris *et al.*, 2022). The FS literature is extensive, but in this work, we only concentrate on the most well-known algorithms that scale to the quantity of data and can generalize to a variety of outcomes and features. In terms of selection, the majority of these high-dimensional algorithms are greedy. We specifically concentrate on the Orthogonal Matching Pursuit (OMP) algorithm, which is widely recognized in the literature on signal processing (Shi *et al.*, 2019). OMP, is a widely used FS technique due to its computational efficiency and ease of use (Shi *et al.*, 2015). Because the next picked feature depends on the residual that is orthogonal to the previously selected feature, OMP, which was first used to choose features for binary-class classification, prefers to select only one among correlated features (Pati, Rezaiifar and Krishnaprasad, 1993).

2.3. Literature Review on FS in Gene Expression Data

FS in gene expression data is crucial for enhancing cancer detection and prognosis by reducing data dimensionality and improving model accuracy. Recent studies have proposed various methods to tackle this challenge. For instance, a study utilized SVMSMOTE for oversampling, followed by dimension reduction and classifier-based feature ranking, revealing that different feature selection methods significantly impact model performance, achieving up to 94.3% accuracy in classification tasks (Petinrin et al., 2023). Additionally, a graph theory-based feature selection approach used mutual information to construct undirected graphs, achieving robust classification results across various genetic datasets. The Ensemble Feature Selection (EFSmarker) method identified twelve critical biomarkers for breast cancer by integrating multiple filter techniques. This method proved effective in early cancer detection, improving classification accuracy to 96.2% (Li et al., 2023).

Another number of innovative techniques have been put forth recently in order to tackle his challenge, including the Max-Relevance-Max-Distance (MRMD) method by (Zou et al. 2005), which chooses features with strong correlation with the labeled and lowest redundancy features subset, and the minimum-Redundancy-Maximum-Relevancy (mRMR) method by (Peng et al. 2011), which chooses features using mutual information as a proxy for computing relevance and redundancy among features.

In addition to that, in high-dimensional data studies, particularly in gene function enrichment analysis, cancer biomarker discovery, and drug targeting identification in precision medicine, feature selection techniques are becoming increasingly important. A novel approach to predict TATA-binding proteins using feature selection and dimensionality reduction strategy was recently proposed by Zou (2016). Classifiers were developed to predict tumor originating sites after putting forth innovative selection procedures to find highly tissue-specific CpG sites (Tang et al. 2017).

OMP has been explored extensively for feature selection in gene expression. For example, a study demonstrated the application of a generalized OMP algorithm for feature selection, showing its scalability and effectiveness in handling high-dimensional data, achieving an accuracy improvement of 12% over traditional methods (Wang & Ye, 2013). Another research applied OMP in a minimum redundancy feature selection context, achieving notable improvements in classifying microarray gene expression data, with accuracy rates reaching 93% (Sun & Qian, 2018).

Matched-Pairs Feature Selection (MPFS) techniques and approaches for bioinformatics research abound. It has been possible for several researchers to incorporate paired data which fall into three categories into their algorithms (Liang *et al.*, 2018). In order to increase model predictive accuracy, a classification strategy is frequently used after the test statistic ranks pertinent characteristics by evaluating significant levels using the original and modified paired t-test. These kinds of approaches can yield preliminary feature selection results, but they are quite time-consuming. Second, a modeling technique that is frequently employed in Matched case-control design (MCCD) investigations to find features strongly linked with case-control status is conditional

logistic regression (CLR). When possible, correlations exist, CLR takes into account how attributes interact with one another to produce superior selection outcomes (BRESLOW *et al.*, 1978).

In Alzheimer's disease research, a hybrid gene selection pipeline combined filter, wrapper, and unsupervised methods, leading to improved classification using deep learning techniques, achieving classification accuracy of 89.4% (Liu et al., 2023).by Another approach integrated XGBoost and multi-objective optimization, resulting in higher accuracy for colon cancer gene expression data, achieving an average accuracy of 95.1% across multiple datasets (Deng et al., 2021). Further studies emphasized hybrid evolutionary approaches, such as integrating Particle Swarm Optimization and Correlation-based FS, which demonstrated enhanced performance in cancer classification tasks, achieving an accuracy improvement of 10-15% over baseline models (Jain et al., 2023). An innovative hybrid FS technique using a micro–Genetic Algorithm on microarray gene expression data significantly improved classification accuracy, reaching up to 91.8% and reducing complexity by 20% (Pragadeesh et al., 2019). Lastly, a comprehensive survey on hybrid FS methods highlighted the integration of bio-inspired metaheuristic and wrapper methods for better classification of breast cancer, showcasing significant advancements in FS techniques, with improvements in accuracy by up to 13% (Naeem et al., 2022).

In conclusion, the reviewed literature highlights the critical importance of FS in processing high-dimensional gene expression data for cancer classification. Advanced techniques such as hybrid models, ensemble methods, and graph theory-based approaches have demonstrated significant improvements in accuracy, efficiency, and robustness. These methods effectively address common challenges such as class imbalance, data redundancy, and noise, thereby enhancing the predictive performance of machine learning models. This body of research aligns closely with the objectives of my thesis, which focuses on optimizing the γ-OMP algorithm for FS in gene expression data. The insights gained from these studies reinforce the relevance of developing sophisticated feature FS methods that can handle the complexity and high dimensionality of gene expression data. By integrating and extending these advanced techniques, my research aims to contribute to more accurate and efficient diagnostic tools for cancer, ultimately improving patient outcomes and advancing the field of bioinformatics.

2.4. AUC-area under the curve

The area under the ROC (Receiver Operating Characteristic) curve, or simply AUC, is a method widely used in statistics and ML for evaluating binary classifiers. It refers to the ROC curve, which is a probability curve, and the AUC represents the degree or measure of separability. In other words, it shows how well the model is capable of distinguishing between classes. The ROC curve is plotted with the TPR (True Positive Rate) on the y-axis against the FPR (False Positive Rate) on the x-axis at various threshold settings (Classification: ROC Curve and AUC, 2022).

- **True Positive Rate (TPR)**: Also known as sensitivity or recall, it is the ratio of correctly predicted positive observations to the actual positives.

$$\frac{TP}{N} = \frac{TP}{TP + TN}$$

- **False Positive Rate (FPR)**: It is the ratio of incorrectly predicted positive observations to the actual negatives.

$$\frac{FP}{N} = \frac{FP}{FP + TN}$$

The higher the AUC, the better the model is at making accurate predictions. AUC, or Area Under the Curve, provides a comprehensive measure of a model's performance across all possible classification thresholds. An AUC of 1 indicates a perfect model that correctly classifies all positive and negative instances. This means that the model has an excellent ability to distinguish between the two classes. An AUC between 0.5 and 1 suggests that the model performs better than random guessing, indicating a decent level of predictive power. Conversely, an AUC of 0.5 means that the model's performance is no better than random guessing, signifying no discriminatory power. If the AUC is below 0.5, it implies that the model performs worse than random guessing, effectively classifying instances in the opposite manner of what is desired. Thus, AUC is a crucial metric for evaluating the effectiveness of a model and understanding its predictive capabilities.

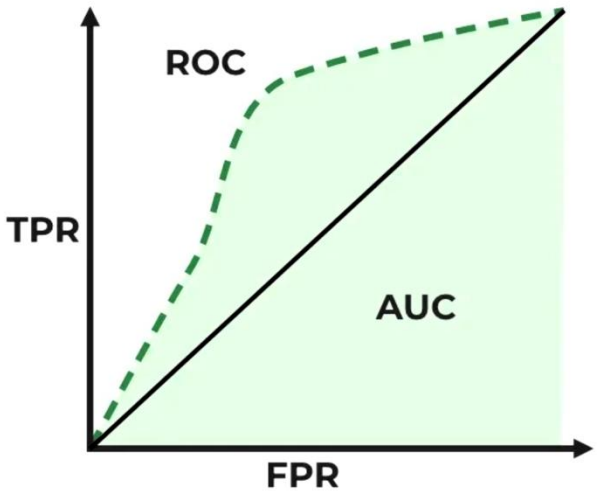

Figure 3. *ROC-AUC Classification Evaluation Metric (*AUC ROC Curve in Machine Learning, *GeeksforGeeks ,2020)*

Overall, AUC is a fundamental aspect of model evaluation in machine learning, contributing to the development of robust and reliable classifiers. It is an important tool for researchers and engineers to better understand the performance of their models and to choose the best models for their needs. It enables a deeper understanding of model performance, facilitates decision-making, and ultimately helps select the most appropriate models to address specific challenges and needs.

2.5. Cross Validation

Cross-validation (CV) is a statistical technique for assessing and comparing learning algorithms by splitting the data into two segments, one for training the model and the other for validating it. In standard CV, the training and validation sets overlap in successive rounds to ensure that each data point is used for validation at least once. The simplest form is k-fold CV, where the data is divided into k subsets, and the model is trained and validated k times, each time using a different subset as the validation set and the remaining k-1 subsets as the training set. Other forms of CV are either variations of k-fold CV or involve multiple rounds of k-fold CV (Refaeilzadeh, Tang, and Liu, 2016).

In biotechnology CV is a highly popular data resampling technique used to evaluate a predictive model's generalization capability and to avoid overfitting. When constructing the final model for predicting actual future cases, the learning function (or learning algorithm) is typically applied to the entire training set. The purpose of CV during the model-building phase is to estimate how well the final model will perform on new, unseen data (Berrar, 2019).

Proper and complete CV helps to avoid optimistic bias when estimating survival risk discrimination for a survival risk model developed using the full data set (Simon et al., 2011). For example, Molinaro (2005) compared different data resampling methods for high-dimensional data sets, which are often encountered in bioinformatics. Their findings indicate that Leave-One-Out CV (LOOCV), 10-fold CV, and the 0.632+ bootstrap method exhibit the smallest bias.

In summary, Simon (2011) believes that when properly implemented, CV methodology can be highly effective for evaluating survival risk modeling and should be more widely adopted. It allows for more efficient use of data in model development and validation compared to fixed sample splitting. However, in data sets with few events, the survival risk models developed may be suboptimal due to the limited data, and the cross-validated Kaplan–Meier curves of risk groups and time-dependent ROC curves may be imprecise (Dobbin and Simon, 2007).

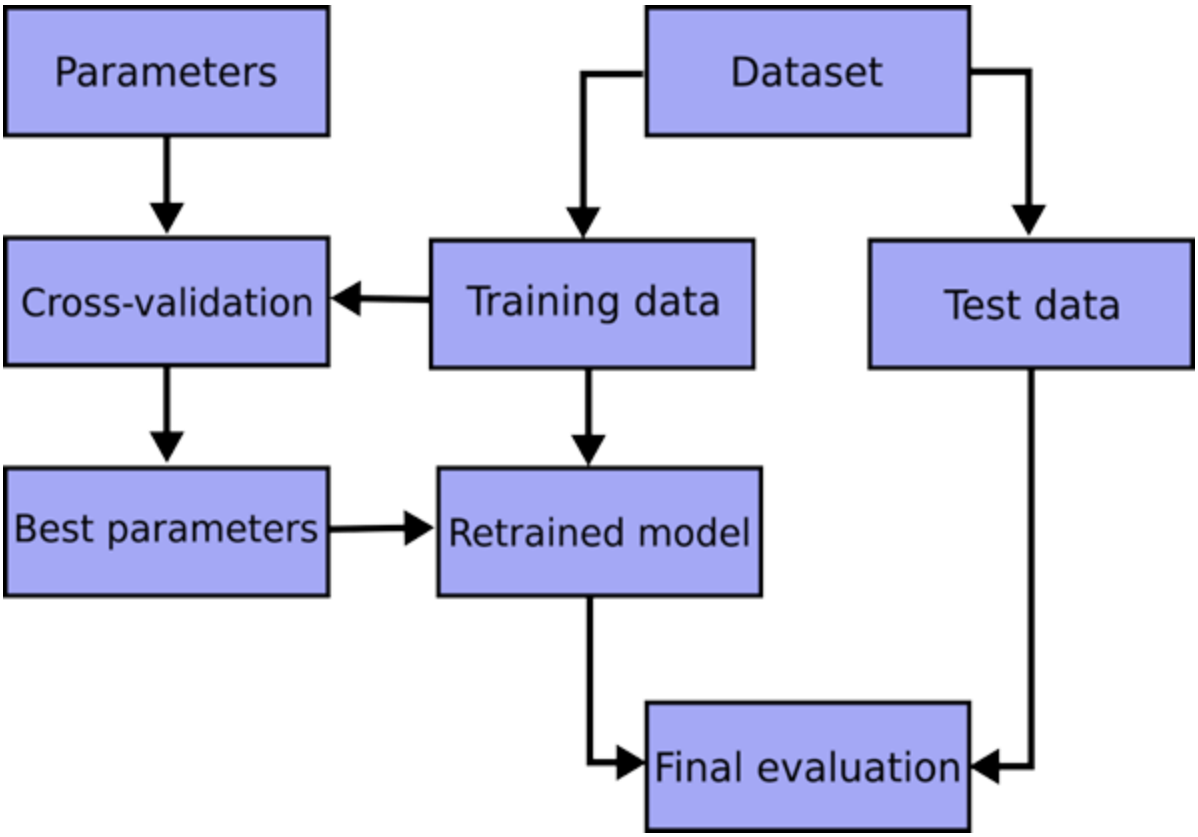

Figure 4. *Cross-Validation pipeline (Farheenshaukat, 2024)*

The Figure above, (Figure 4) illustrates a comprehensive workflow for optimizing and evaluating a machine learning model through CV. The process begins with defining the dataset and the set of parameters to be tuned. The dataset is then split into two groups, the one contains the training data and the other one the test data, ensuring that the model is evaluated on unseen data. The training data is used in the CV process, where different combinations of the defined parameters are tested to find the best performing set. This iterative process helps in identifying the optimal parameters that enhance the model's performance. Once the best parameters are identified, the model is retrained using the entire training dataset with these parameters. Finally, the retrained model is evaluated on the test data to ensure its generalizability and effectiveness.

2.6. Aim & objective

The aim of this study, titled "Optimization of the γ-OMP Algorithm for Feature Selection in Gene Expression Data," is to enhance the performance of the γ-OMP algorithm by developing a new algorithm called dcor-OMP. Both algorithms are a variant of the Orthogonal Matching Pursuit (OMP) widely employed in signal processing (Chen, Billings, & Luo, 1989; Davis, Mallat, & Zhang, 1994), and have been specifically tailored for gene expression data analysis.

Accurate FS in gene expression data is crucial for several reasons. First and foremost, genomic studies generate vast amounts of data, often comprising thousands of genes across multiple samples. Identifying the subset of genes that are most relevant for distinguishing biological conditions (such as disease states or treatment responses) is essential for understanding underlying biological mechanisms. Moreover, precise FS can significantly enhance the development of diagnostic tools and predictive models. By isolating key genes associated with specific outcomes, researchers can improve the accuracy of disease diagnosis, prognosis, and treatment selection. This can lead to more personalized and effective medical interventions, ultimately improving patient outcomes and healthcare efficiency. The dcor-OMP algorithm represents a potential advancement in this field by refining the feature selection process, thereby offering more reliable insights into gene interactions and regulatory networks. Through rigorous comparison with its predecessor γ-OMP, this study aims to validate dcor-OMP as a robust tool capable of addressing the complexities inherent in gene expression data analysis.

Optimizing methods like γ-OMP allows researchers to better use genetic information that will lead to faster discoveries in molecular biology and healthcare. Finally, all of these developments have the potential to change the way illnesses are understood, diagnosed, and treated, which will bring us closer to customized and targeted healthcare solutions.

Chapter 3 Material and methods

3.1. γ-OMP

The γ-OMP, which stands for generalized OMP and is pronounced as γ-OMP, was proposed by Tsagris et al in 2022. It is a FS algorithm based on the OMP algorithm. The γ-OMP is an extension of OMP designed to enhance its performance in specific applications, particularly those involving high-dimensional data, such as gene expression data. This extension addresses the limitations of the traditional OMP by allowing for the selection of groups of features rather than individual features, which can significantly improve the algorithm's effectiveness, especially in the presence of correlated features.

The γ-OMP algorithm is versatile and can handle various types of outcomes and features, including continuous, binary, nominal, ordinal, time-to-event, ratios, and measurements. This versatility makes it suitable for a wide range of applications in computational biology and bioinformatics, where data often come in different forms and structures. Moreover, the algorithm can utilize different regression models, such as linear or logistic regression, and employ various stopping criteria based on statistical measures. It also considers different types of residuals denoted by Resid in lines 7 and 13 of the algorithm presented in Table 1 and correlations to select the most relevant features for predictive analytics. The γ-OMP is scalable, easy to implement, and competitive in terms of predictive performance and computational efficiency, where traditional methods might struggle with the sheer volume and complexity of the information.

The key innovation of the γ-OMP is its ability to select a group of features that jointly contribute to reducing the residual, rather than selecting features one by one. This approach is particularly beneficial when there are groups of correlated features that together have a significant impact on the outcome. By considering these groups, the γ-OMP can more accurately capture the underlying structure of the data. The algorithm is adaptive, determining the optimal group size at each iteration, which enhances its flexibility and performance. Additionally, it includes several optimizations to reduce computational complexity, making it more efficient for large datasets. These optimizations ensure that the γ-OMP can handle extensive datasets without compromising on speed or accuracy.

The algorithm starts with an empty support set and the initial residual as the signal itself. At each iteration, a group of features with the highest joint correlation with the residual is selected. This group is then added to the support set, and the residual is updated by projecting the signal onto the subspace spanned by the current support set. This iterative process continues until the desired criteria are met, ensuring that the most relevant features are selected for the predictive model. The γ-OMP's iterative nature and group selection strategy make it a powerful tool for feature selection in high-dimensional spaces, providing robust and efficient solutions for complex data analysis tasks.

Initially, the algorithm requires input in the form of outcome values γ and a dataset X consisting of potential predictor features. Key functions such as f (model fitting), Resid (residual calculation),

Assoc (association measurement), and Stopping (stopping criterion) are also defined. The process begins by standardizing the data to ensure comparability, centering each feature and the outcome around zero and scaling them to unit norm. The selected features set S is initialized as empty, and initial residuals are set equal to the outcome values. The algorithm then iteratively selects features based on their association with the residuals, updating the current model and residuals with each iteration. Specifically, the feature with the highest association with the residuals is added to the set of selected features S, and a new model is fitted using these features. Residuals are recalculated as the difference between the actual outcome and the predicted outcome using the updated model. This process repeats, with the previous model M' and the current model M being updated in each iteration, until the stopping criterion is met. At this point, the algorithm returns the set of selected features, which represent the most predictive features for the outcome.

Table 1. *The γ-OMP algorithm*

Input: Outcome values y, dataset X
Hyper-Parameters: Functions $f, Resid, Assoc, Stopping$
Output: A subset $\mathcal{S} \subseteq \mathcal{X}$ of selected features.
$\mathcal{S} \leftarrow \emptyset$
Initialize current model $M \leftarrow f(y, X_{\mathcal{S}})$
Initialize previous model $M' \leftarrow \emptyset$
Initialize residuals $r \leftarrow Resid(y, X_{\mathcal{S}}, M)$
while $Stopping(M, M', y, X)$ **do**
    $X_* \leftarrow \arg\max_{i \in \mathcal{X} \setminus \mathcal{S}} Assoc(r, X_i)$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{X_*\}$
    $M' \leftarrow M$
    $M \leftarrow f(y, X_{\mathcal{S}})$
    $r \leftarrow Resid(y, X_{\mathcal{S}}, M)$
end while
return $\mathcal{S}$

3.2. dcor-OMP

In this study we built a unique feature selection and model optimization strategy based on distance correlation (dcor) within the OMP, which led up to our model named dcor-OMP. Our model is organized around three primary functions: 'dcor.omp', 'dcor.omp.path', and 'dcor.omp.cv'. First, the 'dcor.omp' function is intended to perform iterative feature selection by choosing predictors that have the maximum distance correlation with the response variable. This function iteratively picks variables, standardizes the predictors, and uses k-nearest neighbors (k-NN) regression to estimate the response and compute the AUC, essentially refining the model until a defined tolerance level is satisfied. 'dcor.omp's strength is its capacity to handle high-dimensional data while efficiently limiting down relevant features. Secondly, the 'dcor.omp.path' function expands the capabilities of 'dcor.omp' by generating a selection path over several tolerance levels. This function is called 'dcor.omp' with a variety of tolerance levels, capturing the advancement of selected features and associated AUC values. This approach provides a full perspective of the model's performance across various levels of stringency, allowing for a better-informed choice of the ideal tolerance level. Finally, the 'dcor.omp.cv' function incorporates cross-validation into the feature selection process, resulting in a stronger evaluation of model performance. This function divides the data into numerous folds and applies 'dcor.omp.path' to each training set before evaluating the model on the validation set and computing the AUC for each tolerance level. The CV method ensures that the chosen model generalizes effectively to previously unseen data, allowing a reliable assessment of its predicted accuracy. These functions work together to create an improved model and a better feature selection, which is especially useful for high-dimensional data. They use the strength of distance correlation to find relevant features, k-NN regression for flexible modeling, and CV to evaluate performance. This methodology improves both model accuracy and interpretability, making it more efficient over its predecessor, γ-OMP.

3.3. Pearson correlation

The Pearson correlation method is the most common method to use for numerical variables; it assigns a value between -1 and 1, where 0 is no correlation, 1 is total positive correlation, and -1 is total negative correlation. This is interpreted as follows: a correlation value of 0.7 between two variables such as “age” and “disease” would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable “age” increases, then “disease” will also increase, whereas if the value of the correlation is negative, then if “age” increases, “disease” decreases (Faizi and Alvi, 2023).

As such, it is related to distributional assumptions. More specifically, the Pearson is assuming a bivariate normal distribution, or a linear relationship between X and Y. According to Hahs-Vaughn (2023), when linearity is broken, the Pearson correlation is not robust; in fact, it is "very not robust... even a single aberrant point can alter r, the usual estimate of ρ, by a large amount." A useful implication is that even a single outlier can obscure a connection. Apart from outliers, non-

linearity, range restriction, residual magnitude, and rotating points are other data properties that might impact the strength of the Pearson correlation. Strong nonlinear relationships can nonetheless have a Pearson correlation coefficient of zero. Several alternatives to the Pearson have been created as a result of these constraints; some of these alternatives are more reliable than others. A large number of these substitute coefficients are listed in the section on nonparametric correlations (Chen and Anderson, 2023; Hahs-Vaughn, 2023). The formula for calculating the Pearson correlation coefficient is:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

where x and y are the two variables being compared, $\bar{x}$ and $\bar{y}$ are their respective means, and Σ denotes the sum over all observations.

By applying the Pearson correlation, bioinformaticians were able to identify significant relationships between genes. For instance, a positive correlation means that as the expression level of one gene increase, the expression level of another gene also increases. Conversely, a negative correlation means that as the expression level of one gene increase, the expression level of another gene decreases. For instance, de Souto (2008) used the Pearson correlation method to analyze gene expression data. This method is commonly used to measure the strength and direction of the linear relationship between two numerical variables. For their analysis, they calculated the Pearson correlation coefficients between the expression levels of different genes.

3.4. Energy distance, Distance covariance, Distance correlation

3.4.1. Energy distance

Energy distance is a measure of the difference between probability distributions. The term 'energy' is inspired by the concept of potential energy in a gravitational field, where potential energy is zero when two objects are at the same point (the gravitational center) and increases with their separation. This concept can be applied to data as follows: Let X and Y be independent random vectors in $\mathbb{R}^d$ with cumulative distribution functions (CDFs) F and G, respectively. Here, $\| \cdot \|$ denotes the Euclidean norm (length) of its argument, E represents the expected value, and a primed random variable X' denotes an independent and identically distributed (iid) copy of X; similarly, Y and Y' are iid (Székely and Rizzo, 2013; Rizzo and Székely, 2016). The squared energy distance can be defined using the expected distances between these random vectors as:

$$D2\,(F,G) \;=\; 2E||X \;-\; Y|| \;-\; E||X \;-\; X'|| \;-\; E||Y \;-\; Y'|| \;\geq 0$$

and the energy distance between distributions F and G is the square root of D^2 (F, G). It can be demonstrated that the energy distance D (F, G) satisfies all the properties of a metric, specifically

D (F, G) = 0 if and only if F = G. Thus, energy distance provides a way to determine when distributions are equal and offers a theoretical foundation for statistical inference and multivariate analysis based on Euclidean distances. This review covers several key applications and demonstrates their implementation (Rizzo and Székely, 2016).

An important use case for two samples involves testing the independence of random vectors. This involves verifying if the joint distribution of X and Y equals the product of their individual marginal distributions. Notably, the relevant statistics can be represented using a product-moment expression that involves the double-centered distance matrices of the X and Y samples. These distance-based statistics are similar to, but more comprehensive than, product-moment covariance and correlation, leading to the terms distance covariance (dCov) and distance correlation (dCor) which are defined below (Székely and Rizzo, 2012; Rizzo and Székely, 2016).

3.4.2. Distance covariance

The simplest formula for the distance covariance statistic is the square root of;

$$dCov^2(x,y) = \frac{1}{n^2}\sum_{i_{\lambda_j}=1}^{n} \hat{A}_{ij}\hat{B}_{ij}$$

where $\hat{A}$ and $\hat{B}$ are the double-centered distance matrices for the X and Y samples, respectively, with the subscript $_{ij}$ indicating the entry in the i^{th} row and j^{th} column. In the context of distance covariance, the matrices $\widehat{A}$ and $\widehat{B}$ are double-centered distance matrices constructed from the distance matrices of two sets of samples, X and Y. To create these matrices, we first compute the pairwise distance matrices for the samples in X and Y. Next, we calculate the row means, column means, and the overall mean of these distance matrices. Double centering involves adjusting each element in the distance matrices by subtracting the corresponding row and column means and adding the overall mean. This process results in the double-centered matrices $\hat{A}$ and $\hat{B}$, which isolate the variation in distances attributable to dependencies between the samples, thereby removing the effects of their overall location.

Distance covariance has many applications across different fields such as biology. The idea of distance covariance is expanded to assess the dependence between a covariate vector and a right-censored survival endpoint (Edelmann, Welchowski and Benner, 2022). This is achieved by creating an estimator based on an inverse-probability-of-censoring weighted U-statistic. The consistency of this new estimator is demonstrated. A large simulation study reveals that the distance covariance permutation tests perform well in detecting various complex associations. The application of these permutation tests to a gene expression dataset from breast cancer patients highlights its potential utility in biostatistical practice.

3.4.3. Distance correlation

Distance correlation, on the other hand, is a measure of association between two random variables that takes into account both linear and nonlinear relationships. It was introduced by Székely, Rizzo and Bakirov (2007) as a way to overcome some of the limitations of traditional correlation measures. The distance correlation between two random variables X and Y is defined as:

$$dCor(x,y) = \frac{dCov(x,y)}{\sqrt{dVar(x)dVar(y)}}$$

Distance correlation ranges from 0 to 1, in contrast to Pearson correlation, which can range from -1 to 1. It is zero if and only if X and Y are independent. This measure also has the advantage of being invariant to monotonic transformations of the variables, which allows it to identify nonlinear connections that conventional correlation measures would overlook. Additionally, it is symmetric, meaning that if and only if X and Y are independent, then dCor (X, Y) = dCor (Y, X) and it is zero. The distance variance is simply the distance covariance of a variable with itself:

$$d\,Var^2(X) = d\,Cov^2(X,X) = \frac{1}{n^2}\sum_{i=1}\sum_{j=1}\hat{A}_{ij}\hat{A}_{ij}$$

Similarly, for Y:

$$d\,Var^2(Y) = d\,Cov^2(Y,Y) = \frac{1}{n^2}\sum_{i=1}\sum_{j=1}\hat{B}_{ij}\hat{B}_{ij}$$

dVar(X) is the distance variance of X, which is equivalent to the distance covariance of X with itself. as well as for Y where Var(Y) is the distance variance of Y, which is equivalent to the distance covariance of Y with itself.

Distance correlation has the advantage of being invariant to monotonic transformations of the variables, which allows it to identify nonlinear connections that conventional correlation measures would overlook. Applications of distance correlation can be found in many domains, such as genetics, machine learning, and statistics. It can be applied to high-dimensional data to perform FS, assess the similarity across datasets, and find nonlinear correlations between variables (Hou *et al.*, 2022).

3.5. k-NN algorithm

k-NN algorithm is a supervised machine learning method used for classification and regression problems (Dadi, 2024). As a nonparametric classification technique, k-NN makes no assumptions about the underlying data distribution, making it renowned for its simplicity and efficiency. To forecast the class of unlabeled data, a labeled training dataset with data points classified into multiple classes is used. Thus, k-NN utilizes this labeled training data to categorize new data points based on the majority class of their nearest neighbors, identified using Euclidean distance (Zhang et al., 2018).

k-NN classification involves two main steps: the learning step, where a classifier is constructed using the training data, and the assessment step, where the classifier is evaluated. The algorithm classifies new unlabeled data by analyzing which classes the nearest neighbors belong to. The value of 'k' determines how many neighbors are considered in the classification process. When encountering a new unlabeled data point, k-NN first identifies the k nearest neighbors based on the Euclidean distance. The Euclidean distance can be calculated as follows: If two vectors x_i and x_j are given, where $x_i = (x_{i1},\ x_{i2}, x_{i3}, x_{i4},\ x_{i5}, \dots, x_{ip})$ and $x_j = (x_{j1}, x_{j2}, x_{j3}, x_{j4}, x_{j5}, \dots, x_{jp})$.

The difference between x_i and x_j is:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{jl})^2}$$

It then classifies the new data point into the class that the majority of these neighbors belong to. Factors affecting k-NN's performance include the value of 'k', the distance metric used, and the normalization of parameters. When 'k' is set to 1, the data point is assigned to the class of its nearest neighbor, resulting in zero error for training data but potential overfitting. To balance accuracy and overfitting, a larger value of 'k' is usually preferred, but the optimal choice depends on the specific dataset and its characteristics (Taunk et al., 2019). The algorithm's performance heavily depends on the choice of 'k', the number of neighbors considered. k-NN requires storing the entire training dataset and performs classification by calculating distances at prediction time, making it computationally intensive but straightforward to implement. Additionally, the computational cost of k-NN is relatively high because all calculations are performed during classification rather than during the training phase. This characteristic makes it a "lazy learning" algorithm, as it stores the training data without any processing until a prediction is required. Consequently, the entire training dataset is needed during the testing phase since the algorithm does not perform generalization (Taunk et al., 2019).

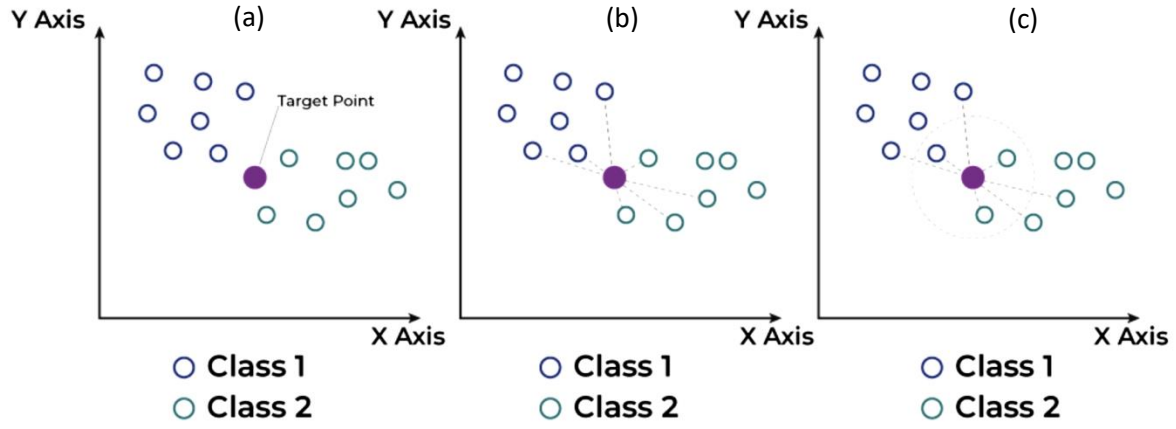

Figure 5. *Illustration of the k-Nearest Neighbors (k-NN) algorithm in action (Dadi, 2024)*

Figure 5 illustrates the k-NN algorithm in action across three stages. In the first panel, (a), we see a target point (purple) situated among two classes, represented by Class 1 (blue) and Class 2 (green). The middle panel, (b), demonstrates the process of finding the nearest neighbors of the target point using Euclidean distance, with dashed lines connecting the target point to its neighbors. Finally, the right panel, (c) shows the classification decision: the target point is classified based on the majority class of its nearest neighbors, depicted within a dashed circle.

The k-NN algorithm is widely used in bioinformatics for tasks such as gene expression analysis, protein function prediction, and disease classification. In a study by Dhawan, Selvaraja, and Duan (2010), k-NN was applied to classify functional categories in microarray data, demonstrating its effectiveness in handling biological datasets and providing reliable classification results. The study achieved an average accuracy of 95% across various datasets. Additionally, Vengateshkumar, Sanmugavel, and Raj (2019) emphasized the utility of k-NN in handling complex biological data, achieving classification accuracies ranging from 85% to 92% depending on the dataset and normalization techniques used. These studies underscore k-NN's value in bioinformatics for accurately interpreting complex datasets and facilitating scientific discoveries, highlighting k-NN's robustness and reliability in bioinformatics application.

3.6. OMP and its variants

In this last section of material and methods, we will highlight and focus on the main distinct features between the OMP algorithm and its variants algorithms; γ-OMP, and the newly developed dcor-OMP. With The standard OMP algorithm serving as a baseline for the other two. These distinct features are represented in Table 2 below, as the following; Correlation Type, Model type, Stopping criterion, Target Type, and, Algorithm Complexity.

On the first hand, OMP, a greedy well-used algorithm popular for feature selection due to its computational efficiency and fast execution. The OMP uses Pearson correlation in order to find the best correlation between the gene expression and the cancer type, and it uses the linear regression as the model type. The algorithm stops the computation based on the SSE, making it

ideal for numerical targets. Because it is a simple algorithm the OMP has the fastest execution between the three. On the other hand, γ-OMP expands upon the traditional OMP, it is designed to enhance its performance in specific applications, particularly those involving high-dimensional non-linear data, such as gene expression data. The key point of γ-OMP is that it can treat numerous types of outcome variables employing various regression models by using any pairwise association and supports various models, including non-/semi-/parametric ones, its stopping criteria is also based on general criteria like log-likelihood or BIC making it the most flexible algorithm among the three, allowing for a more divers options to detect a correlation between gene expression and cancer type .The computational efficiency of γ-OMP made it suitable for handling our data without compromising performance for any target type. Finally, The dcor-OMP algorithm is formulated based on the γ-OMP algorithm that integrates both, the distance correlation and the k-NN methods to improve feature selection. By using distance correlation, the dcor-OMP algorithm can identify meaningful correlations between features in high-dimensional datasets, which is necessary in our case, aiming in better model performance and interpretability. The use of k-NN aids in estimating responses and computing residuals, improving the selection process. The dcor-OMP will stop computing when the AUC stop improving. Although the dcor-OMP is the most computationally intensive among the three due to distance calculations, and its dataset must be Binary/numerical target, this combination of distinct features makes it a promising tool for analyzing gene expression data surpassing the other two algorithms.

Table 2 *Comparison of OMP and its Variants*

	OMP	**γ-OMP**	**dcor-OMP**
Correlation Type	Pearson correlation	Any type of pairwise association	Distance correlation
Model type	Linear regression	Non-/Semi-/Parametric models (Generalized Additive Models, kernel based)	k-NN
Stopping criterion	Sum of Residual Squares (SSE)	General criterion (log-likelihood, BIC)	Area Under the Curve (AUC) (when the AUC stop improving)
Target Type	Numerical target	Any type of target	Binary/numerical target
Algorithm Complexity	Simple, faster execution	Computationally efficient	Computationally intensive due to distance calculations

Chapter 4 Data analysis and results

4.1. GSE (gene expression data series)

GSE (gene expression data series) data refers to datasets from the Gene Expression Omnibus (GEO), a public repository for high-throughput gene expression and other functional genomics datasets. The GSE data typically includes gene expression profiles obtained from microarray or RNA-seq experiments, which are extensively used in bioinformatics and computational biology research.

Our database comprises 7 GSE datasets for cancer. All the following information was sourced from BioDataome, an exceptional and comprehensive database developed by the University of Crete. BioDataome contains an extensive collection of uniformly preprocessed and annotated datasets, covering approximately 5,600 datasets and 260,000 samples related to around 500 diseases, including gene expression data, RNA-Seq, and DNA methylation data. The annotation of these datasets with disease ontology terms facilitates large-scale experiments and meta-analyses. From this significant database, we selected the necessary data for our thesis, leveraging its rich capabilities for analyzing and processing biological data, thereby enhancing our research with reliable and well-documented information. Our dataset is composed of 4 types of cancers lung cancer, breast cancer, gastric cancer, and colorectal cancer. In order to reduce errors and increase the performance of the model, the size of the sample for each dataset will be at least 100 samples.

GSE10780

The “GSE10780” dataset titled "Proliferative genes dominate malignancy-risk gene signature in histologically-normal breast tissue," is extracted from the Homo sapiens species. It includes 143 completely histologically-normal breast tissues, leading to the identification of a gene signature associated with malignancy risk, potentially serving as a marker for subsequent breast cancer development. The design involves RNA extraction from micro dissected frozen breast tissues for gene array analysis.

GSE20465

The dataset “GSE20465”, titled "Her2/Neu breast cancer mouse model whole tissue transcriptome" was generated from the species Mus musculus, a Her2-driven mouse model of breast cancer, mirroring human breast cancer. Biospecimens from this mouse model are freely available through a sample repository, eliminating the need for breeding animals and collecting biospecimens for researchers testing biological hypotheses. Experimental design entails twelve datasets, comprising 841 LC-MS/MS experiments (plasma and tissues) and 255 microarray analyses across various tissues (thymus, spleen, liver, blood cells, and breast). Cases and controls

were meticulously paired to prevent bias. Results show the identification of 18,880 unique peptides, with 3,884 and 1,659 non-redundant protein groups in plasma and tissue datasets, respectively. Notably, 61 protein groups overlap between cancer plasma and tissue.

GSE29272

The “GSE29272” dataset, titled “Affymetrix gene expression array data for cardiac and non-cardia gastric cancer samples' ' is generated from the species of Homo sapiens. The GSE identifies different and common dysregulated genes in cardiac and non-cardia gastric cancer in the two types of gastric cancer. The design consists of cardiac and non-cardia gastric tumors and normal glands and it consists of 268 samples 134 from the adjacent tissue normal glands 72 from the tumor tissue non-cardia of gastric and 62 from tumor tissue cardia of gastric.

GSE31210

The dataset “GSE31210” titled "Gene expression data for pathological stage I-II lung adenocarcinomas" focuses on identifying genes up-regulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. To delineate molecular characteristics, 226 primary lung adenocarcinomas of pathological stage I-II were examined for EGFR, KRAS, and ALK mutations. Genome-wide expression profiling revealed genes up-regulated in ALK-mutated lung adenocarcinomas and those lacking EGFR, KRAS, and ALK mutations. Among 174 up-regulated genes specifically identified in 79 cases without EGFR and KRAS mutations, ALK was noteworthy. These cases were further categorized into ALK-positive ADCs, Group A triple-negative ADCs, and Group B triple-negative ADCs based on expression patterns of the 174 genes. ALK-positive ADCs exhibited significant overexpression of 30 genes, including ALK and GRIN2A. Group A triple-negative ADC cases demonstrated worse prognoses compared to cases with EGFR, KRAS, or ALK mutations, and Group B triple-negative ADC cases. Nine genes, including DEPDC1, were significantly up-regulated in Group A cases, critical for prognosis prediction. These genes may aid in selecting patients for adjuvant chemotherapy post-surgical resection of stage I-II triple-negative ADCs and inform the development of molecular targeting therapies for these patients. The overall design involved expression profiling of 226 lung adenocarcinomas, including cases with EGFR mutation, KRAS mutation, EML4-ALK fusion, and triple-negative cases.

GSE35978

The dataset “GSE35978” titled "Expression data from the human cerebellum and parietal cortex brain" comprises a SuperSeries consisting of several SubSeries. These SubSeries likely contain detailed expression profiling data from specific experiments or conditions within the human cerebellum and parietal cortex brain regions. The overall design of this SuperSeries refers to the individual Series within it, providing a comprehensive exploration of gene expression patterns in these brain regions. It used 312 samples from which 158 represents samples from the parietal cortex and 131 from the cerebellum.

GSE41258

The following “GSE41258” dataset titled "Expression data from colorectal cancer patients" entails a study conducted on patients diagnosed with colonic neoplasms at Memorial Sloan-Kettering Cancer Center between 1992 and 2004. Biological specimens utilized in the study encompass primary colon adenocarcinomas, adenomas, metastases, and corresponding normal mucosae. The overall design comprises 390 expression arrays, from which 186 representing Primary Tumor, 53 Normal colon, 49 polyp, 47 liver Metastasis, 20 lung Metastasis, 13 normal liver, 12 cell line, 7 normal lung, and 2 Microadenoma.

GSE44077

The dataset “GSE44077” titled "Gene expression profiling of the adjacent airway field cancerization in early-stage NSCLC" focuses on characterizing the transcriptomic landscape of adjacent airway field cancerization in non-small cell lung cancer (NSCLC), a phenomenon where lung tumors and nearby normal tissues exhibit specific abnormalities relevant to lung cancer development. The study aims to elucidate the molecular architecture of adjacent airway field cancerization alongside tumors, providing deeper insights into lung cancer biology and oncogenesis. Using the Affymetrix Human Gene 1.0 ST platform, the transcriptome of matched NSCLC tumors, multiple normal airway epithelia at varying distances from tumors, and uninvolved normal lung tissues were analyzed. The overall design involves analyzing the transcriptomic profiles of adjacent airways to identify global differentially expressed cancerization patterns and airway profiles potentially influenced by proximity to tumors. 226 samples were used from which 96 samples belonged to airway samples from field cancerization cases, 65 normal lung tissue, 56 NSCLC tissue, and 9 samples from lung carcinoid tissue.

4.2. Result and Discussion

Our results consist of a comparison between the dcor-OMP and its predecessor, the γ-OMP. To achieve this, we start by calculating the ROC curve for each gene expression dataset and then extract the AUC from it for each of the two models. The AUC determines the performance of the model; the higher the AUC, the more accurate the model is in classifying the correlation of gene expression to the cancer under investigation. We then compare the AUC for each gene expression dataset and compute the mean AUC difference. A higher mean AUC difference indicates better performance of the dcor-OMP compared to the γ-OMP. we end up with a box plot graph as seen in the figure below (figure 6).

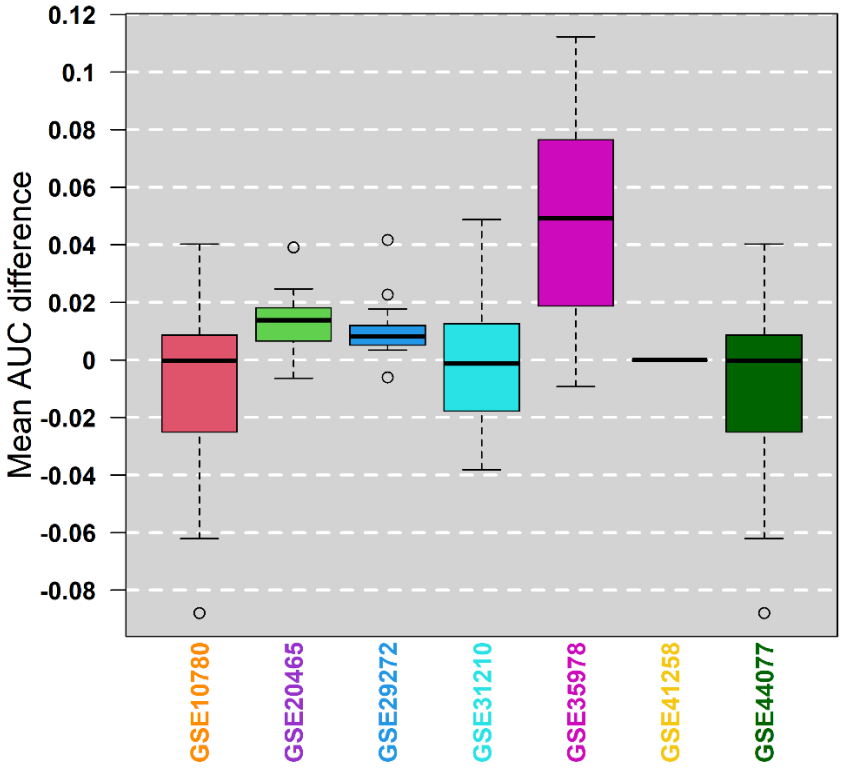

Figure 6 AUC mean difference boxplot

Figure 6, shows box plots comparing the mean difference of the Area Under the Curve (AUC) of dcor-OMP and γ-OMP for feature selection across seven datasets. The Y-axis shows the difference in mean AUC between dcor-OMP and γ-OMP across each dataset. Values above 0 show dcor-OMP outperformed γ-OMP. Values with 0 mean value imply equivalent performance. Values below 0 suggest dcor-OMP performed poorer than γ-OMP.

The horizontal lines within the boxes show the median difference in AUC, which is used to analyze the distribution of differences. The median in the box plot was acquired by repeating the k fold

CV 20 times for each the γ-OMP and the dcor-OMP then doing the difference between the mean AUC of each GSE dataset. The boxes include the highest and lower quartiles of the differences, illustrating how the data is spread out or dispersed. Furthermore, the whiskers extend outward to the nearest data points within an interquartile range (IQR) of the corresponding quartiles, thus displaying the range of variability beyond the quartiles. Overall, this representation offers a comprehensive view of the performance disparities between using dcor-OMP and γ-OMP across multiple datasets explained bellow;

In the box plot we can see that three of the seven datasets have a mean value above zero, whereas the other four have values of zero and none below zero. The gene expression dataset "GSE35978" named "Expression data from the human cerebellum and parietal cortex brain" had the greatest mean value of 0.05 and the biggest IQR, indicating that the dcor-OMP outperformed the γ-OMP on this specific dataset. Overall, the dcor-OMP outperformed the γ-OMP, since no dataset had a value below zero. Outliers are situations in which performance departed dramatically from the median. This might be owing to special properties of some samples in the datasets, which benefit more or less from the model. The difference in performance among datasets can be due to the differences in sample size, feature dimensionality, and noise levels, with datasets with greater noise or fewer samples showing more substantial gains using sophisticated feature selection methods like dcor-OMP. This shows that dcor-OMP may better capture complicated links and interactions in the data than γ-OMP. Moreover, dcor-OMP consistently outperforms γ-OMP across all datasets, demonstrating its dependability as a feature selection approach for heterogeneous gene expression data.

Table 3. *Information about the GSE Data and mean AUC difference*

GSE	Disease	Sample size	Features	Cases (%)	Mean AUC difference
GSE10780	breast cancer	226	33252	194 (85.84%)	0
GSE20465	breast cancer	250	45101	125 (50.00%)	0.018
GSE29272	gastric cancer	268	22283	134 (50.00%)	0.008
GSE31210	lung cancer	246	54675	226 (91.87%)	0
GSE44077	lung cancer	230	33252	190 (82.61%)	0
GSE35978	brain cancer	305	33297	205 (67.21%)	0.05
GSE41258	colorectal cancer	178	22283	48 (26.97%)	0

Table 3 summaries our dataset, it displays for each GSE the type of cancer, its sample size, the number of features, the number positive case in percentage, and finally, the AUC mean difference between the γ-OMP and the dcor-OMP. On the one hand, we can observe that sample size might be connected with AUC mean difference. This suggests that the outperformance of the dcor-OMP is due to sample size. As the sample size increase the mean AUC difference increase, starting from a sample size of 250 until 305. On the second side, we can observe that the kind of sickness is not strictly related to the model's performance. For example, all lung cancers have an AUC mean difference of zero, but breast cancers have a mean AUC difference of 0 and 0.018. On the third hand, we can see that the performance of the model might be linked to the % of cases within the dataset, we can see from the table that the only difference of the AUC mean is when the % of positive cases is 50% or close to it. Overall, the average AUC mean difference is 0.011 which indicates that on average the dcor-OMP will have a better AUC of 0.011 than the γ-OMP, therefore performed better.

Chapter 5 Conclusion

Our focuses on optimizing the γ-OMP method for feature selection in gene expression data, a crucial work in bioinformatics that improves cancer detection systems' accuracy and efficiency. Our work improved the γ-OMP method by integrating distance correlation to construct the dcor-OMP model in addition with k-NN, which performed better with high-dimensional gene expression data. The dcor-OMP performed better than its predecessor, γ-OMP, in through comparisons of seven cancer-related datasets, with higher mean AUC values across three datasets and a better average mean difference of 0.11. This difference may be related to dcor-OMP's capacity to better capture complex interactions and linkages in data, making it an effective feature selection tool. Overall, this work is important because it has the potential to improve customized treatment by offering more precise models for diagnosis and prognosis. The upgraded feature selection approach, dcor-OMP not only detects critical genetic markers, but it also minimizes data redundancy and noise, resulting in more accurate predictions. As a result, this technique has the potential to greatly affect the development of tailored medicines while also improving patient outcomes. Overall, when it comes to OMP, γ-OMP, and dcor-OMP, OMP is straightforward, using Pearson correlation to understand linear relationships, and it performs best with linear regression models, making it quick and efficient for non-binary numerical data. It ends when the improvements in model errors, as measured by residual sums, stop. γ-OMP is more versatile and compatible with many generalized models, from simple to complicated. It is designed for binary datasets and uses more generalized stopping criteria, such as log-likelihood, which increases the computational complexity. Finally, dcor-OMP is excellent at finding nonlinear interactions via distance correlation and frequently use the k-NN technique. It is the most computationally intensive, especially because it stops changing when no further increases in the AUC are detected, making it ideal for binary datasets. Each strategy has various advantages depending on the complexity and kind of data under consideration with dcor-OMP performing better than its predecessor on the same dataset. To ensure the dcor-OMP algorithm's reliability and usability in the future, it must be evaluated on a larger variety of datasets. We also expect that combining dcor-OMP with other machine learning models can considerably improve its prediction value. Beyond gene expression data, applying this method to other bioinformatics areas such as proteomics and metabolomics might lead to exciting new prospects and benefits.

Bibliography

1. Abumsimir, B., Al-Qaisi, T.S. and Kasmi, Y. (2022) 'Rereading the genetic origin of cancer: the puzzle of all eras', *Future Science OA*, 8(5), pp. FSO799. Available at: https://doi.org/10.2144/fsoa-2022-0014.
2. Cancer (IARC), T.I.A. for R. on (2024) *Global Cancer Observatory*. Available at: https://gco.iarc.fr/ (Accessed: 2 June 2024).
3. *Cancer Statistics - NCI* (2015). Available at: https://www.cancer.gov/about-cancer/understanding/statistics (Accessed: 2 June 2024).
4. *Cancer Types | Find Your Cancer Type | American Cancer Society* (2024). Available at: https://www.cancer.org/cancer/types.html (Accessed: 2 June 2024).
5. Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome Biology*, 17(1), pp. 13. Available at: https://doi.org/10.1186/s13059-016-0881-8.
6. Gilad, Y. *et al.* (2000) 'Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes', *Nature Genetics*, 26(2), pp. 221–224. Available at: https://doi.org/10.1038/79957.
7. Hasty, P. and Montagna, C. (2014) 'Chromosomal Rearrangements in Cancer: Detection and potential causal mechanisms', *Molecular & Cellular Oncology*, 1(1). Available at: https://doi.org/10.4161/mco.29904.
8. Kim, I.-J., Kang, H.C. and Park, J.-G. (2004) 'Microarray Applications in Cancer Research', *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, 36(4), pp. 207–213. Available at: https://doi.org/10.4143/crt.2004.36.4.207.
9. Lilenbaum, R.C. *et al.* (2001) 'Phase II trial of weekly docetaxel in second-line therapy for non small cell lung carcinoma', *Cancer*, 92(8), pp. 2158–2163. Available at: https://doi.org/10.1002/1097-0142(20011015)92:8<2158::aid-cncr1558>3.0.co;2-2.
10. Martel, C. de *et al.* (2020) 'Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis', *The Lancet Global Health*, 8(2), pp. e180–e190. Available at: https://doi.org/10.1016/S2214-109X(19)30488-7.
11. *Microarray Technology* (2024). Available at: https://www.genome.gov/genetics-glossary/Microarray-Technology (Accessed: 2 June 2024).
12. *Microarray Technology: An introduction to DNA Microarray* (2024). Available at: https://www.premierbiosoft.com/tech_notes/microarray.html (Accessed: 2 June 2024).
13. Mortazavi, A. *et al.* (2008) 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nature Methods*, 5(7), pp. 621–628. Available at: https://doi.org/10.1038/nmeth.1226.

14. Ozsolak, F. and Milos, P.M. (2011) ‘RNA sequencing: advances, challenges and opportunities’, *Nature Reviews Genetics*, 12(2), pp. 87–98. Available at: https://doi.org/10.1038/nrg2934.

15. Stark, R., Grzelak, M. and Hadfield, J. (2019) ‘RNA sequencing: the teenage years’, *Nature Reviews Genetics*, 20(11), pp. 631–656. Available at: https://doi.org/10.1038/s41576-019-0150-2.

16. The AACR Project GENIE Consortium *et al.* (2017) ‘AACR Project GENIE: Powering Precision Medicine through an International Consortium’, *Cancer Discovery*, 7(8), pp. 818–831. Available at: https://doi.org/10.1158/2159-8290.CD-17-0151.

17. *The Genetics of Cancer - NCI* (2015). Available at: https://www.cancer.gov/about-cancer/causes-prevention/genetics (Accessed: 2 June 2024).

18. Guyon, I. and Elisseeff, A. (2003) 'An Introduction to Variable and Feature Selection', Journal of Machine Learning Research, 3, pp. 1157-1182. Available at: https://doi.org/10.1093/nar/gky1049.

19. *Types of Cancer Treatment - NCI* (2017). Available at: https://www.cancer.gov/about-cancer/treatment/types (Accessed: 2 June 2024).

20. Dressler, L. et al. (2022) ‘Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource’, Genome Biology, 23(1), pp. 35. Available at: https://doi.org/10.1186/s13059-022-02607-z.

21. Narrandes, S. and Xu, W. (2018) ‘Gene Expression Detection Assay for Cancer Clinical Use’, Journal of Cancer, 9, pp. 2249–2265. Available at: https://doi.org/10.7150/jca.24744.

22. Nourbakhsh, M. et al. (2024) ‘Prediction of cancer driver genes and mutations: the potential of integrative computational frameworks’, Briefings in Bioinformatics, 25(2), pp. bbad519. Available at: https://doi.org/10.1093/bib/bbad519.

23. Tang, Y.-Y. et al. (2021) ‘Identification of driver genes based on gene mutational effects and network centrality’, BMC Bioinformatics, 22(3), pp. 457. Available at: https://doi.org/10.1186/s12859-021-04377-0.

24. Vicente-Dueñas, C. et al. (2013) ‘Function of oncogenes in cancer development: a changing paradigm’, The EMBO Journal, 32(11), pp. 1502–1513. Available at: https://doi.org/10.1038/emboj.2013.97.

25. *Understanding What Cancer Is: Ancient Times to Present | American Cancer Society* (2018). Available at: https://www.cancer.org/cancer/understanding-cancer/history-of-cancer/what-is-cancer.html (Accessed: 2 June 2024).

26. Wang, Z., Gerstein, M. and Snyder, M. (2009) ‘RNA-Seq: a revolutionary tool for transcriptomics’, *Nature Reviews Genetics*, 10(1), pp. 57–63. Available at: https://doi.org/10.1038/nrg2484.

27. *What Is Cancer? - NCI* (2021). Available at: https://www.cancer.gov/about-cancer/understanding/what-is-cancer (Accessed: 2 June 2024).

28. Petinrin, O.O., Saeed, F., Salim, N., Toseef, M., Liu, Z. and Muyide, I.O., (2023). Dimension Reduction and Classifier-Based Feature Selection for Oversampled Gene Expression Data and Cancer Classification. Processes, 11(7), pp.1940. Available at: https://doi.org/10.3390/pr11071940.

29. Li, L., Algabri, Y.A., and Liu, Z.P., (2023). Identifying Diagnostic Biomarkers of Breast Cancer Based on Gene Expression Data and Ensemble Feature Selection. Current Bioinformatics, 18(3), pp. 232-246. Available at: https://doi.org/10.2174/1574893618666230111153243.

30. Wang, J., & Ye, J. (2013). Generalized Orthogonal Matching Pursuit for Feature Selection. arXiv preprint arXiv:1308.0887.

31. Sun, J., & Qian, M. (2018). Minimum Redundancy Feature Selection Based on Orthogonal Matching Pursuit. IEEE Transactions on Neural Networks and Learning Systems, 29(4), 960-975.

32. Deng, X., Li, M., Deng, S., and Wang, L., (2021). Hybrid Gene Selection Approach Using XGBoost and Multi-Objective Optimization Genetic Algorithm for Cancer Classification. arXiv. Available at: https://arxiv.org/abs/2106.05841.

33. Liu, Y., Chen, X., and Wu, J., (2023). Improving the Classification of Alzheimer's Disease Using Hybrid Gene Selection Pipeline and Deep Learning. Frontiers in Genetics, 14, pp. 100-115. Available at: https://doi.org/10.3389/fgene.2023.00234.

34. Pragadeesh, C., Jeyaraj, R., Siranjeevi, K., Abishek, R., and Jeyakumar, G., (2019). Hybrid Feature Selection Using Micro Genetic Algorithm on Microarray Gene Expression Data. Journal of Intelligent & Fuzzy Systems, 36(3), pp. 2241-2246. Available at: https://doi.org/10.3233/JIFS-169935.

35. Kumar, S., Gupta, D., and Sharma, V., (2023). A Hybrid Machine Learning Feature Selection Model (HMLFSM) to Enhance Accuracy in Cancer Classification. PLoS One, 18(4), pp. e0278851.

36. Naeem, M., Qureshi, M.A., and Ali, R., (2022). A Comprehensive Survey of Recent Hybrid Feature Selection Methods in Gene Expression Data for Cancer Classification. IEEE Access, 10, pp. 57291-57310. Available at: https://doi.org/10.1109/ACCESS.2022.3182444.

37. Alsberg, B.K. *et al.* (1998) 'Variable selection in wavelet regression models', *Analytica Chimica Acta*, 368(1), pp. 29–44. Available at: https://doi.org/10.1016/S0003-2670(98)00194-9.

38. Ang, J.C. *et al.* (2016) 'Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection', *IEEE/ACM Transactions on Computational Biology and*

Bioinformatics, 13(5), pp. 971–989. Available at: https://doi.org/10.1109/TCBB.2015.2478454.

39. *'NGS sequencing' – Grafiati* (2021). Available at: https://www.grafiati.com/en/literature-selections/ngs-sequencing/ (Accessed: 16 January 2024).

40. Bowtell, D.D.L. (1999) 'Options available — from start to finish — for obtaining expression data by microarray', *Nature Genetics*, 21(1), pp. 25–32. Available at: https://doi.org/10.1038/4455.

41. Chandrashekar, G. and Sahin, F. (2014) 'A survey on feature selection methods', *Computers & Electrical Engineering*, 40(1), pp. 16–28. Available at: https://doi.org/10.1016/j.compeleceng.2013.11.024.

42. Chen, D. and Anderson, C.J. (2023) 'Categorical data analysis', in R.J. Tierney, F. Rizvi, and K. Ercikan (eds) *International Encyclopedia of Education (Fourth Edition)*. Oxford: Elsevier, pp. 575–582. Available at: https://doi.org/10.1016/B978-0-12-818630-5.10070-3.

43. Berrar, D. (2019) 'Cross-Validation', in S. Ranganathan et al. (eds) *Encyclopedia of Bioinformatics and Computational Biology*. Oxford: Academic Press, pp. 542–545. Available at: https://doi.org/10.1016/B978-0-12-809633-8.20349-X.

44. *Classification: ROC Curve and AUC | Machine Learning* (2022) *Google for Developers*. Available at: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc (Accessed: 8 June 2024).

45. Dobbin, K.K. and Simon, R.M. (2007) 'Sample size planning for developing classifiers using high-dimensional DNA microarray data', *Biostatistics*, 8(1), pp. 101–117. Available at: https://doi.org/10.1093/biostatistics/kxj036.

46. Edelmann, D., Welchowski, T. and Benner, A. (2022) 'A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing', *Biometrics*, 78(3), pp. 867–879. Available at: https://doi.org/10.1111/biom.13470.

47. Faizi, N. and Alvi, Y. (2023) 'Correlation∗', in N. Faizi and Y. Alvi (eds) *Biostatistics Manual for Health Research*. Academic Press, pp. 109–126. Available at: https://doi.org/10.1016/B978-0-443-18550-2.00002-5.

48. Farheenshaukat (2024) 'Cross Validation', *Medium*, 2 February. Available at: https://medium.com/@farheenshaukat_19/cross-validation-6ce5702e8eed (Accessed: 8 June 2024).

49. *K-Nearest Neighbor (KNN) Algorithm* (2017) *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/k-nearest-neighbours/ (Accessed: 8 June 2024).

50. *National Center for Biotechnology Information* (2024). Available at: https://www.ncbi.nlm.nih.gov/ (Accessed: 9 June 2024).

51. AUC ROC Curve in Machine Learning (2020) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/auc-roc-curve/ (Accessed: 22 June 2024).

52. BRESLOW, N.E. et al. (1978) 'Estimation of multiple relative risk functions in matched case-control studies', American Journal of Epidemiology, 108(4), pp. 299–307. Available at: https://doi.org/10.1093/oxfordjournals.aje.a112623.

53. Liang, S. et al. (2018) 'A Review of Matched-pairs Feature Selection Methods for Gene Expression Data Analysis', Computational and Structural Biotechnology Journal, 16, pp. 88–97. Available at: https://doi.org/10.1016/j.csbj.2018.02.005.

54. Refaeilzadeh, P., Tang, L. and Liu, H. (2016) 'Cross-Validation', in L. Liu and M.T. Özsu (eds) *Encyclopedia of Database Systems*. New York, NY: Springer, pp. 1–7. Available at: https://doi.org/10.1007/978-1-4899-7993-3_565-2.

55. Rizzo, M.L. and Székely, G.J. (2016) 'Energy distance', *WIREs Computational Statistics*, 8(1), pp. 27–38. Available at: https://doi.org/10.1002/wics.1375.

56. Simon, R.M. *et al.* (2011) 'Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data', *Briefings in Bioinformatics*, 12(3), pp. 203–214. Available at: https://doi.org/10.1093/bib/bbr001.

57. de Souto, M.C. *et al.* (2008) 'Clustering cancer gene expression data: a comparative study', *BMC Bioinformatics*, 9(1), pp. 497. Available at: https://doi.org/10.1186/1471-2105-9-497.

58. Székely, G.J. and Rizzo, M.L. (2012) 'On the uniqueness of distance covariance', *Statistics & Probability Letters*, 82(12), pp. 2278–2282. Available at: https://doi.org/10.1016/j.spl.2012.08.007.

59. Székely, G.J. and Rizzo, M.L. (2013) 'Energy statistics: A class of statistics based on distances', *Journal of Statistical Planning and Inference*, 143(8), pp. 1249–1272. Available at: https://doi.org/10.1016/j.jspi.2013.03.018.

60. Zou Q, Wan S, Ju Y, Tang J, Zeng X. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. BMC (Biomedcentral.com) Syst Biol 2016;10. https://doi.org/10.1186/s12918-016-0353-5.

61. Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q. Tumor origin detection with tissue specific miRNA and DNA methylation mon markers. Bioinformatics (2017). https://doi.org/ 10.1093/bioinformatics/btx622

62. Chuang, L.-Y. *et al.* (2008) 'Improved binary PSO for feature selection using gene expression data', *Computational Biology and Chemistry*, 32(1), pp. 29–38. Available at: https://doi.org/10.1016/j.compbiolchem.2007.09.005.

63. Danasingh, A.A., Balamurugan, S. and EPIPHANY, J.L. (2016) 'Literature Review on Feature Selection Methods for High-Dimensional Data', *International Journal of Computer Applications*, 136. Available at: https://doi.org/10.5120/ijca2016908317.

64. Das, A. (2020) ‘Logistic Regression’, in F. Maggino (ed.) *Encyclopedia of Quality of Life and Well-Being Research*. Cham: Springer International Publishing, pp. 1–2. Available at: https://doi.org/10.1007/978-3-319-69909-7_1689-2.

65. Dash, S. *et al.* (2019) ‘Big data in healthcare: management, analysis and future prospects’, *Journal of Big Data*, 6(1), pp. 54. Available at: https://doi.org/10.1186/s40537-019-0217-0.

66. Dopazo, J. *et al.* (2001) ‘Methods and approaches in the analysis of gene expression data’, *Journal of Immunological Methods*, 250(1), pp. 93–112. Available at: https://doi.org/10.1016/S0022-1759(01)00307-6.

67. Eisele, M. and Kappelmann-Fenzl, M. (2021) ‘NGS Technologies’, in M. Kappelmann-Fenzl (ed.) *Next Generation Sequencing and Data Analysis*. Cham: Springer International Publishing (Learning Materials in Biosciences), pp. 47–58. Available at: https://doi.org/10.1007/978-3-030-62490-3_4.

68. Fonti, V. and Belitser, E. (2017) ‘Paper in Business Analytics Feature Selection using LASSO’, in. Available at: https://www.semanticscholar.org/paper/Paper-in-Business-Analytics-Feature-Selection-using-Fonti-Belitser/24acd159910658223209433cf4cbe3414264de39 (Accessed: 22 January 2024).

69. Guyon, I. and Elisseeff, A. (2003) 'An Introduction to Variable and Feature Selection' Journal of Machine Learning Research, 3, pp. 1157–1182. https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf.

70. Hahs-Vaughn, D.L. (2023) ‘Foundational methods: descriptive statistics: bivariate and multivariate data (correlations, associations)’, in R.J. Tierney, F. Rizvi, and K. Ercikan (eds) *International Encyclopedia of Education (Fourth Edition)*. Oxford: Elsevier, pp. 734–750. Available at: https://doi.org/10.1016/B978-0-12-818630-5.10084-3.

71. Hou, J. *et al.* (2022) ‘Distance correlation application to gene co-expression network analysis’, *BMC Bioinformatics*, 23, pp. 81. Available at: https://doi.org/10.1186/s12859-022-04609-x.

72. *Initial sequence of the chimpanzee genome and comparison with the human genome | Nature* (2005). Available at: https://www.nature.com/articles/nature04072 (Accessed: 23 January 2024).

73. *Initial sequencing and analysis of the human genome | Nature* (2001). Available at: https://www.nature.com/articles/35057062 (Accessed: 23 January 2024).

74. International Human Genome Sequencing Consortium (2004) ‘Finishing the euchromatic sequence of the human genome’, *Nature*, 431(7011), pp. 931–945. Available at: https://doi.org/10.1038/nature03001.

75. Kalinin, A.A. *et al.* (2018) ‘Deep learning in pharmacogenomics: from gene regulation to patient stratification’, *Pharmacogenomics*, 19(7), pp. 629–650. Available at: https://doi.org/10.2217/pgs-2018-0008.

76. Kappelmann-Fenzl, M. (2021) ‘NGS Data’, in M. Kappelmann-Fenzl (ed.) *Next Generation Sequencing and Data Analysis*. Cham: Springer International Publishing (Learning Materials in Biosciences), pp. 79–104. Available at: https://doi.org/10.1007/978-3-030-62490-3_7.

77. LaValley, M.P. (2008) ‘Logistic Regression’, *Circulation*, 117(18), pp. 2395–2399. Available at: https://doi.org/10.1161/CIRCULATIONAHA.106.682658.

78. Liu, H. and Motoda, H. (2012) *Feature Selection for Knowledge Discovery and Data Mining*. Springer Science & Business Media.

79. Mahesh, B. (2020) 'Machine Learning Algorithms - A Review,' International Journal of Science and Research, 9(1), pp. 381–386. https://doi.org/10.21275/art20203995.

80. Marcos-Zambrano, L.J. *et al.* (2021) ‘Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment’, *Frontiers in Microbiology*, 12. Available at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.634511.

81. Mohammadi, A.T. *et al.* (2023) *The Latest Advances in Genetics and biology*. Nobel Sciences.

82. Muthukrishnan, R. and Rohini, R. (2016) ‘LASSO: A feature selection technique in predictive modeling for machine learning’, in *2016 IEEE International Conference on Advances in Computer Applications (ICACA). 2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, pp. 18–20. Available at: https://doi.org/10.1109/ICACA.2016.7887916.

83. Nick, T.G. and Campbell, K.M. (2007) ‘Logistic Regression’, in W.T. Ambrosius (ed.) *Topics in Biostatistics*. Totowa, NJ: Humana Press (Methods in Molecular Biology™), pp. 273–301. Available at: https://doi.org/10.1007/978-1-59745-530-5_14.

84. Pati, Y.C., Rezaiifar, R. and Krishnaprasad, P.S. (1993) ‘Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition’, in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*. *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pp. 40–44 vol.1. Available at: https://doi.org/10.1109/ACSSC.1993.342465.

85. Pudil, P. *et al.* (1995) ‘Feature selection based on the approximation of class densities by finite mixtures of special type’, *Pattern Recognition*, 28(9), pp. 1389–1398. Available at: https://doi.org/10.1016/0031-3203(94)00009-B.

86. Ray, S. (2019) ‘A Quick Review of Machine Learning Algorithms’, in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 35–39. Available at: https://doi.org/10.1109/COMITCon.2019.8862451.

87. Reuter, J.A., Spacek, D.V. and Snyder, M.P. (2015) 'High-Throughput Sequencing Technologies', *Molecular Cell*, 58(4), pp. 586–597. Available at: https://doi.org/10.1016/j.molcel.2015.05.004.

88. Saeys, Y., Inza, I. and Larrañaga, P. (2007) 'A review of feature selection techniques in bioinformatics', *Bioinformatics*, 23(19), pp. 2507–2517. Available at: https://doi.org/10.1093/bioinformatics/btm344.

89. Schloss, J.A. (2008) 'How to get genomes at one ten-thousandth the cost', *Nature Biotechnology*, 26(10), pp. 1113–1115. Available at: https://doi.org/10.1038/nbt1008-1113.

90. Shi, X. *et al.* (2015) 'A Framework of Joint Graph Embedding and Sparse Regression for Dimensionality Reduction', *IEEE Transactions on Image Processing*, 24(4), pp. 1341–1355. Available at: https://doi.org/10.1109/TIP.2015.2405474.

91. Shi, X. *et al.* (2019) 'Structured orthogonal matching pursuit for feature selection', *Neurocomputing*, 349, pp. 164–172. Available at: https://doi.org/10.1016/j.neucom.2018.12.030.

92. Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007) 'Measuring and testing dependence by correlation of distances', *The Annals of Statistics*, 35(6), pp. 2769–2794. Available at: https://doi.org/10.1214/009053607000000505.

93. Taunk, K. *et al.* (2019) 'A Brief Review of Nearest Neighbor Algorithm for Learning and Classification', in *2019 International Conference on Intelligent Computing and Control Systems (ICCS). 2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pp. 1255–1260. Available at: https://doi.org/10.1109/ICCS45141.2019.9065747.

94. Tsagris, M. *et al.* (2022) 'The γ-OMP Algorithm for Feature Selection With Application to Gene Expression Data', *IEEE/ACM transactions on computational biology and bioinformatics*, 19(2), pp. 1214–1224. Available at: https://doi.org/10.1109/TCBB.2020.3029952.

95. Turner, J.R. (2020) 'Area Under the Curve (AUC)', in M.D. Gellman (ed.) *Encyclopedia of Behavioral Medicine*. Cham: Springer International Publishing, pp. 146–146. Available at: https://doi.org/10.1007/978-3-030-39903-0_986.

96. V, D.B. (1991) 'Nearest neighbor (NN) norms: NN pattern classification techniques', *IEEE Computer Society Tutorial* [Preprint]. Available at: https://cir.nii.ac.jp/crid/1572261550010307072.

97. Venkatesh, B. and Anuradha, J. (2019) 'A Review of Feature Selection and Its Methods', *Cybernetics and Information Technologies*, 19(1), pp. 3–26.

98. Wettschereck, D. and Dietterich, T.G. (1995) 'An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms', *Machine Learning*, 19(1), pp. 5–27. Available at: https://doi.org/10.1007/BF00994658.

99. Xuan, J. *et al.* (2013) 'Next-generation sequencing in the clinic: Promises and challenges', *Cancer Letters*, 340(2), pp. 284–295. Available at: https://doi.org/10.1016/j.canlet.2012.11.025.

100. Yang, A. *et al.* (2020) 'Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA', *Frontiers in Bioengineering and Biotechnology*, 8. Available at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.01032.

101. Zitnik, M. *et al.* (2019) 'Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities', *Information Fusion*, 50, pp. 71–91. Available at: https://doi.org/10.1016/j.inffus.2018.09.012.

102. Euclidean Distance | Formula, Derivation & Solved Examples (2024) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/euclidean-distance/.