PREDICTING THE PRIMARY TISSUES OF CANCERS OF UNKNOWN PRIMARY USING MACHINE LEARNING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

KAMRAN KARIMOV

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE
IN
BIOINFORMATICS

DECEMBER 2023

PREDICTING THE PRIMARY TISSUES OF CANCERS OF UNKNOWN PRIMARY USING MACHINE LEARNING

submitted by **KAMRAN KARIMOV** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics Department, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç
Dean, **Graduate School of Informatics** _______________

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics** _______________

Assist. Prof. Dr. Aybar Can Acar
Supervisor, **Health Informatics, METU** _______________

Examining Committee Members:

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics, METU _______________

Assist. Prof. Dr. Aybar Can Acar
Health Informatics, METU _______________

Assoc. Prof. Dr. Bala Gür Dedeoğlu
Institute of Biotechnology, Ankara University _______________

Date: 18.01.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Kamran Karimov

Signature :

ABSTRACT

PREDICTING THE PRIMARY TISSUES OF CANCERS OF UNKNOWN PRIMARY USING MACHINE LEARNING

Karimov, Kamran
M.S., Department of Bioinformatics
Supervisor: Assist. Prof. Dr. Aybar Can Acar

December 2023, 53 pages

Cancers of Unknown Primary (CUP) origin are metastases where the primary source of the tumor cannot be detected and only the secondary tumor is evident. This can cause problems in treatment since the tissue of origin defines the base characteristics of the tumor and most therapeutic methods are specific to these characteristics. We built three machine learning models to predict the primary tissue of CUPs based on gene expression profile similarities between the primary tumor and its metastases. The models are trained on 8798 cancer cases across 14 different cancer types obtained from the TCGA program. The specific cancer types are annotated in the data used. During the process, we tried origin prediction based on specific gene types and compared these results with each other and with overall accuracy. The trained model can assist in CUP diagnoses, by further development, using more data.

Keywords: CUP prediction, cancer, machine learning, logistic regression

ÖZ

PRİMERİ BİLİNMEYEN KANSERLERİN PRİMER DOKULARININ MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TAHMİNİ

Karimov, Kamran
Yüksek Lisans, Biyoenformatik Bölümü
Tez Yöneticisi: Dr. Öğr. Üyesi. Aybar Can Acar

Aralık 2023, 53 sayfa

Primeri Bilinmeyen Kanserler (PBK), tümörün birincil kaynağının tespit edilemediği ve yalnızca ikincil tümörün belirgin olduğu metastazlardır. Köken doku tümörün temel özelliklerini tanımladığından ve çoğu tedavi yöntemi bu özelliklere özgü olduğundan, bu durum tedavide sorunlara neden olabilmektedir. Primer tümör ve onun metastazları arasındaki gen anlatım profili benzerliklerine dayanarak PBK'ların primer dokusunu tahmin etmek için üç farklı makine öğrenmesi modeli oluşturduk. Modeller, TCGA programından elde edilen 14 farklı kanser türünde 8798 kanser vakası üzerinde eğitilmektedir. Kullanılan verilerde spesifik kanser türleri belirlidir. Süreç boyunca belirli gen tiplerine dayalı olarak köken tahminini denedik, bu sonuçları birbirleriyle ve genel doğrulukla karşılaştırdık. Kurulan model, daha fazla veri sayesinde geliştirilirse PBK teşhislerine yardımcı olabilir.

Anahtar Kelimeler: PBK tahmini, kanser, makine öğrenimi, lojistik regresyon

To My Family and Loved Ones

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Assist. Prof. Dr. Aybar Can Acar for providing suggestions and feedback on my thesis study, and organizing my thesis committee. I would also like to thank Assoc. Prof. Dr.Yeşim Aydın Son, Assoc. Prof. Dr. Bala Gür Dedeoğlu for taking their time to review this work and being on my committee, and for providing their insightful comments and suggestions for both this study and future studies.

I would like express my endless gratitude my close friends Balagardash Bashirov and Berkay Günay for their support and contributions. I am grateful for my students' motivational support at ADA School in Azerbaijan.

Last, but not least, I am very grateful to have my family members' support and belief in me, throughout not only the thesis preparation period, but my whole life.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
CUP	Cancers of Unknown Primary
EMT	Epithelial-Mesenchymal Transition
ICGC	International Cancer Genome Consortium
lncRNA	long non-coding RNA
miRNA	microRNA
ML	Machine Learning
PCA	Principal Component Analysis
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TEC	Tyrosine-protein kinase
TME	Tumour MicroEnvironment

CHAPTER 1

INTRODUCTION

The aberrant and unchecked division of cells is called cancer, which is one of the leading death causes in the world. The uncontrollable dividing cells are invasive, and they may undergo a transition where they gain mesenchymal characteristics and infiltrate the circulatory systems in the body. This may lead to spreading of the malignancy to various sites in the body by a process called metastasis, hence formation of secondary tumors.

Cancers of unknown primary (CUP) are secondary tumors detected in the body, where the original site is not identified, due to either small size or loss of the primary tumor. They constitute roughly 3-5% of all cancer cases in the world [1]. Treatment methods in cancer involves prescription of drugs based on the type of the cancer, in other words, based on its primary site. The ambiguity of the primary site in CUP impedes specialized treatment, and this in turn, reduces the survival rate of the patients. According to Cancer Research UK (2023), among the people who were diagnosed with CUP in England between the years 2012 and 2016, only 16% could survive for 1 year or longer. Only 10% of them had survival rate above 3 years. According to American Cancer Society, average survival time is 9 to 12 months from the time of diagnosis with CUP.

Identification of primary site in CUP holds a great importance for an efficacious treatment process. However, traditional histological methods frequently fail at this challenge. In one review, it was reported that when immunohistochemistry was applied to identify a single tissue of origin, the prediction accuracy was 10.8-51% [**?**]. Another review reports identification in less than 30% of the cases with unknown primary [2].

Several studies have been done to investigate the genomic profile in metastases of primary tumors. There was found higher similarity between the metastases of breast cancer and its primary tumor, rather than other cancers with the same metastasis site by means of expression profile [3]. The same conclusion was reached in a similar study on breast cancer [4]. Drawbacks regarding use of immunohistochemistry in primary site prediction and promising results from gene expression-based studies lead to emergence of primary tissue identification methodologies that involve gene expression pattern. These include usage of biomarkers for lncRNAs, microRNAs and other several gene types, CUP classifiers, or integrated methods which merges expression data with immunohistochemistry or radiology results.

Main CUP classifiers that are developed involve a branch of artificial intelligence called machine learning. Machine learning (ML) constitutes a series of algorithms that take in data, learns from the data and makes predictions based on the patterns in the data. Because of the gene expression similarities of primary and secondary tumors, these models were trained on already annotated cancer data obtained from databases such as TCGA and ICGC. Different research teams have utilized DNA methylation profile, mutation profile or gene expression profile for the same goal. One of the major studies could achieve 96.70% accuracy across 32 cancer types using gene expression profiles. This result was followed by application of the model for data obtained from two different hospitals, and accuracy easily surpassed that of obtained by immunohistochemistry in the literature [5].

Similar to the other existing works, we aim to utilize the genomic profile similarity between the primary tumor and its metastases to predict the cancer type with over 90% accuracy. We built three different ML models and compared their accuracy in predicting the origin tissue across 14 different cancer types. The used models are logistic regression, support vector machine and random forest classifier. Therefore, we trained the models on already annotated cancer cases obtained from the TCGA program. The Logistic Regression model could achieve 96% accuracy, which is a competitive result in comparison to the other existing works in this subject. Apart from the overall accuracy, impact of the expression profile of specific gene types, e.g., miRNA is investigated and compared to the overall accuracy.

Although significant genomic profile similarity is proven between primary and secondary tumors, steps in the life of cancer cells, such as, epithelial-mesenchymal transition and colonization in the metastasis site results in some differences in mutation accumulation and gene expression, known as tumor heterogeneity. Another challenge is guaranteeing that the genetic expression data obtained from the patient belongs to a single type of cell, i.e., the secondary tumor. These factors can induce noise in the identification process, hence need to be approached carefully.

In Chapter 2, we review nature cancer, mutations that lead to cancer, metastasis and CUP. Here, we also explain machine learning and its applications in identifying CUP. Similar works on the topic are briefly discussed.

In Chapter 3, we present the steps of the experimental work. Gene expression data obtained from TCGA program, is cleaned and processed into a format to be input into the planned ML models.

In Chapter 4, results of the experimental work are presented. The results include the top 7 gene types that act as discriminators in identifying the cancers. Then, we compare this data with a more complex analysis which deploys whole expression profile-based classification.

In Chapter 5, we discuss the findings through comparison to other works available in literature. We also review the limitations of this work, followed by possible suggestions on further improvement of it.

CHAPTER 2

LITERATURE REVIEW

2.1 Cancer and its types

The uncontrolled development of cells is the underlying characteristic of a complicated set of disorders known as cancer [6]. In contrast to normal cells, which follow specific lifecycles, cancer cells avoid apoptosis, a type of programmed cell death, and multiply unchecked. Tumours are lumps of tissue that can cause disruptions to bodily functioning due to disobedience of growth rules. But not all cancers manifest as physical masses; leukaemia, for instance, multiply in the bone marrow and blood [7].

At the molecular level, tumorigenesis results from changes in multiple biological pathways. Several stages are involved in turning a healthy cell malignant, and genetic abnormalities are frequently the first to cause this change [8]. These mutations result from environmental exposure to cancer-causing substances like tobacco smoke or UV light, or they can be inherited. Mutations can cause tumour suppressor genes, which normally regulate cell division, to become inactive or activate oncogenes, which normally drive cell development [9].

A selection process is akin to natural evolution as these aberrant cells multiply and pick up further mutations, some of which benefit survival. Through a process known as clonal expansion, a population of cells becomes more and more different from what it was initially and more able to live, multiply, and invade other tissues [10]. The molecular profile of cancer is not constant; it varies greatly depending on the tissue from which it originated and the mutations it carries. Since what works for one patient or even one form of cancer does not work for another, this variety creates obstacles to treatment [11]. As such, molecular assessment of a patient's cancer has emerged as a key component of contemporary oncology, allowing customized therapeutic strategies [12].

The hallmarks of cancer encompass six distinct biological capacities gained throughout the complex and progressive evolution of human tumours. The factors above encompass sustaining proliferative signalling, avoiding growth suppressors, resisting the death of cells, enabling replicative immortality, generating angiogenesis, and activating invasions and metastasis [13]. Within the framework of these distinctive characteristics, scholars and practitioners can systematically classify and address the ailment with more efficacy [14].

2.1.1 Mutations and different cancer types

To comprehend the correlation between mutations and cancer, it is important to thoroughly examine the various forms of genetic modifications that can lay the foundation for this intricate ailment. Mutations can be categorized into various classes, including insertions, point mutations, deletions, and chromosomal rearrangements [15].

Point mutations represent the most elementary type of genetic alteration involving modifying, inserting, or deleting a solitary nucleotide base [16]. Mutations can significantly impact the functionality of proteins, particularly when they manifest in crucial segments of the genetic material, such as the active site of an enzyme. Insertions and deletions (indels) refer to adding or removing nucleotide bases within the DNA sequence [17]. These genetic alterations can result in frameshift mutations, causing a significant modification to the amino acid sequence located downstream of the mutation site. Chromosomal rearrangements encompass genetic alterations, including translocations, inverted positions, duplications, and large-scale deletions. These rearrangements can impair normal gene function or generate fusion genes that encode proteins with carcinogenic properties [18].

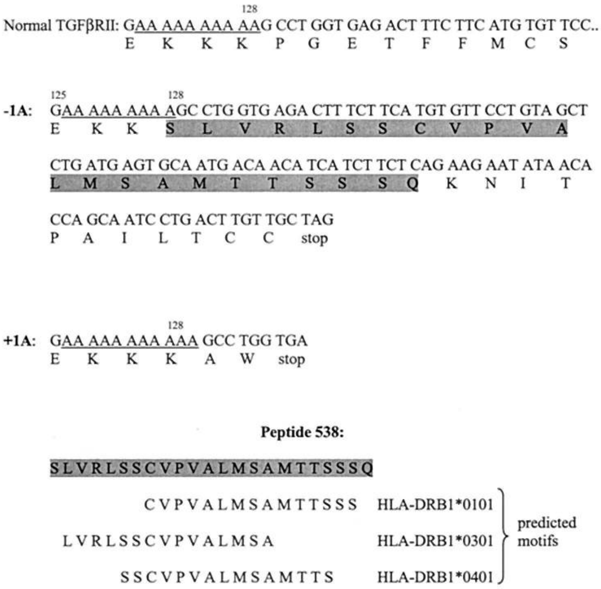

128
Normal TGFβRII: GAA AAA AAA AAG CCT GGT GAG ACT TTC TTC ATG TGT TCC..
E K K K P G E T F F M C S
125
128
-1A: GAA AAA AAA AGC CTG GTG AGA CTT TCT TCA TGT GTT CCT GTA GCT
E K K S L V R L S S C V P V A
CTG ATG AGT GCA ATG ACA ACA TCA TCT TCT CAG AAG AAT ATA ACA
L M S A M T T S S S Q K N I T
CCA GCA ATC CTG ACT TGT TGC TAG
P A I L T C C stop
128
+1A: GAA AAA AAA AAA GCC TGG TGA
E K K K A W stop
Peptide 538:
SLVRLSSCVPVALMSAMTTSSSQ
CVPVALMSAMTTSSS HLA-DRB1*0101
LVRLSSCVPVALMSA HLA-DRB1*0301
SSCVPVALMSAMTTS HLA-DRB1*0401
predicted motifs

Figure 2.1: Frameshift mutations in TGFβRII. The human gene coding for TGFβRII contains a poly(A) sequence (A10). Insertion or deletion of adenine causes a frameshift in the sequence which results in a completely different amino acid sequence downstream of the mutation. Adapted from [19]

Mutations can potentially induce oncogenes' activation or tumour suppressor genes' inactivation. Oncogenes are genetic elements that, upon undergoing mutations, can induce the transformation of healthy cells into malignant cells [14]. An illustrative example involves a point mutation occurring in the gene that encodes the RAS protein. This protein typically governs the regulation of cell development. However, the presence of a point mutation can result in a variant of the RAS protein that is persistently active, hence facilitating unregulated cell division [20]. In contrast, tumour suppressor genes, such as TP53, function as cell growth regulators by exerting inhibitory effects [21]. The loss of control over cell division and subsequent cancer development can occur when such genes are deleted or inactive [22].

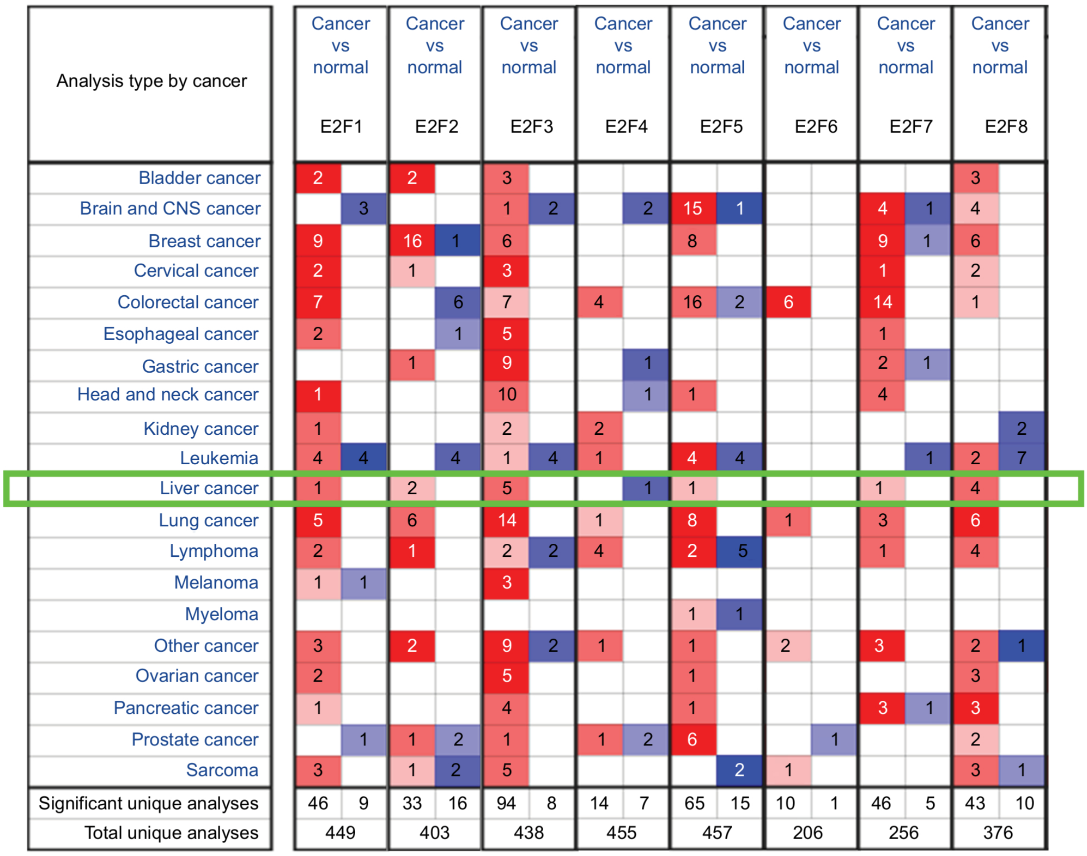
The correlation between particular genetic alterations and various forms of cancer has been well-established in the scientific literature [23]. Specific mutations are distinctive characteristics of specific malignancies, functioning as diagnostic markers and possible targets for therapeutic interventions [24]. An elevated susceptibility to breast and ovarian cancers is closely linked to genetic abnormalities occurring in the BRCA1 and BRCA2 genes. When these genes are inactive due to mutation, the cell's capacity to effectively repair DNA damage is hindered. Consequently, this raises the probability of more genetic mistakes occurring, which can ultimately result in the development of cancer [25].

The relationship between mutations and the likelihood of developing cancer is complex and multifaceted. Certain individuals who possess mutations in high-risk genes do not experience the development of cancer, suggesting that other influential elements are at play [26]. These extra factors encompass environmental exposure, and lifestyle choices, contributing significantly to cancer formation [27]. Examining these genetic alterations has enhanced the comprehension of the biological mechanisms behind cancer and has also facilitated the advancement of precise therapeutic interventions [28]. For example, PARP inhibitors are a class of medications that target the vulnerability of cancer cells with BRCA mutations, as these cells heavily depend on the PARP enzyme for DNA repair mechanisms [29]. Through the inhibition of poly (ADP-ribose) polymerase (PARP), these pharmaceutical agents successfully produce a state of synthetic lethality, thereby selectively eliminating cancerous cells harbouring BRCA mutations while preserving the viability of healthy cells [30].

2.1.2 Gene expression patterns in cancer types

Gene expression profiling has become a revolutionary method in examining cancer biology, offering an unparalleled understanding of the molecular characteristics of many forms of tumours [31]. This methodology simultaneously quantifies the expression levels of several genes, providing a comprehensive overview of the biological processes occurring within a tumour. Researchers can discern patterns that indicate malignancy by comparing gene expression profiles between malignant cells and normal tissue [32].

Every form of cancer possesses a unique gene expression profile that signifies its cellular source and the precise genetic modifications it contains [33]. For example, breast cancer cells demonstrate an elevated expression of the oestrogen receptor gene, whereas lung cancer cells manifest mutations that result in the excessive activation of growth signalling pathways [34]. The disparities mentioned have ramifications that extend beyond academia, as they significantly impact the classification and treatment of tumours [35].



Analysis type by cancer	Cancer vs normal		Cancer vs normal		Cancer vs normal		Cancer vs normal		Cancer vs normal		Cancer vs normal		Cancer vs normal		Cancer vs normal	
	E2F1		E2F2		E2F3		E2F4		E2F5		E2F6		E2F7		E2F8	
Bladder cancer	2		2		3										3	
Brain and CNS cancer		3			1	2		2	15	1			4	1	4	
Breast cancer	9		16	1	6				8				9	1	6	
Cervical cancer	2		1		3								1		2	
Colorectal cancer	7			6	7		4		16	2	6		14		1	
Esophageal cancer	2			1	5								1			
Gastric cancer			1		9			1					2	1		
Head and neck cancer	1				10			1	1				4			
Kidney cancer	1				2		2									2
Leukemia	4	4		4	1	4	1		4	4				1	2	7
Liver cancer	1		2		5			1	1				1		4	
Lung cancer	5		6		14		1		8		1		3		6	
Lymphoma	2		1		2	2	4		2	5			1		4	
Melanoma	1	1			3											
Myeloma									1	1						
Other cancer	3		2		9	2	1		1		2		3		2	1
Ovarian cancer	2				5				1						3	
Pancreatic cancer	1				4				1				3	1	3	
Prostate cancer		1	1	2	1		1	2	6			1			2	
Sarcoma	3		1	2	5					2	1				3	1
Significant unique analyses	46	9	33	16	94	8	14	7	65	15	10	1	46	5	43	10
Total unique analyses	449		403		438		455		457		206		256		376	

Figure 2.2: E2F gene expression differences across 20 different cancer types. The first part of each column corresponds to cancer tissue, whereas the second part corresponds to normal tissue. Red color indicates overexpression, and blue color indicates under-expression. Adapted from [36]

Regarding therapeutic interventions, the manifestation of specific genes can indicate a tumour's potential responsiveness to particular pharmaceutical agents [37]. An example of this is the correlation between the overexpression of the HER2 gene in breast cancer and the likelihood of a positive response to trastuzumab, a medication that specifically targets this gene [38]. The expression of the KRAS gene can determine the suitability of cetuximab treatment in colorectal cancer. The utilization of gene expression profiling holds significant value within the domain of cancers of unknown primary, wherein the specific location from which the malignancy originates remains unidentified. The treatment of these tumours poses a substantial difficulty due to the necessity of customizing therapy based on the specific tissue from which the cancer originates [39]. Machine learning algorithms can analyze intricate gene expression data and make predictions on the primary site, thus guiding treatment decisions [40].

2.2 Metastasis

Metastasis refers to the mechanism through which malignant cells metastasize from the site of origin to remote organs and tissues [41]. It is a multifaceted occurrence involving numerous stages that substantially contribute to the illness and death linked to cancer. Metastatic dissemination transpires when malignant cells acquire the capacity to infiltrate adjacent tissues, infiltrate the circulatory system, and generate secondary tumours at novel sites. This phenomenon not only signifies the intrinsic virulence of the neoplastic cells but also the dynamic interaction between said cells and the physiological milieu of the host [42].

It is impossible to emphasize the clinical impact that metastasis has. Metastatic disease, which is frequently the primary cause of mortality among cancer patients, is considerably more difficult to control. Localized therapies, such as radiation or surgery, are frequently sufficient to treat the primary tumour. Effective systemic therapies designed to prevent or treat metastatic spread require a comprehensive understanding of the mechanisms underlying metastasis. Moreover, an understanding of tumour biology, including tumour heterogeneity, modification, and evolution throughout cancer progression, is gained through the investigation of metastasis [43].

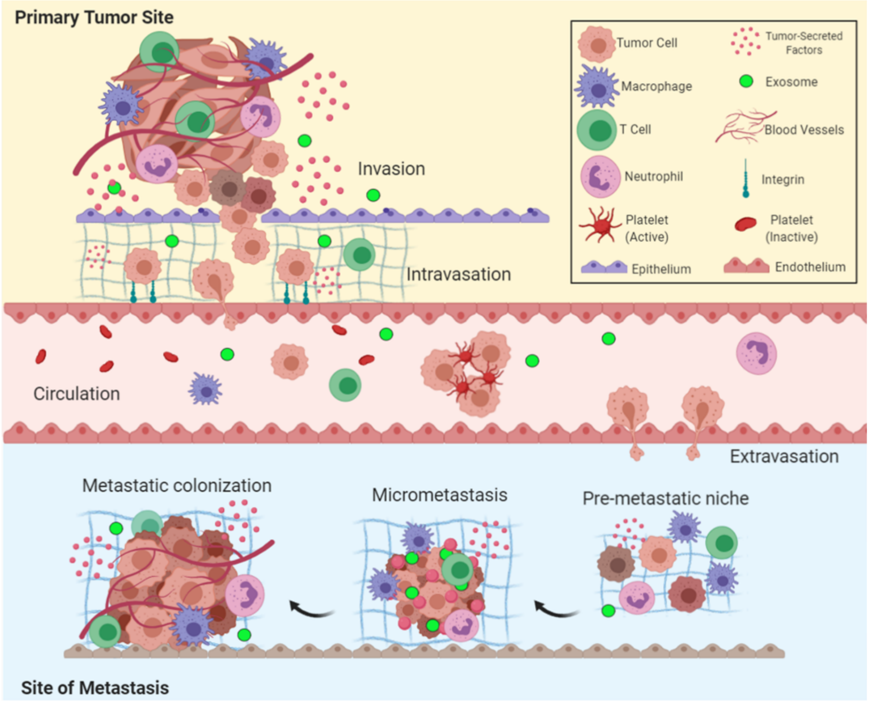


Figure 2.3: Depiction of metastasis. Metastatic cascade includes local invasion, intravasation, transport via the circulatory systems, extravasation, and colonization. Adapted from [44]

2.2.1 Mechanisms of Metastasis

Metastasis comprises a series of critical stages: local invasion, intravasation, transport via the circulatory systems, extravasation and colonization. Cancer cells invade adjacent tissue after detaching from the primary tumour and degrading extracellular matrix barriers. Subsequently, cancer cells enter blood vessels and lymphatic systems via intravasation. After entering the bloodstream, these cells must endure the hostile environment, avoid immune monitoring, and ultimately leave the circulation via extravasation at remote sites. Establishing secondary tumours in novel organ sites constitutes the ultimate stage, colonization, which necessitates adjustment and transformation of the microenvironment [45].

The tumour microenvironment (TME) is an indispensable factor in each stage of the metastatic progression. Fibroblasts, immune cells, endothelial cells, and other noncancerous cells are among its constituents. Signalling molecules and extracellular matrix elements are also present. The TME and cancer cell interaction promote the metastatic cascade. TME can facilitate intravasation and extravasation, promote motility and invasiveness (via EMT), and provide a conducive niche for the survival and proliferation of metastatic cells, among other modifications. Furthermore, the TME play a role in the emergence of therapy resistance, thereby complicating the management of metastatic disease [46].

2.2.2 Epithelial-Mesenchymal Transition (EMT)

The epithelial-to-mesenchymal transition (EMT) is a biological mechanism by which an epithelial cell acquires the phenotype of a mesenchymal cell through numerous biochemical changes. This metamorphosis leads to enhanced invasiveness, migratory capability, and resistance to apoptosis. A network of signalling pathways involving numerous molecules, including growth factors (e.g., TGF-), transcription factors (e.g., Snail, Slug, Twist), and microRNAs, regulates the induction of EMT in cancer cells. These molecules induce an upregulation of mesenchymal markers (e.g., vimentin) and a downregulation of epithelial markers (e.g., E-cadherin). All of these factors contribute to the EMT process: reorganization of cytoskeletal structures, modification of cell-matrix interactions, and alteration of cell-cell adhesion properties result from activating these pathways [47].

The successful accomplishment of EMT confers enhanced migratory and invasive capabilities to cancer cells. By reorganizing the extracellular matrix, these transformed cells facilitate passage through tissues. Additionally, they demonstrate enhanced resistance to necrosis and can evade the immune system with greater efficacy. The heightened invasive capability is pivotal in advancing localized tumours towards malignant metastasis [48].

2.2.3 Challenges in Treating Metastatic Cancer

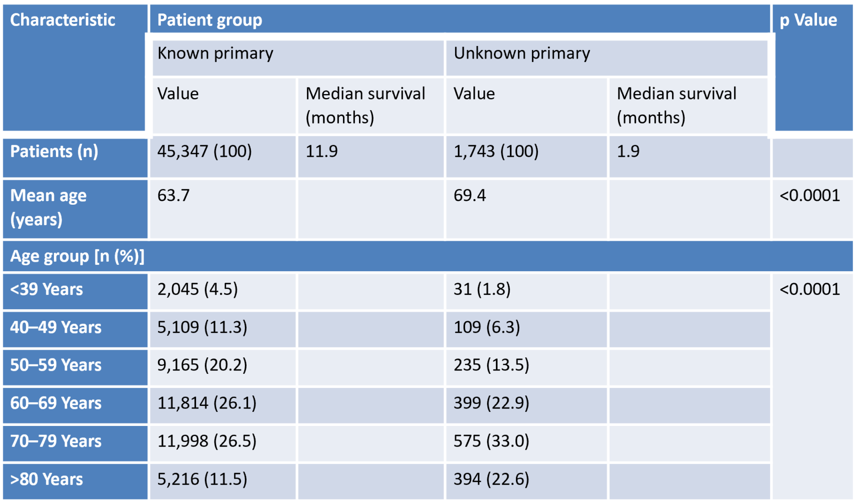
Detecting and selectively targeting metastatic cells represents a fundamental obstacle in treating metastatic cancer. In addition to their low abundance of specific biomarkers, these cells frequently metastasize throughout the body before diagnosing the primary tumour, making their detection difficult [49]. Moreover, the presence of metastatic cells in various anatomical sites, which are frequently inaccessible, complicates the implementation of targeted therapy [50].

Resistance to conventional therapies is a common characteristic of metastatic cancer cells. Acquired through selective pressures exerted by therapeutic agents or intrinsic genetic and epigenetic variations present within the tumour, resistance can manifest in various ways [51]. Further complicating the development of successful treatments are the heterogeneity, adaptability, and evolution of metastatic tumours in response to the surroundings and treatment. In the effective treatment of metastatic cancer, overcoming these resistance mechanisms continues to be a formidable obstacle [52].

2.3 Cancers of Unknown Primary Origin

Cancers of unknown primary origin (CUP) are a heterogeneous group of metastatic malignancies for which, despite extensive clinical investigation, the primary site of origin remains unidentified. They constitute an estimated 3-5% of the total number of cancer diagnoses and pose a substantial clinical complication. CUPs exhibit various histological characteristics, aggressive behaviour, and early dissemination. In the absence of a detectable primary tumour, late-stage presentations are common; therefore, these malignancies constitute an essential domain of oncological investigation [49].

Table 2.1: Survival rate comparison between cancers of known and unknown primary in Ontario. Adapted from [53]

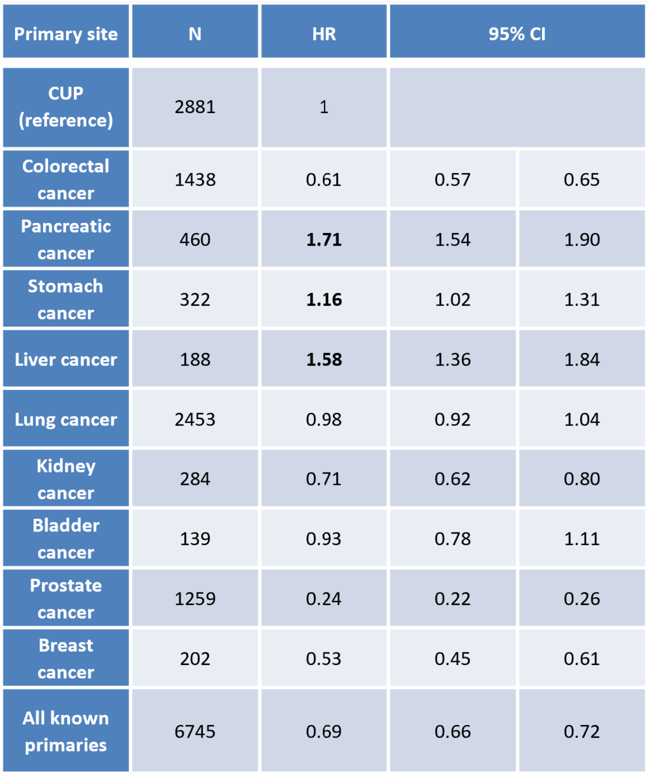


Characteristic	Patient group				p Value
	Known primary		Unknown primary		
	Value	Median survival (months)	Value	Median survival (months)	
Patients (n)	45,347 (100)	11.9	1,743 (100)	1.9	
Mean age (years)	63.7		69.4		<0.0001
Age group [n (%)]					
<39 Years	2,045 (4.5)		31 (1.8)		<0.0001
40–49 Years	5,109 (11.3)		109 (6.3)		
50–59 Years	9,165 (20.2)		235 (13.5)		
60–69 Years	11,814 (26.1)		399 (22.9)		
70–79 Years	11,998 (26.5)		575 (33.0)		
>80 Years	5,216 (11.5)		394 (22.6)		

CUP is typically diagnosed exclusionarily, following the exclusion of known primary malignancies via an extensive battery of diagnostic tests. Diagnosis by exclusion is carried out when scientific knowledge is scarce. Due to this diagnostic ambiguity, treatment planning is significantly complicated. Conventional cancer therapies frequently target particular cancer types; therefore, treatment options become constrained and generally less productive in the absence of knowledge regarding the primary site. Furthermore, the heterogeneous characteristics of CUPs introduce an extra stratum of intricacy, given that individuals with CUP exhibit varying responses to identical treatment protocols [54].

In CUP cases, it is critical to identify the primary tissue of origin to develop potentially more effective and individualized treatment strategies. By considering the distinct attributes of the tissue thought to be the source of the malignancy, it is possible to optimize therapeutic interventions. For example, therapeutic approaches for breast cancer and lung cancer are notably distinct; therefore, ascertaining the probable aetiology can facilitate the implementation of treatment protocols that are more suitable and productive [55]. In addition, comprehension of the primary site can aid in anticipating possible metastatic patterns and surveillance for recurrence, thus enabling more proactive and individualized patient care [56].

Table 2.2: Comparison of hazard ratio between various cancer types and CUPs. Patients were diagnosed and died between 2002 and 2008. Only patients with one primary cancer and a positive metastasis status were considered. Adapted from [57]



Primary site	N	HR	95% CI	
CUP (reference)	2881	1		
Colorectal cancer	1438	0.61	0.57	0.65
Pancreatic cancer	460	**1.71**	1.54	1.90
Stomach cancer	322	**1.16**	1.02	1.31
Liver cancer	188	**1.58**	1.36	1.84
Lung cancer	2453	0.98	0.92	1.04
Kidney cancer	284	0.71	0.62	0.80
Bladder cancer	139	0.93	0.78	1.11
Prostate cancer	1259	0.24	0.22	0.26
Breast cancer	202	0.53	0.45	0.61
All known primaries	6745	0.69	0.66	0.72

2.3.1 Historical Approaches to Identifying CUPs

At one time, conventional diagnostic techniques were predominantly utilized to ascertain the primary site of cancers of unknown primary (CUPs). This process included physical examinations, histopathological analysis, and imaging modalities such as CT, MRI, and PET scans [58]. Histopathology, which entailed the investigation of biopsy samples under a microscope, served as the foundation of these endeavours. Pathologists analyze tissue attributes, including cell variety and structure, to formulate informed conjectures regarding the potential genesis of the malignancy. The conventional approaches to tracing the origins of CUP encountered considerable constraints and difficulties. Numerous conventional diagnostic instruments were deficient in specificity. Tumour markers, for example, exhibited variability across conditions and were not solely associated with a specific form of cancer; thus, the outcomes were indeterminate [59]. In one review, it was reported that when immunohistochemistry was applied to identify a single tissue of origin, the prediction accuracy was 10.8-51% [**?**]. Another review reports identification in less than 30% of the cases with unknown primary. This result was also overshadowed by the existence of inconsistencies with other big immunohistochemistry studies [2].

Biopsies and similar invasive procedures posed potential challenges for patients, particularly those who were in the advanced stages of their illnesses. Additionally, the utilization of these traditional approaches to diagnosis frequently resulted in lengthy process times, which caused a delay in the commencement of treatment [60].

2.3.2 Genomic Profiling in CUP Identification

Profiling of the genome has emerged as a fundamental technique in the study of CUPs. Practical genomic profiling, which identifies distinctive genetic signatures that may indicate the origin of a tumour by sequencing significant portions of the genome, has been the subject of numerous studies [61]. Certain investigations have examined the mutational landscape of CUPs, wherein they have identified distinct patterns that exhibit correlations with established primary tumour sites. Similarity between primary and secondary tumours extends to gene expression profiles. It was found that breast cancer primary tumours grouped with their metastases when clustering was applied [3]. Individuals have also identified potential therapeutic targets and deduced a tumour's source through targeted gene panels arranged to detect actionable mutations [62].

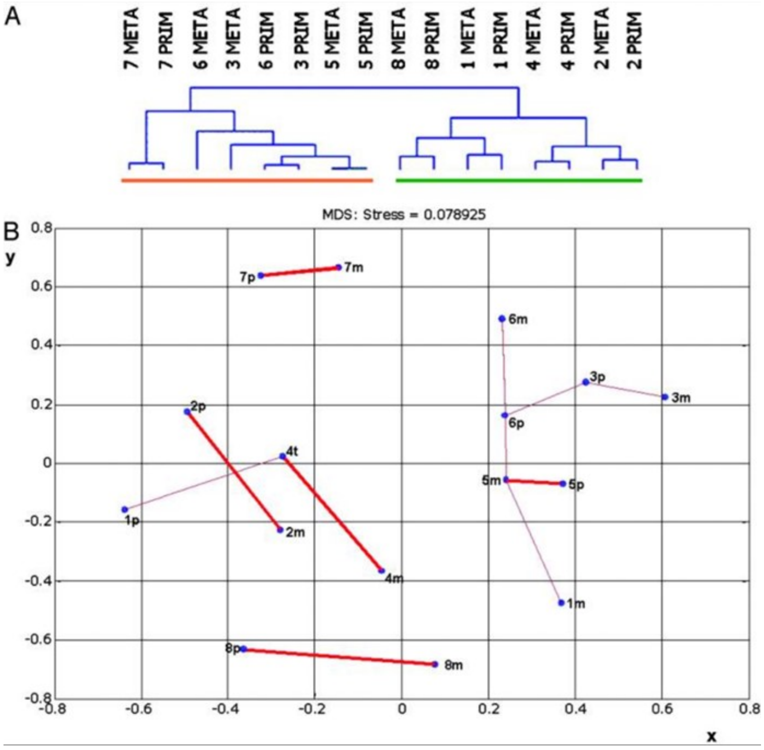


Figure 2.4: Dendrogram showing the relationship between 8 primary tumors and their metastases by means of gene expression profiles [3].

The efficacy of genomic profiling in detecting CUP has been noteworthy, specifically in situations where conventional diagnostic approaches failed to reach a definitive conclusion. Genomic profiling can comprehensively comprehend the tumour's molecular composition, thereby presenting insights that facilitate a more precise localization of the main site. However, this approach is not without its limitations [63]. Tracing genomic signatures to a primary site is not always possible for all tumours, and the analysis of genomic information can be difficult and demands specialized knowledge. Moreover, certain clinical environments may be unable to afford comprehensive genomic profiling due to its prohibitive cost and limited accessibility. Notwithstanding these obstacles, genomic profiling remains a potent instrument in examining CUPs, substantially contributing to developing personalized cancer treatment approaches [64].

2.4 The Cancer Genome Atlas (TCGA)

Cancer genome databases are indispensable in the ongoing battle against the disease. They function as extensive storage of genetic data, offering comprehensive analyses of the genomic modifications distinctive of numerous types of cancer [65]. These tools empower scientists to conduct extensive investigations, establishing correlations between genetic variations and cancer phenotypes that are statistically significant [66].

The National Human Genome Research Institute (NHGRI) and the National Cancer Institute (NCI) joined forces in 2006 to establish The Cancer Genome Atlas (TCGA). The objective was to promote a more profound comprehension of cancer and assist in the formulation of treatment approaches that are more effective. TCGA swiftly broadened its scope to incorporate more than 33 cancer types, including hematologic malignancies and solid tumours [67]. About cancer genome databases, TCGA is among the most expansive and all-encompassing on a global scale. The database contains gene expression profiles, millions of genetic mutations, copy number variations, and epigenetic modifications, among other information of an unprecedented magnitude. This extensive repository has emerged as a fundamental component of cancer genomic research, providing scientists and clinicians around the globe with a wealth of information to comprehend the molecular underpinnings of cancer and devise targeted therapies [68].

2.4.1 Content and Data Types in TCGA

The exhaustive compilation of genomic data types contained in TCGA offers a wide-ranging overview of genetic alterations associated with cancer. Comprehensive data on genetic mutations, such as single nucleotide polymorphisms (SNPs) and structural variations within cancer genomes, is made available through Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES). TCGA obtains insights into the contributions of gene expression alterations to cancer by capturing the presentation levels of genes across various cancer types via RNA sequencing (RNA-seq). This contains data regarding histone modifications and DNA methylation patterns, which are essential for modulating gene expression without modifying the underlying DNA sequence [69].

In addition to genomic information, the TCGA incorporates a wealth of clinical data and metadata, encompassing:

Patient Demographics: Details including gender, age, and ethnicity.

Clinical Outcomes: Survival rates, treatment response information, and recurrence data.

Pathological Information: Specifics regarding the tumour's grade, histology, and staging.

Environmental and Lifestyle Factors: Information regarding alcohol consumption, smoking status, and other pertinent variables.

2.4.2 TCGA's Contribution to Cancer Research

TCGA has propelled the comprehension of the molecular biology of cancer to tremendous heights. It has facilitated the identification of many molecular subtypes and genetic mutations that span varieties of cancer. The revelation of particular gene mutations, such as BRAF in melanoma and IDH1 in glioma, has significantly transformed our comprehension of these malignancies [70].

The extensive data from TCGA has provided insights into intratumor heterogeneity, which refers to the genetic diversity within tumours. This comprehension is essential in developing targeted, more efficacious treatment strategies [71]. By analyzing genomic data, TCGA has assisted in identifying potential new drug targets and critical pathways involved in oncogenesis. As an illustration, elucidating the function of the PI3K/AKT/mTOR pathway in numerous malignancies has facilitated the creation of targeted therapeutic interventions. The utilization of TCGA data has been crucial in numerous scientific investigations, including the following: The application of TCGA data in research has facilitated the development of targeted therapies by enhancing our comprehension of the molecular categories of breast cancer [72]. Utilizing TCGA data, researchers have delineated the genomic environment of colorectal cancer, identifying biomarkers for treatment response and disease progression and novel mutations [73].

Oncologists can develop more individualized treatment strategies by utilizing TCGA data to characterize tumours at the molecular level. It is possible to match patients with particular genetic mutations with targeted therapies with a higher probability of success [61].

2.4.3 Other Relevant Databases and Comparative Analysis

The mission of the International Cancer Genome Consortium (ICGC) is to comprehend the genomic alterations that occur in various forms of cancer. The objective is to produce genomic data of superior quality encompassing a wide range of cancer types and populations. A multi-institutional effort, AACR Project GENIE (Genomics Evidence Neoplasia Information Exchange) consolidates clinical and genomic cancer data to accelerate translational research and enhance patient treatment decision-making [74].

Although all three databases contain genomic sequencing data, TCGA offers a more extensive collection of data types, such as epigenomic and transcriptomic information. ICGC provides comparable categories of genomic data, albeit with a greater emphasis on global diversity. Genomic data associated with clinical outcomes constitute the forefront of GENIE's efforts.

2.5 Machine Learning and Cancer Biology

A subfield of artificial intelligence (AI), machine learning (ML) is concerned with creating statistical models and algorithms that enable computers to execute operations without explicit commands. The process entails instructing computers to acquire knowledge from data and generate predictions or decisions. Algorithms capable of machine learning can discern patterns and insights within massive datasets; their performance and precision improve with time and exposure to additional data. This discipline integrates computer science, statistics, and data analytics components to develop models capable of efficiently processing intricate data [75].

2.5.1 Introduction to Machine Learning in Cancer Biology

ML is progressively emerging as an indispensable instrument in cancer biology [76]. Conventional analytical techniques are frequently insufficient in light of the proliferation of high-throughput methods and the accumulation of enormous datasets (e.g., genomic, transcriptomic, and proteomic data) in cancer research. In response to this deficiency, ML offers robust capabilities for analyzing, interpreting, and extracting significant insights from such intricate datasets. Its applications range from predicting responses from patients to various treatments to identifying genetic mutations associated with malignancy [77]. ML plays a pivotal role in cancer biology by not only comprehension of the molecular mechanisms that underlie cancer but also facilitating the formulation of individualized treatment approaches—thereby enhancing patient care and outcomes. As we move towards more data-driven, precise, and personalized oncology, this incorporation of machine learning into cancer research signifies a paradigm shift [78].

ML operates on the fundamental principle of constructing models capable of receiving input data, predicting outputs through statistical analysis, and updating outputs in real-time as new data becomes accessible. Using algorithms and extensive data sets, these models are "trained" to acquire the necessary skills to execute the given task [79].

2.5.2 Types of Machine Learning Algorithms and Their Brief Comparison

Machine learning algorithms can be broadly classified into three categories: supervised learning, unsupervised learning and reinforcement learning.

In unsupervised learning, the algorithms undergo training using annotated data, which consists of input-output pairs supplied to the model. As the model acquires the ability to map inputs to outputs, it becomes viable for classification and regression tasks. Some examples include neural networks, machines with support vectors, and decision trees [80]. It is ideal in applications for forecasting future events using historical data. The utilization of known datasets is prevalent in cancer diagnosis and prognosis.

Unsupervised learning involves utilizing unlabeled data to train the model. It attempts to detect patterns or grouping inherent in the data, which benefits association and clustering duties. Principal component analysis (PCA) and k-means clustering are two prevalent algorithms [81]. It is optimal for investigating the structure and distribution of data, such as unlabeled identification of cancer subtypes or patterns [82].

By utilizing a system of penalties and rewards, reinforcement learning forces the computer to solve a problem independently. In robotics or gaming, for instance, it is especially beneficial when the algorithm must make choices with uncertain outcomes [83]. Despite its rarity, while not prevalent in cancer research, this approach holds promise in domains such as adaptive therapies and treatment regimen optimization—where data is scarce, and decision-making is sequential [84].

2.5.3 Machine Learning in Identifying Primary Tissue

Machine learning (ML) has significantly transformed the domain of cancer studies and diagnosis through its wide-ranging applications. It facilitates early identification of cancer through the more precise analysis of medical images, including mammography, MRIs, and CT scans, compared to conventional methods [85]. In these images, ML algorithms can detect subtle patterns that may signify the existence of tumours. ML also facilitates the automated classification and analysis of samples of tissues in diagnostic pathology, thereby enhancing the speed and precision of cancer diagnoses [86].

Asserting the principal tissue in CUP cases may involve applying various machine-learning techniques. These encompass supervised learning algorithms such as random forests and support vector machines, which can be trained on established cancer cases to categorize CUPs according to their molecular and genetic characteristics. The application of deep learning methodologies, specifically convolutional neural networks, to analyze intricate patterns in imaging data is growing in popularity. Additional methods employ clustering and reduction of dimensionality algorithms to identify novel subtypes of cancer that may be associated with particular primary sites [87].

An additional noteworthy investigation utilized deep learning techniques, specifically neural networks, to examine data on gene expression. This methodology facilitated the detection of intricate patterns and intergeneric interactions, thereby augmenting the precision of primary tissue prognosis [88].

Research at CUP has also demonstrated the potential of combining machine learning with radiomics to extract quantitative characteristics from medical images. Extraction of an exhaustive collection of radiomic characteristics from imaging data (such as CT scans or MRIs) of patients with previously diagnosed primary malignancies has been the focus of research in this field [89]. The features were subsequently correlated with particular forms of cancer using machine learning models. When applied to

CUP cases, these models could suggest potential primary sites based on imaging characteristics, thereby facilitating differential diagnosis.

2.5.4 CUP-AI-Dx

The number of studies employing machine learning (ML) to forecast CUPs aetiology has increased significantly in recent years. In one article, authors utilized 1D convolutional neural networks to develop an RNA-classifier [5]. The model was named CUP-AI-Dx, and it was trained on transcriptome profiling of 18,217 patient entries from TCGA and ICGC. It was able to classify 32 different cancer types from TCGA database with 96.70% accuracy. The model was further applied to patient data from two unrelated clinical institutions, and authors had obtained 86.96% and 72.46% accuracy.

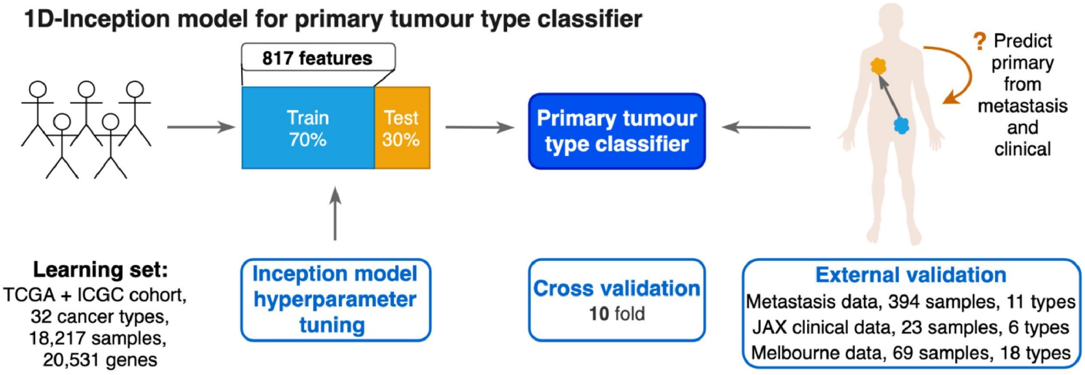

1D-Inception model for primary tumour type classifier
817 features
Train 70%
Test 30%
Primary tumour type classifier
? Predict primary from metastasis and clinical
Learning set:
TCGA + ICGC cohort,
32 cancer types,
18,217 samples,
20,531 genes
Inception model hyperparameter tuning
Cross validation
10 fold
External validation
Metastasis data, 394 samples, 11 types
JAX clinical data, 23 samples, 6 types
Melbourne data, 69 samples, 18 types

Figure 2.5: Prediction workflow of CUP-AI-Dx model. The model is trained by TCGA and ICGC data, and tested on both the same data and external data. Adapted from [5]

2.5.5 Integrative Approaches Combining Multiple Data Types

An emerging trend in CUP identification is the integrative approach, which integrates genomic, transcriptomic, and clinical data for a more comprehensive analysis [90]. By leveraging the capabilities of genomic sequencing to detect mutations in genetics transcriptomic analysis to comprehend gene expression patterns and clinical data to offer contextual insights about the tumour, this multidimensional approach effectively utilizes its potential. Several studies, for instance, have integrated these data categories using sophisticated bioinformatics tools, resulting in the development of complex models that can more accurately predict the primary site of CUPs than based on only one of the data types. Adopting an integrative approach, this method provides a better comprehension of the tumour's biology by taking into account not only its genetic composition but also its behaviour and interactions with the patient's body [91].

CHAPTER 3

MATERIALS AND METHODS

3.1 Data retrieval and processing

The used dataset is provided by GDC data portal of National Cancer institute. It hosts around 89000 cases aggregated across 82 projects. For consistency, cases are limited to only gene expression quantification data from TCGA program. Gene expression quantification refers to transcript counts across numerous genes in a cell.

GDC data web portal is used to download the data, alongside provided terminal client. Firstly, user needs to go to the portal and select appropriate data type ("transcriptome profiling" under data category and "Gene Expression Quantification" under data type) under "files" tab and select cancer type and program. After that, portal generates a manifest file which can be used by the client program.

The client downloads raw files which need further processing to be useful for data analysis. Therefore, python scripts are created to massage them into right format. Individual's case file is in tab separated values (tsv) format with metadata attached to it at the beginning. Then, each row in a file corresponds to a single gene and its "unstranded", "stranded_first", "stranded_second", "tpm_unstranded", "fpkm_unstranded", "fpkm-_uq_unstranded" expression values. In total, 8798 cases are downloaded which span 14 different cancer types.

Since these columns are highly correlated values and can be derived from each other, only "tpm_unstranded" values are used, and the rest is discarded. Moreover, there are about 60000 transcripts belonging to 40 gene types in the dataset and loading them all would make data frames too big, so, a script is made so that it creates a data frame for single gene type which can later be merged together if needed. These 40 gene types refers to the 40 unique values under the column named "gene_types" in RNA sequencing files. After the data is processed into a usable format, it is split into train, validation and test parts (80%-10%-10% of data respectively, see Table 3.1 for number of cases). Only training data is used to fit the models, and validation data is used to make decisions on feature selection and hyperparameter tuning which will be detailed on later sections. Test data will only be used to report on performance of the final models. This ensures that no bias is introduced while developing the models. For processing the data, visualization and building of models, python programming language was used alongside pandas, scikit-learn and seaborn libraries.

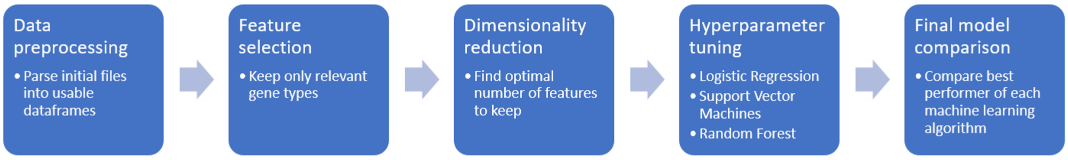

Data preprocessing
• Parse initial files into usable dataframes
Feature selection
• Keep only relevant gene types
Dimensionality reduction
• Find optimal number of features to keep
Hyperparameter tuning
• Logistic Regression
• Support Vector Machines
• Random Forest
Final model comparison
• Compare best performer of each machine learning algorithm

Figure 3.1: Overview of methodology employed.

Data dimensionality is reduced through use of feature selection and dimensionality reduction steps to discard irrelevant gene types and any features that do not help the models in those gene types. The overview of all the steps is given on Figure 3.1 and will be explained in detail in next sections.

Table 3.1: Number of cancer types for training, validation and test data.

	Train	**Validation**	**Test**	**Total**
bladder	344	46	41	**431**
brain	566	68	77	**711**
breast	969	130	133	**1232**
colon	435	49	43	**527**
corpus uteri	464	59	58	**581**
kidney	823	101	106	**1030**
liver	370	43	51	**464**
lung	935	118	101	**1154**
ovary	346	40	44	**430**
pancreas	143	18	22	**183**
prostate gland	439	57	58	**554**
skin	385	50	38	**473**
stomach	375	30	49	**454**
thyroid gland	444	71	59	**574**
Total	**7038**	**880**	**880**	**8798**

3.2 Feature selection

There are 40 gene types in the whole data and not all of them are useful. A feature selection step is applied involving a simple classification method on each gene type to select the most useful gene types. The goal of this initial classifier is not to produce the best results but provide clues to what features to keep. Logistic regression classifier is used in this step as it is the simplest model and does not require any hyperparameter tuning. The process is as follows:

1. Load data for single gene type.

2. Fit PCA and keep top 20 components.

3. Train logistic regression on PCA features.

4. Repeat steps 1-3 for all 40 gene types.

5. Rank the gene types based on initial accuracy scores.

6. Select the most important gene types that rank at the top.

Between steps 2 and 3, Min-Max scaling is applied to the features before feeding it into a model. It sets maximum and minimum value of a feature into user defined range and linearly scales intermediate values. This step does not change the underlying data as it is simply scaling the values. However, it helps the models during the learning stage since large positive and negative values can lead to numerical instabilities in the algorithms. All the features in the dataset are “squished” into [-5,5] range. The process is shown pictorially on Figure 3.2.

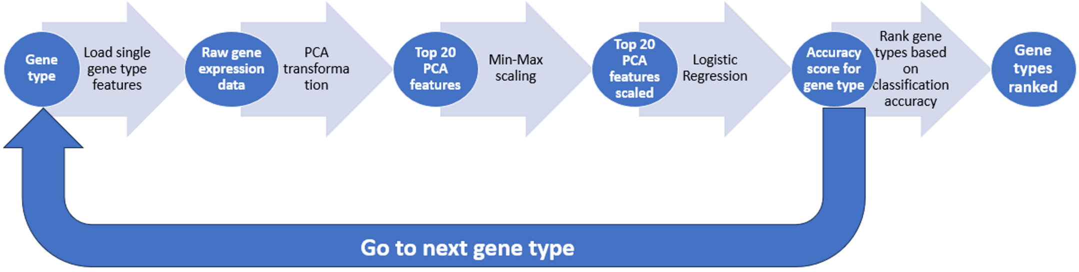


Figure 3.2: Depiction of feature selection process.

After the gene types are selected, the correlation matrix which shows the correlation coefficient between each pair of the top PCA feature of each gene type are examined. Correlation coefficient ranges from -1 to 1 and indicates how related two variables are. Value closer to 1 means that when one variable increases, the other variable increases as well. Negative values indicate variables move in opposite direction, when one increases, the other decreases. Correlation coefficient of 0 indicates that the variables are independent. It is usually beneficial to discard highly correlated variables in the dataset before feeding them into learning methods. Therefore, this step is used to ensure that the chosen gene types do not show any high correlation.

3.3 Dimensionality reduction

Once the relevant gene types are selected, their raw data are concatenated into a final form. It results in a data frame with more than 41000 gene features (transcripts). Therefore, another PCA is applied again to the combined dataset to reduce dimensionality. A logistic regression is trained using top 20 feature of the obtained dataset (same min-max scaling in the feature selection was also applied here as well) and accuracy is compared against that of models in the feature selection to make sure the performance is not deteriorating. After that, more PCA features are included progressively (until 120), and performance of the classifier is monitored. "Elbow method" is used to choose the number of features to keep. Using this method, one looks for a point where the accuracy plateaus. After choosing the appropriate number of components to keep, other machine learning methods are applied to the dataset and best performing hyperparameters are selected using validation data.

3.4 Modelling

Machine learning methods further employed in this work are Logistic Regression with L2 regularization, Support Vector Machine (SVM) classifier and Random Forest classifier. scikit-learn package provides parameter C to control regularization strength, higher values meaning weaker regularization. Its value was varied between 0.2 and 10 to tune logistic regression and SVM classifiers. In addition, linear and radial-basis function kernels were tried when choosing appropriate SVM model. For Random Forest models two hyperparameters were varied, number of trees in ensemble - between 50 and 500 and maximum depth of each tree – between 10 and 40. Based on their performance on the validation data, best hyperparameter combination are chosen for these three methods and their accuracy are presented using test data at the end.

3.5 Measuring classification performance

Various classification algorithms are intended to be used on this dataset and their performances are compared. One of the issues with the data is class imbalance. Class imbalance makes it hard for the learner and it also requires caution with accuracy analysis. Simple classification accuracy may not reflect the true picture. Therefore, in addition to classification accuracy, so-called confusion matrix is reported as well. Confusion matrix can be used to calculate the precision and the recall values. Precision refers to the proportion of actual positive classes among cases identified as positive by a model, while recall measures the proportion of positive cases identified by the model among all positive cases.

This provides greater insight into the behavior of our models from clinical standpoint as we can assess, for example, how confident one can be when identifying certain cancer type, as well as coverage of the given cancer type in the population. Precision and recall of classification model can be combined to estimate F1 score which is a value between 0 and 1. Higher values of F1 score corresponds to a better classifier. F1 score will be used as the main benchmark when discussing relative performance of the various models.

CHAPTER 4

RESULTS

4.1 Accuracy comparison across single gene type-based classifications

There were about 60000 different transcripts for each cancer patient in the data. PCA was applied across 40 gene types, and using 20 features, accuracy was obtained as seen in Figure 4.1. 7 out of 40 gene types outperform the others significantly by means of the accuracy they could yield (Figure 4.1). These include long non-coding RNA (lncRNA) genes, tyrosine-protein kinase (TEC) genes, protein coding genes, three pseudogene types and microRNA genes. In the further steps of the procedure, these 7 gene types were utilized.

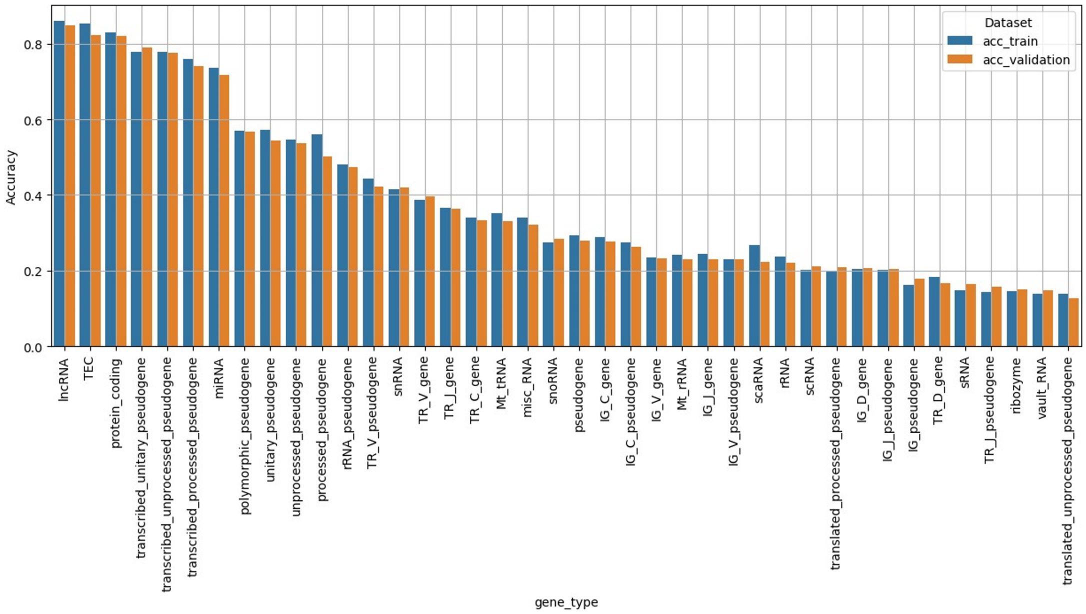


Figure 4.1: The bar plot shows accuracy of logistic regression when each gene type with 20 features was used.

In the next step, to check the significance of specific gene types in varying cancer types, the top 3 gene types were selected to see their importance in defining the cancer type. From Table 4.1, it can be seen that long non-coding RNA genes expression constitutes a notable part in transcriptome profile of brain and prostate gland cancer. Prediction solely based on lncRNA could forecast 96% and 95% of all brain and prostate cancers respectively with no false positive results. On the other hand, pancreatic cancer does not have a type-specific profile of lncRNA transcription as only 11% of the whole type could be identified. The highest false positive rate was obtained for lung cancer, due to wrong classification of numerous bladder, breast, pancreatic and thyroid gland cancer cases.

Table 4.1: Confusion Matrix for long non-coding RNA expression data across 14 cancer types with recall and precision values.

		PREDICTION														
		bladder	**brain**	**breast**	**colon**	**corpus uteri**	**kidney**	**liver**	**lung**	**ovary**	**pancreas**	**prostate gland**	**skin**	**stomach**	**thyroid gland**	**recall**
ACTUAL	**bladder**	**34**	0	0	1	0	0	0	11	0	0	0	0	0	0	0.74
	brain	0	**65**	0	0	1	0	0	0	2	0	0	0	0	0	0.96
	breast	0	0	**103**	0	0	1	0	23	1	0	0	0	2	0	0.79
	colon	0	0	0	**47**	1	0	0	1	0	0	0	0	0	0	0.96
	corpus uteri	0	0	1	2	**48**	2	1	3	1	0	0	0	1	0	0.81
	kidney	1	0	1	0	0	**92**	1	4	0	0	0	0	0	2	0.91
	liver	0	0	0	0	1	0	**39**	3	0	0	0	0	0	0	0.91
	lung	5	0	1	4	3	1	0	**98**	0	0	0	1	3	2	0.83
	ovary	0	0	0	0	6	0	0	1	**32**	0	0	0	1	0	0.80
	pancreas	0	0	0	1	1	0	0	14	0	**2**	0	0	0	0	0.11
	prostate gland	0	0	0	0	0	0	0	3	0	0	**54**	0	0	0	0.95
	skin	0	0	3	0	0	0	0	1	0	0	0	**46**	0	0	0.92
	stomach	1	0	0	2	0	0	0	2	0	0	0	0	**25**	0	0.83
	thyroid gland	0	0	0	0	0	1	0	8	0	0	0	0	0	**62**	0.87
	precision	0.83	1.00	0.94	0.82	0.79	0.95	0.95	0.57	0.89	1.00	1.00	0.98	0.78	0.94	**0.85**

The same two cancer types scored the top results, when the prediction was based on tyrosine-protein kinase (TEC) gene expression quantification. Thyroid gland cancer could be identified 93% of the time, with 0.93 precision. Pancreatic cancer and bladder cancer scored the lowest, where only 30% and 11% of the whole typeset could be identified respectively. The results also represent high value of false positives (Table 4.2).

Table 4.2: Confusion matrix for tyrosine-protein kinase expression data across 14 cancer types with recall and precision values.

		PREDICTION														
		bladder	**brain**	**breast**	**colon**	**corpus uteri**	**kidney**	**liver**	**lung**	**ovary**	**pancreas**	**prostate gland**	**skin**	**stomach**	**thyroid gland**	**recall**
ACTUAL	**bladder**	**14**	0	10	2	1	0	0	19	0	0	0	0	0	0	0.30
	brain	1	**66**	0	0	1	0	0	0	0	0	0	0	0	0	0.97
	breast	4	1	**100**	2	8	5	0	7	1	0	1	0	0	1	0.77
	colon	0	0	1	**46**	0	0	0	0	0	0	0	0	2	0	0.94
	corpus uteri	0	0	6	2	**46**	3	0	0	1	1	0	0	0	0	0.78
	kidney	1	0	1	0	2	**96**	0	0	1	0	0	0	0	0	0.95
	liver	0	0	4	0	1	1	**35**	1	1	0	0	0	0	0	0.81
	lung	3	0	2	5	0	1	1	**106**	0	0	0	0	0	0	0.90
	ovary	0	0	0	0	5	0	0	0	**35**	0	0	0	0	0	0.88
	pancreas	1	0	1	0	0	1	0	12	0	**2**	0	0	0	1	0.11
	prostate gland	1	0	2	0	1	0	0	0	0	0	**50**	0	0	3	0.88
	skin	1	1	1	0	1	1	2	1	0	1	0	**41**	0	0	0.82
	stomach	0	0	0	5	0	1	0	1	2	0	0	0	**21**	0	0.70
	thyroid gland	1	0	0	0	0	2	0	2	0	0	0	0	0	**66**	0.93
	precision	0.52	0.97	0.78	0.74	0.70	0.86	0.92	0.71	0.85	0.50	0.98	1.00	0.91	0.93	**0.82**

In the confusion matrix of protein coding genes, it could be seen that brain cancer, liver cancer, prostate gland cancer and thyroid gland cancer scored the highest precision and recall values (Table 4.3). All prostate gland cancers could be identified with only 2% being false positives. Pancreatic cancer scored higher compared to lncRNA and TEC expression profile, 84% of it could be identified with 0.94 precision. The lowest recall and precision rate was obtained for bladder cancer.

Table 4.3: Confusion matrix for protein coding genes across 14 cancer types with recall and precision values.

		PREDICTION														
		bladder	**brain**	**breast**	**colon**	**corpus uteri**	**kidney**	**liver**	**lung**	**ovary**	**pancreas**	**prostate gland**	**skin**	**stomach**	**thyroid gland**	**recall**
ACTUAL	**bladder**	**16**	1	10	3	0	3	0	8	0	0	0	1	2	2	0.35
	brain	0	**65**	0	0	1	2	0	0	0	0	0	0	0	0	0.96
	breast	2	1	**107**	3	3	7	0	5	0	0	0	2	0	0	0.82
	colon	1	0	2	**37**	3	3	0	1	1	0	0	1	0	0	0.76
	corpus uteri	1	0	3	4	**44**	3	0	1	3	0	0	0	0	0	0.75
	kidney	1	2	1	0	2	**84**	0	2	1	1	1	5	1	0	0.83
	liver	0	1	1	0	0	1	**40**	0	0	0	0	0	0	0	0.93
	lung	3	0	5	1	1	0	0	**103**	2	0	0	3	0	0	0.87
	ovary	1	0	2	1	2	1	0	3	**30**	0	0	0	0	0	0.75
	pancreas	0	1	1	0	0	1	0	0	0	**15**	0	0	0	0	0.83
	prostate gland	0	0	0	0	0	0	0	0	0	0	**57**	0	0	0	1.00
	skin	2	0	5	1	0	2	0	5	0	0	0	**35**	0	0	0.70
	stomach	0	0	1	1	0	2	0	2	0	0	0	0	**24**	0	0.80
	thyroid gland	0	0	0	0	0	2	0	1	0	0	0	1	1	**66**	0.93
	precision	0.59	0.92	0.78	0.73	0.79	0.76	1.00	0.79	0.81	0.94	0.98	0.73	0.86	0.97	**0.82**

Cancer types such as brain, prostate gland and thyroid gland cancer has higher F1 scores (>0.9) in classification based on the top 3 gene types (Table 4.4). 7 out of all cancer types had lower than 0.9 F1 scores for all the gene types. Among them, bladder, corpus uteri, lung, stomach and pancreatic cancer are the main ones to touch upon. The obtained results so far necessitate utilization of all 7 gene types for classification and lower scoring cancer types will be tracked in further parts of the procedure.

Table 4.4: F1 scores for obtained for top 7 genes that scored the highest when the used for the single gene type-based classification across 14 cancer types.

	lncRNA	TEC	protein coding	miRNA	trans. unproc. pseu.	trans. proc. pseu.	trans. unit. pseu.
bladder	0.78	0.38	0.44	0.63	0.42	0.70	0.68
brain	0.98	0.97	0.94	0.99	0.88	0.96	0.96
breast	0.86	0.78	0.80	0.66	0.73	0.84	0.80
colon	0.89	0.83	0.74	0.80	0.71	0.61	0.83
corpus uteri	0.80	0.74	0.77	0.45	0.71	0.67	0.71
kidney	0.93	0.91	0.79	0.82	0.89	0.82	0.79
liver	0.93	0.86	0.96	0.63	0.67	0.91	0.72
lung	0.68	0.79	0.83	0.58	0.58	0.81	0.73
ovary	0.84	0.86	0.78	0.91	0.87	0.59	0.79
pancreas	0.20	0.18	0.88	0.20	0.00	0.56	0.53
prostate gland	0.97	0.93	0.99	0.57	0.94	0.94	0.86
skin	0.95	0.90	0.71	0.79	0.67	0.87	0.90
stomach	0.81	0.79	0.83	0.78	0.88	0.21	0.71
thyroid gland	0.91	0.93	0.95	0.83	0.85	0.76	0.82

Before loading the features into the models for training, the correlation between the top scoring 7 gene types was investigated. It was found that there is not any significant correlation between gene types (Table 4.5). This enables the procedure forward into the modelling step without concerns. If there was a high correlation between any gene types, one of each correlated pair needed to be discarded.

Table 4.5: Correlation matrix between gene types.

	lncRNA	TEC	protein coding	transcribed unitary pseudogene	transcribed unprocessed pseudogene	transcribed processed pseudogene	miRNA
lncRNA	1.000	0.394	-0.022	0.000	0.067	-0.081	0.025
TEC	0.394	1.000	0.008	0.003	0.011	-0.075	0.043
protein coding	-0.022	0.008	1.000	-0.011	-0.103	-0.187	-0.008
transcribed unitary pseudogene	0.000	0.003	-0.011	1.000	0.000	-0.013	0.000
transcribed unprocessed pseudogene	0.067	0.011	-0.103	0.000	1.000	0.022	-0.006
transcribed processed pseudogene	-0.081	-0.075	-0.187	-0.013	0.022	1.000	0.013
miRNA	0.025	0.043	-0.008	0.000	-0.006	0.013	1.000

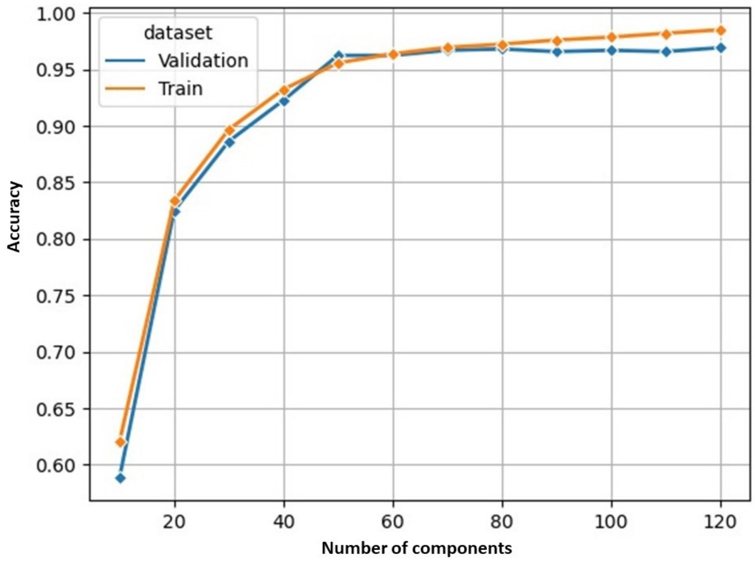


Figure 4.2: Accuracy versus number of components obtained by PCA. Accuracy reaches plateau around 60 components

4.2 Logistic Regression results

In the next step, principal component analysis was performed for reducing the dimensionality of the data from 59444 transcripts down to maximum of 120 features. Based on the accuracy versus number of components analysis, it was chosen that 60 features yielded the best result based on three considerations: 1. accuracy score was not getting significantly higher, 2. training and validation scores were moving in the opposite direction, 3. higher number of features demands higher computational resources (Figure 4.2).

Table 4.6: Logistic regression results with precision and recall values. Accuracy score is given at the bottom right corner of the table (0.96).

		PREDICTION														
		bladder	**brain**	**breast**	**colon**	**corpus uteri**	**kidney**	**liver**	**lung**	**ovary**	**pancreas**	**prostate gland**	**skin**	**stomach**	**thyroid gland**	**Recall**
ACTUAL	**bladder**	33	0	3	0	0	0	0	3	0	1	0	0	1	0	0.80
	brain	0	77	0	0	0	0	0	0	0	0	0	0	0	0	1.00
	breast	2	1	125	0	0	1	0	3	1	0	0	0	0	0	0.94
	colon	0	0	0	42	0	0	0	0	0	1	0	0	0	0	0.98
	corpus uteri	1	0	2	0	54	0	0	0	1	0	0	0	0	0	0.93
	kidney	0	0	0	0	0	106	0	0	0	0	0	0	0	0	1.00
	liver	0	0	0	0	0	1	50	0	0	0	0	0	0	0	0.98
	lung	1	0	3	0	0	0	0	95	0	1	0	0	1	0	0.94
	ovary	0	0	0	0	0	0	0	0	44	0	0	0	0	0	1.00
	pancreas	0	0	0	0	0	0	0	1	1	19	0	1	0	0	0.86
	prostate gland	0	0	0	0	0	0	0	0	0	0	58	0	0	0	1.00
	skin	0	0	0	0	0	0	0	0	0	0	0	38	0	0	1.00
	stomach	1	0	0	0	0	0	0	1	0	0	0	0	47	0	0.96
	thyroid gland	0	0	0	0	0	0	1	1	0	0	0	0	0	57	0.97
	Precision	0.87	0.99	0.94	1.00	1.00	0.98	0.98	0.91	0.94	0.86	1.00	0.97	0.96	1.00	0.96

Classification by logistic regression based on the features obtained from PCA obtained 96% accuracy. Bladder, breast and lung cancer has the highest number of false positives (Table 4.6). When the F1 scores for different cancer types were compared with those obtained by single gene type-based classification, it was observed a significant increase (Table 4.7). Bladder cancer scored 0.85 in validation step and 0.84 in testing step. Despite this improvement, it was the lowest scoring cancer type. The recall rate was 0.80, which means 20% of the time it could not be detected despite being present.

Table 4.7: Comparison of overall accuracy and F1 scores of cancer types obtained for single gene type-based classification and all genes-based classification by logistic regression.

	lncRNA	**TEC**	**protein coding**	**miRNA**	**trans. unproc. pseu.**	**trans. proc. pseu.**	**trans. unit. pseu.**	**all genes validation**	**all genes test**
bladder	0.78	0.38	0.44	0.63	0.42	0.70	0.68	0.85	0.84
brain	0.98	0.97	0.94	0.99	0.88	0.96	0.96	0.99	0.99
breast	0.86	0.78	0.80	0.66	0.73	0.84	0.80	0.96	0.94
colon	0.89	0.83	0.74	0.80	0.71	0.61	0.83	0.98	0.99
corpus uteri	0.80	0.74	0.77	0.45	0.71	0.67	0.71	0.96	0.96
kidney	0.93	0.91	0.79	0.82	0.89	0.82	0.79	0.96	0.99
liver	0.93	0.86	0.96	0.63	0.67	0.91	0.72	0.95	0.98
lung	0.68	0.79	0.83	0.58	0.58	0.81	0.73	0.94	0.93
ovary	0.84	0.86	0.78	0.91	0.87	0.59	0.79	0.99	0.97
pancreas	0.20	0.18	0.88	0.20	0.00	0.56	0.53	0.97	0.86
prostate gland	0.97	0.93	0.99	0.57	0.94	0.94	0.86	1.00	1.00
skin	0.95	0.90	0.71	0.79	0.67	0.87	0.90	0.96	0.99
stomach	0.81	0.79	0.83	0.78	0.88	0.21	0.71	0.98	0.96
thyroid gland	0.91	0.93	0.95	0.83	0.85	0.76	0.82	0.98	0.98
overall accuracy	0.85	0.82	0.82	0.71	0.74	0.78	0.79	0.97	0.96

Pancreatic cancer had very low F1 scores across all the gene types, except protein coding genes, and it scored 0.97 in validation step and 0.86 in testing step. Other cancer types, such as corpus uteri, colon and stomach cancer, despite not having over 0.90 F1 score previously, scored over 0.95 in the test results. Lung and breast cancer stayed just under 0.95. Prostate cancer proved to be the most easily identifiable type of cancer by means of its gene expression profile, followed by brain, colon, kidney and skin cancer.

4.3 Comparison of Logistic Regression results with SVM and Random Forest classifier results

After obtaining logistic regression results, two more models were applied on the data for comparison. These models include a support vector machine and a random forest classifier. Support vector model yielded the same accuracy as the logistic regression model; however, random forest classifier yielded a lower accuracy (Table 4.8).

Table 4.8: Comparison of F1 scores of cancer types and overall accuracy for logistic regression, support vector machine and random forest classifier.

	Logistic regression	**SVM**	**Random Forest**
bladder	0.84	0.86	0.81
brain	0.99	0.99	0.99
breast	0.94	0.94	0.91
colon	0.99	0.99	0.94
corpus uteri	0.96	0.98	0.94
kidney	0.99	0.99	0.94
liver	0.98	0.98	0.97
lung	0.93	0.93	0.86
ovary	0.97	0.98	0.98
pancreas	0.86	0.78	0.83
prostate gland	1.00	1.00	0.99
skin	0.99	0.96	0.95
stomach	0.96	0.98	0.85
thyroid gland	0.98	0.97	0.98
overall accuracy	0.96	0.96	0.93

Logistic regression classifier outperformed SVM in brain, skin, thyroid and pancreatic cancer prediction. SVM performed better at predicting bladder, corpus uteri, ovary and stomach cancer. The difference in the accuracy scores was less than 0.01 in the remaining cancers. Lowest accuracy score for logistic regression classifier was in bladder cancer, and for SVM, it was pancreatic cancer. Both logistic regression and SVM models bypassed 0.95 accuracy threshold for 10 out of 14 cancer types, however random forest could only achieve this result for 5 out of 14 cancer types (Figure 4.2).

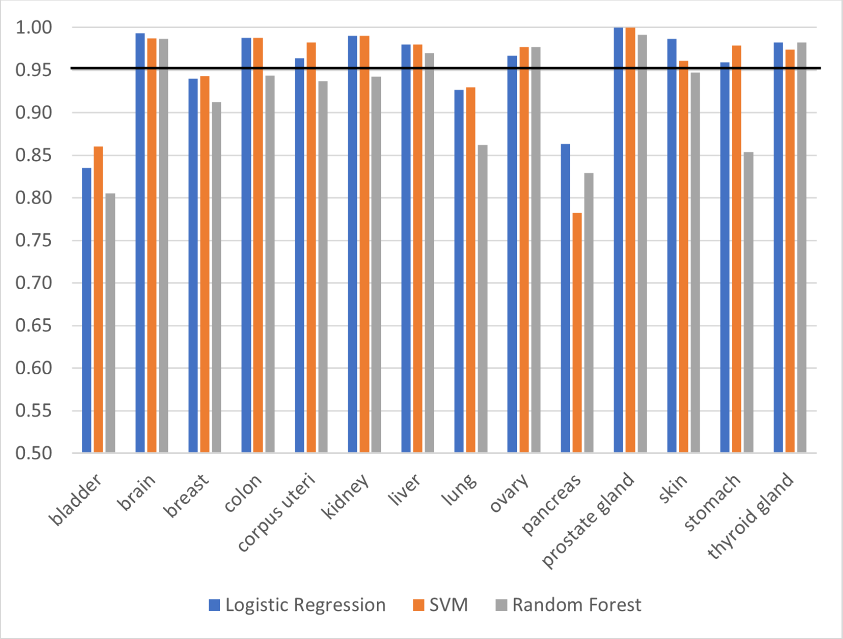


Figure 4.3: Bar plot of F1 scores for logistic regression, SVM and random forest model across 14 cancer types. Black line represents 0.95 threshold as a reference.

CHAPTER 5

DISCUSSION

5.1 LncRNA and miRNA expression in cancers

It was demonstrated in the results that long non-coding RNA (lncRNA) expression profile is unique across many cancer types. Despite most of the human genome is transcribed, only a small portion of the transcriptome can encode for proteins. LncRNAs are among the transcripts that cannot code for any proteins. They regulate transcription level in the cell through different means. This includes modifying chromatin state to activate or repress genes and interacting with ribonucleoprotein complexes and transcription factors [92]. Due to their importance in DNA methylation profiles, the obtained results compels us to look into methylation profile differences in cancer types and how they are conserved in the metastases of the primary tumours.

In the obtained results, it was shown that brain cancer demonstrated an easily identifiable expression profile for lncRNA genes (Table 4.1). This result is well supported by the works in literature. Some lncRNAs that have oncogenic induction in gliomas include HOTAIR, lncRNA-ATB, CRNDE, ECONEXIN. Another group of lncRNAs are more context dependent and can demonstrate an elevated or lowered expression level, such as, HOTTIP and MALAT1 [93].

Most of the lncRNAs regarding prostate cancer are tumor-promoting, and only a few of them are tumor-suppressing [94]. Especially, PCA3 is the main biomarker lncRNA that is upregulated in this cancer type. This expression profile allows easy identification of this malignancy type, as shown in Table 4.1.

One of the noticeable points in the confusion matrix of lncRNA based classification is that many of bladder, pancreas and thyroid gland cancers were classified as lung cancer. This could be either due to the non-unique expression profile of specific cancer type for lncRNA, or an actual relationship. Lungs are the most common metastasis site for breast cancer, thyroid cancer. Additionally, in a study by Shinagare et al. (2011), 37% of patients with bladder cancer had lungs as a metastasis site [95]. TCGA program mainly focuses on data collection from primary tumours. However, if we do not rule out the existence of data from secondary tumours, this could mean that bladder cancer may change their lncRNA expression profile to that of the site they metastasize to. This statistics could also exist due to similarities between some subtypes of these two cancers types.

In Table 4.4, miRNA profile can be distinguished in ovary cancer more easily. Ovarian cancer is well-known for its miRNA overexpression rooted causes. MiRNAs can be used as biomarkers for ovarian cancer prognosis [96]. On the contrary, prostate cancer had a very low F1 score for miRNA expression-based classification.

5.2 TEC and pseudogene expression in cancers

TEC genes code for non-receptor tyrosine kinases that are located in the cell cytoplasm and mediate intracellular signaling. Overexpression of this class of tyrosine kinases has been linked to hematological malignancies. Byrone's tyrosine kinase (BTK) is a common target for drugs such as Ibrutinib [97]. Based on Figure 4.1, it can be said that TEC kinases are not only related to hematological cancers, since it has a recognizable expression pattern in the cancer types classified in this work. Yue et al., (2017) identified an overexpression of BTK in gliomas [98]. It is plausible that BTK expression profile has played a significant role in identification of brain tumors based on TEC genes in this work.

Expression profile of all three types of pseudogenes were among the top discriminators of cancer types. Pseudogenes could be processed, unprocessed (also known as duplicated) or unitary. Processed pseudogenes occur via retrotransposons, a genetic component that copies itself into the functional protein coding gene followed by accumulation of the mutations that disable the gene. Unprocessed pseudogenes are formed via gene duplication followed by disabling mutations. Unitary pseudogenes are formed by accumulation of disabling mutations. Pseudogenes are highly linked to various cancer types. They interfere with other transcripts either by hybridizing with them or by competing for the same RNA-binding protein complexes [99].

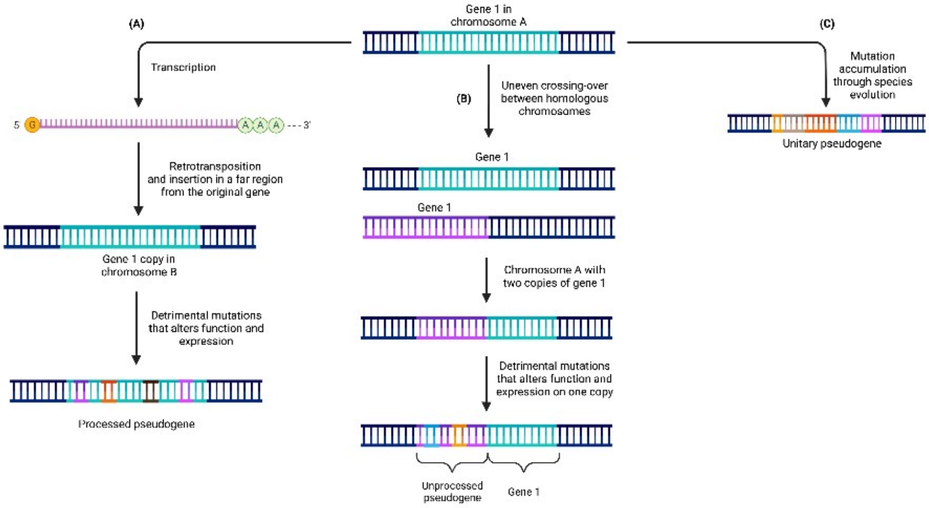

Gene 1 in chromosome A
(A)
Transcription
5'
G
A A A
3'
Retrotransposition and insertion in a far region from the original gene
Gene 1 copy in chromosome B
Detrimental mutations that alters function and expression
Processed pseudogene
(B)
Uneven crossing-over between homologous chromosomes
Gene 1
Gene 1
Chromosome A with two copies of gene 1
Detrimental mutations that alters function and expression on one copy
Unprocessed pseudogene
Gene 1
(C)
Mutation accumulation through species evolution
Unitary pseudogene

Figure 5.1: Pseudogenes are classified into three groups based on their formation mechanism. A) Retrotransposition of an mRNA molecule of a distant gene, B) Gene duplication, C) Accumulation of mutations in the original gene. [99]

5.3 Genomic profile of pancreatic cancer

Pancreatic cancer has one of the highest mortality rates of all malignant cases [100]. It has a considerable potential of metastasis because of its location. According to U.S. Cancer Statistics Working Group (2023), 48.8% of all pancreatic cancer cases were diagnosed at distant stage, which means the malignancy had already metastasized to other regions into body at the time of diagnosis [101]. Therefore, pancreatic cancer cases retrieved from TCGA must have been mostly at a distant stage. In this work, this cancer type proved to be one of the most difficult to identify using a single type of genes. This could be due to the smaller number of cases, i.e., class imbalance, or it might indicate a change in gene expression pattern throughout metastasis process. However, it scored significantly higher (0.88) F1 score when expression of protein coding genes was used. In other works, it was found that the best three single gene discriminators in pancreatic adenocarcinomas are KRT17, COL10A1 and CTHRC1 genes [102]. All these genes are protein-coding genes, as KRT17 codes for keratin chain, and the ladder participates in collagen synthesis [103].

F1 score differs greatly between validation and testing. It went from 0.97 in the validation to 0.86 in the testing, and even lower in the other models. This strongly implies the effects of class imbalance in this cancer type. We need to increase the amount of patient data to get a more reliable result, however, TCGA program is especially scarce on pancreatic cancer data. Increasing the coverage of this work to include ICGC program, may improve the reliability of the results. Very low survival rate in pancreatic cancer, lowers the importance of its identification in CUPs. It has hazard ratio of 1.71 compared to CUPs. In other words, the problem lies in detection of the malignancy as a cancer, rather than being detected with an unknown the primary.

5.4 Analysis of confusion matrix obtained by logistic regression

In the confusion matrix of the logistic regression results, highest false positive rates belong to pancreatic, lung and breast cancer. We should exclude pancreatic cancer from this list because of very little number of cases in this cancer type, in other of words, class imbalance. There exists an interesting pattern among the false positive rates, which can be expressed in two items: 1) Many breast and lung cancer cases are misclassified as each other 2) Many bladder cancer cases are misclassified as breast and lung cancers. Frequent metastasis of breast cancer to lungs is a well-known fact. It is reported that more than 60% of the metastatic breast cancer causes secondary tumours at lungs or bones [104]. Moreover, in a study carried out in Taiwan, it was found that patients with breast cancer had higher risk of developing second primary lung cancers [105]. Hence, we can make two conclusions based on the above mentioned information, either breast cancer metastasized to lungs gains gene expression similarity to that of the metastasis site, or some cases from lung and breast cancer are annotated incorrectly.

If we refer to previous works to decipher the similarity between breast and bladder cancer, it can be found that basal-like subtype of both malignancies share strong similarities by means of the genes involved [106]. When it comes to the similarity between bladder and lung cancer, it is reported that small cell carcinomas of bladder are lung share a convergent pathogenesis [107]. However, it must be noted that small cell carcinomas of bladder are very rare and account for less than 1% of bladder cancers. Another possible cause for this similarity may be due to subtypes of these two cancer types. Squamous cell carcinoma (SCC) is a malignancy which can develop in organs that are covered with squamous epithelium. Lungs and bladder are among such organs [108]. It is possible that some cases in our data belong to groups SCCs of lung and bladder that share a very similar genomic profile.

5.5 Comparative analysis of the final results

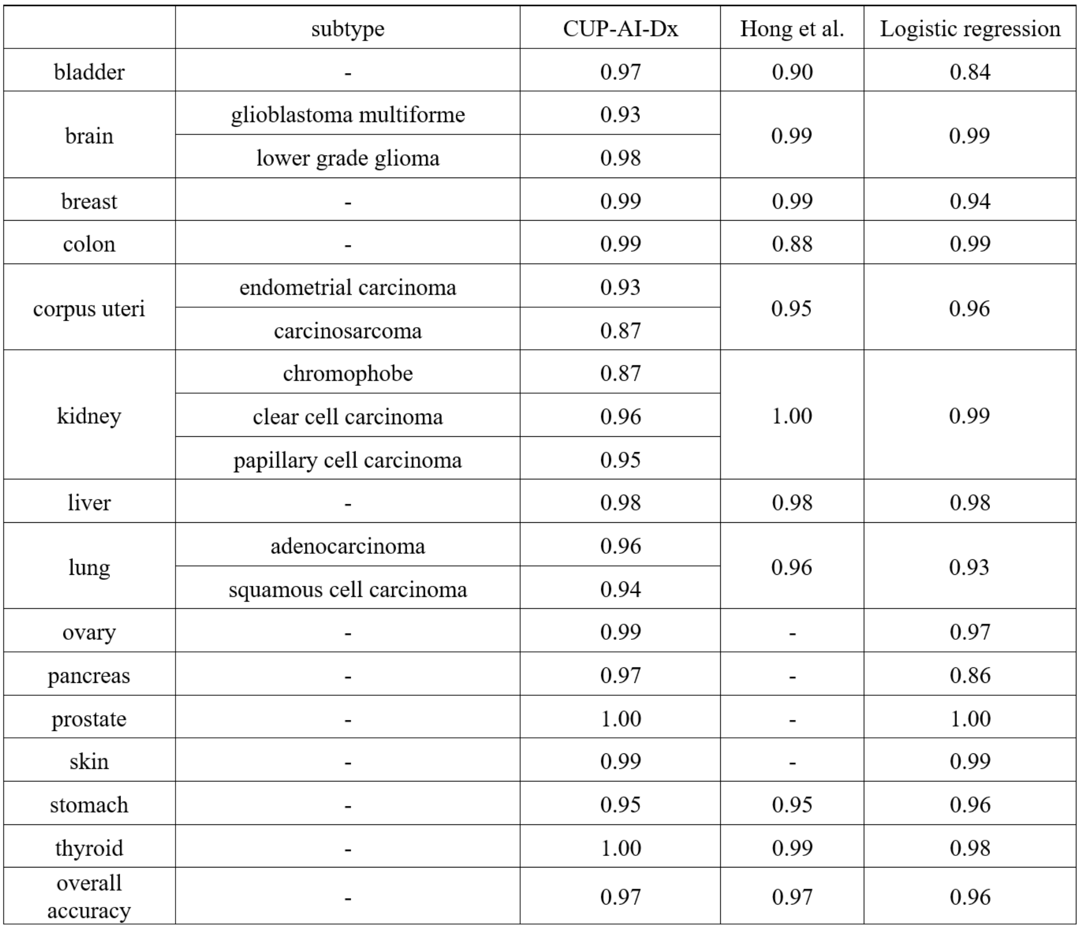
When combination of all biomarkers used, classification accuracy was improved. Only 60 components derived by PCA were used instead of around 41000 different transcripts. This was done to alleviate the stress on computational and time resources. Linear models which include logistic regression and SVM classifier achieved a higher accuracy than the random forest classifier, which is a non-linear model. Support vector classifier yielded the same accuracy as logistic regression, with similar F1 scores per cancer type. SVM predominates logistic regression in bladder, corpus uteri, ovary and stomach cancer, whereas, logistic regression performs better in identifying pancreatic, skin and thyroid cancer. Accuracy scores for all types of cancers are very similar for these two models, except for pancreatic cancer, which may be because of low abundance of the cases.

Random Forest classifiers has been regarded as among the top classification algorithms in biology and medicine [109]. Despite this fact, it scored the lowest accuracy in this work compared to the other two models. A possible explanation could be that the number of components to be used in the modeling stage was chosen based on the performance by logistic regression. This was because logistic regression was the primary choice for the classification task, due to its higher computation speed than decision tree algorithms. This also explains the accuracy similarity between logistic regression and SVM, since they are both linear models.

However, when we varied the number of components in random forest classifier, it also performed the best at 60 components like the other two models. Another reason could be related to the loss function utilized in random forest algorithm. Despite all hyperparameters being exhausted, only a maximum of 92% accuracy could be obtained. In a study related to diabetes diagnosis by ML, logistic regression was reported to score a better accuracy than random forest [110]. In another comparative ML study, it was similarly found that logistic regression performed superior [111]. These findings do not rule out the possibility of logistic regression outperforming random forest classifier in a classification-based task.

Logistic regression is referred to as the main model in this study. For comparison, two other works from literature are chosen. One of them involves a 1D inception convolutional neural network named CUP-AI-Dx [5], and the other one is 5-layered convolutional neural network [112]. CUP-AI-Dx includes subtypes of the cancer and has been tested on clinical data, with promising results. When compared to these two models, our model seems to have achieved a competitive classification accuracy (Table 10 5.1). Because of the inclusion of cancer subtypes it is not possible to make a direct comparison of F1 scores between CUP-AI-Dx and logistic regression in 4 cancer types. However, in 9 out of other 10 cancer types, close F1 scores are observed. In case of bladder cancer our model has performed inferior.

Table 5.1: Comparison of logistic regression F1 scores achieved across 14 cancer types in this work with the F1 scores by CUP-AI-Dx model and Hong team. Cancer subtypes were classified separately in CUP-AI-Dx model, hence there are separate F1 scores that correspond to a single F1 score in the other columns [5]. Some of the cancer types used in this work were absent in the study by Hong team [112].



	subtype	CUP-AI-Dx	Hong et al.	Logistic regression
bladder	-	0.97	0.90	0.84
brain	glioblastoma multiforme	0.93	0.99	0.99
	lower grade glioma	0.98		
breast	-	0.99	0.99	0.94
colon	-	0.99	0.88	0.99
corpus uteri	endometrial carcinoma	0.93	0.95	0.96
	carcinosarcoma	0.87		
kidney	chromophobe	0.87	1.00	0.99
	clear cell carcinoma	0.96		
	papillary cell carcinoma	0.95		
liver	-	0.98	0.98	0.98
lung	adenocarcinoma	0.96	0.96	0.93
	squamous cell carcinoma	0.94		
ovary	-	0.99	-	0.97
pancreas	-	0.97	-	0.86
prostate	-	1.00	-	1.00
skin	-	0.99	-	0.99
stomach	-	0.95	0.95	0.96
thyroid	-	1.00	0.99	0.98
overall accuracy	-	0.97	0.97	0.96

When we look at the results by the Hong team, some cancer types used in our work is not covered. Among the comparable F1 scores, there are 2 main differences. Like the previous comparison, the F1 score for bladder cancer in our work is lower. However, it must be noted that this time the difference is slighter, which signifies the challenge in identification of bladder cancer. In case of colon cancer, our model has performed significantly better than the neural network model by the Hong team.

5.6 Limitations

In this study, we aimed to utilize the similarity between primary tumors and their metastases by means gene expression profiles. Because of this assumption, cancers with already known types were used for training the models on. Since the data was retrieved from the TCGA program, it was significantly clean. However, it must be mentioned that the obtained results may not represents itself in clinical applications to several factors. Firstly, despite the similarity, there also exists heterogeneity between some tumors. This means the same tumor type in the organism may not represent the genomic profile similarity. Despite scarce in numbers, some secondary tumors have been reported to accumulate a different line of mutations compared to the corresponding primary tumor [113]. Secondly, it could be challenging to get a clean expression data depending on how difficult it is to obtain the tumor sample to run RT-qPCR on. This can induce high levels of noise which, in turn, can lower the accuracy of the built models in predicting the primary tissue. For example, when applied on data retrieved from two different hospitals, one ML model achieved only 86.96% and 72.46% accuracy, compared to the accuracy achieved on clean TCGA data, which was 96.70% [5].

Another limitation in this work was related to lack of enough computational resources to address a more variety of algorithms and data types. Models were built on a limited number of components transformed by dimensionality reduction, because of huge numbers of genes. Comparative analysis could be investigated between two scenarios, where dimensionality reduction is applied or not, respectively.

Although machine learning can revolutionize the identification of primary tissues in malignancies of unknown primary (CUP), existing methodologies have several drawbacks. The scarcity of annotated datasets of superior quality, for example, pancreatic cancer in this work, represents a principal obstacle. An inadequate degree of generalizability characterizes numerous machine learning models. Although they exhibit satisfactory performance on the training data, their ability to generate precise predictions diminishes when confronted with data from diverse populations or medical centers. Variations in patient demographics, data collection methodologies, and disease presentations partially contribute to this constraint [114].

The perception of numerous sophisticated machine learning models, particularly deep learning models, as "black boxes" is common. The absence of transparency poses a substantial obstacle in clinical environments, where comprehension of the reasoning behind a model's prognosis is vital for establishing confidence and making informed decisions [115].

5.7 Future Directions

One of the major issues of this study was regarding the variation in the data by means of demography and data collection centers. In other words, data from only TCGA program is utilized. Although being advantageous because of its superior quality, data scope could be extended to other programs such as ICGC, besides TCGA. This could not only increase the size of training and test sample, but also introduce a realistic variation. Increase in the size and variation of the data may also compel us to upgrade our modeling approach to using neural networks. Consequently, built models ability to allow for more possibilities and infer the most essential patterns in the data would increase. In the further steps, the built models should be applied to clinical data to check the performance variation.

TCGA program collects data mainly from primary cancers. Despite the reported genetic similarities between primary and secondary cancers, including the differences while building the models can yield more reliable outcomes. Hence, incorporation of metastatic expression data to training and test samples can prove significant. Next, if a similar pattern exists for deferentially expressed genes, we could try to minimize this effect to build stronger relationship between primary and secondary tumours. However, it must be noted that, unfortunately, there are no well-known active databases that collect metastatic transcriptomics data.

Gene expression is not the only genetic element that is found to be conserved between primary and secondary tumours. Several studies have shown the importance of DNA methylation profiles in identification of primary sites in diagnostics [116, 117, 118]. In this work, only two out of seven top discriminator gene types were coding for proteins. Especially, lncRNAs, our top scoring discriminator, are well known for their gravity in transcriptional regulation by modifying chromatin state [92]. Following this motive, methylation patterns of specific cancers can be further investigated and incorporated into our models for a better decision making process. Another type of data that could prove useful for integrating into our model, is mutational profiles specific to cancers. However, integration of all the above steps may also aggravate the demand on complexity of the data to be obtained from the patients, which might prolong the diagnosis period.

REFERENCES

[1] K. Hemminki, M. Bevier, A. Hemminki, and J. Sundquist, "Survival in cancer of unknown primary site: population-based analysis by site and histology," *Annals of oncology*, vol. 23, no. 7, pp. 1854–1863, 2012.

[2] F. A. Monzon, F. Medeiros, M. Lyons-Weiler, and W. D. Henner, "Identification of tissue of origin in carcinoma of unknown primary with a microarray-based gene expression test," *Diagnostic pathology*, vol. 5, no. 1, pp. 1–9, 2010.

[3] B. Weigelt, A. M. Glas, L. F. Wessels, A. T. Witteveen, J. L. Peterse, and L. J. van't Veer, "Gene expression profiles of primary breast tumors maintained in distant metastases," *Proceedings of the National Academy of Sciences*, vol. 100, no. 26, pp. 15901–15905, 2003.

[4] S. Kim, D. Shin, A. Min, M. Kim, D. Na, H.-B. Lee, H. S. Ryu, Y. Yang, G.-U. Woo, K.-H. Lee, *et al.*, "Genomic profile of metastatic breast cancer patient-derived xenografts established using percutaneous biopsy," *Journal of Translational Medicine*, vol. 19, no. 1, pp. 1–17, 2021.

[5] Y. Zhao, Z. Pan, S. Namburi, A. Pattison, A. Posner, S. Balachander, C. A. Paisie, H. V. Reddi, J. Rueter, A. J. Gill, *et al.*, "Cup-ai-dx: A tool for inferring cancer tissue of origin and molecular subtype using rna gene-expression data and artificial intelligence," *EBioMedicine*, vol. 61, 2020.

[6] K. A. Preethi, G. Lakshmanan, and D. Sekar, "Antagomir technology in the treatment of different types of cancer," *Epigenomics*, vol. 13, no. 07, pp. 481–484, 2021.

[7] D. L. Starmer, "Medical student exposure to cancer patients whilst on clinical placement: A retrospective analyses of clinical log books," *Journal of Cancer Education*, vol. 34, pp. 671–676, 2019.

[8] M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, and G. Getz, "Discovery and saturation analysis of cancer genes across 21 tumour types," *Nature*, vol. 505, no. 7484, pp. 495–501, 2014.

[9] S. V. Vasaikar, P. Straub, J. Wang, and B. Zhang, "Linkedomics: analyzing multi-omics data within and across 32 cancer types," *Nucleic acids research*, vol. 46, no. D1, pp. D956–D963, 2018.

[10] D. E. Brash, "How do mutant clones expand in normal tissue?," *Frontiers in cancer research: evolutionary foundations, revolutionary directions*, pp. 61–98, 2016.

[11] C. Compton and C. Compton, *Cancer initiation, promotion, and progression and the acquisition of key behavioral traits*. Springer, 2020.

[12] M. Arnold, C. C. Abnet, R. E. Neale, J. Vignat, E. L. Giovannucci, K. A. McGlynn, and F. Bray, "Global burden of 5 major types of gastrointestinal cancer," *Gastroenterology*, vol. 159, no. 1, pp. 335–349, 2020.

[13] K. Yizhak, F. Aguet, J. Kim, J. M. Hess, K. Kübler, J. Grimsby, R. Frazer, H. Zhang, N. J. Haradhvala, D. Rosebrock, *et al.*, "Rna sequence analysis reveals macroscopic somatic clonal expansion across normal tissues," *Science*, vol. 364, no. 6444, p. eaaw0726, 2019.

[14] D. Tamborero, A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, C. Kandoth, J. Reimand, M. S. Lawrence, G. Getz, G. D. Bader, L. Ding, *et al.*, "Comprehensive identification of mutational cancer driver genes across 12 tumor types," *Scientific reports*, vol. 3, no. 1, p. 2650, 2013.

[15] I. A. Prior, F. E. Hood, and J. L. Hartley, "The frequency of ras mutations in cancer," *Cancer research*, vol. 80, no. 14, pp. 2969–2974, 2020.

[16] M. J. Williams, B. Werner, C. P. Barnes, T. A. Graham, and A. Sottoriva, "Identification of neutral tumor evolution across cancer types," *Nature genetics*, vol. 48, no. 3, pp. 238–244, 2016.

[17] C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, *et al.*, "Mutational landscape and significance across 12 major cancer types," *Nature*, vol. 502, no. 7471, pp. 333–339, 2013.

[18] I. R. Watson, K. Takahashi, P. A. Futreal, and L. Chin, "Emerging patterns of somatic mutations in cancer," *Nature reviews Genetics*, vol. 14, no. 10, pp. 703–718, 2013.

[19] I. Sæterdal, J. Bjørheim, K. Lislerud, M. K. Gjertsen, I. K. Bukholm, O. C. Olsen, J. M. Nesland, J. A. Eriksen, M. Møller, A. Lindblom, *et al.*, "Frameshift-mutation-derived peptides as tumor-specific antigens in inherited and spontaneous colorectal cancer," *Proceedings of the National Academy of Sciences*, vol. 98, no. 23, pp. 13255–13260, 2001.

[20] J. R. Pon and M. A. Marra, "Driver and passenger mutations in cancer," *Annual Review of Pathology: Mechanisms of Disease*, vol. 10, pp. 25–50, 2015.

[21] B. J. Aubrey, A. Strasser, and G. L. Kelly, "Tumor-suppressor functions of the tp53 pathway," *Cold Spring Harbor perspectives in medicine*, vol. 6, no. 5, 2016.

[22] F. Supek, B. Miñana, J. Valcárcel, T. Gabaldón, and B. Lehner, "Synonymous mutations frequently act as driver mutations in human cancers," *Cell*, vol. 156, no. 6, pp. 1324–1335, 2014.

[23] S. H. Hassanpour and M. Dehghani, "Review of cancer from perspective of molecular," *Journal of cancer research and practice*, vol. 4, no. 4, pp. 127–129, 2017.

[24] M. Uhlen, C. Zhang, S. Lee, E. Sjöstedt, L. Fagerberg, G. Bidkhori, R. Benfeitas, M. Arif, Z. Liu, F. Edfors, *et al.*, "A pathology atlas of the human cancer transcriptome," *Science*, vol. 357, no. 6352, p. eaan2507, 2017.

[25] U. Varol, Y. Kucukzeybek, A. Alacacioglu, I. Somali, Z. Altun, S. Aktas, and M. O. Tarhan, "Brca genes: brca 1 and brca 2," *apoptosis*, vol. 13, p. 19, 2018.

[26] Z. Kleibl and V. N. Kristensen, "Women at high risk of breast cancer: Molecular characteristics, clinical presentation and management," *The Breast*, vol. 28, pp. 136–144, 2016.

[27] D. Provenzale, S. Gupta, D. J. Ahnen, T. Bray, J. A. Cannon, G. Cooper, D. S. David, D. S. Early, D. Erwin, J. M. Ford, *et al.*, "Genetic/familial high-risk assessment: colorectal version 1.2016, nccn clinical practice guidelines in oncology," *Journal of the National Comprehensive Cancer Network*, vol. 14, no. 8, pp. 1010–1030, 2016.

[28] M. B. Yurgelun, B. Allen, R. R. Kaldate, K. R. Bowles, T. Judkins, P. Kaushik, B. B. Roa, R. J. Wenstrup, A.-R. Hartman, and S. Syngal, "Identification of a variety of mutations in cancer predisposition genes in patients with suspected lynch syndrome," *Gastroenterology*, vol. 149, no. 3, pp. 604–613, 2015.

[29] M. B. Daly, R. Pilarski, J. E. Axilbund, S. S. Buys, B. Crawford, S. Friedman, J. E. Garber, C. Horton, V. Kaklamani, C. Klein, *et al.*, "Genetic/familial high-risk assessment: breast and ovarian, version 1.2014," *Journal of the National Comprehensive Cancer Network*, vol. 12, no. 9, pp. 1326–1338, 2014.

[30] H. R. Aslanian, J. H. Lee, and M. I. Canto, "Aga clinical practice update on pancreas cancer screening in high-risk individuals: expert review," *Gastroenterology*, vol. 159, no. 1, pp. 358–362, 2020.

[31] E. Budinska, V. Popovici, S. Tejpar, G. D'Ario, N. Lapique, K. O. Sikora, A. F. Di Narzo, P. Yan, J. G. Hodgson, S. Weinrich, *et al.*, "Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer," *The Journal of pathology*, vol. 231, no. 1, pp. 63–76, 2013.

[32] Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, and R. Shen, "Pattern discovery and cancer gene identification in integrated cancer genomic data," *Proceedings of the National Academy of Sciences*, vol. 110, no. 11, pp. 4245–4250, 2013.

[33] Z. Tang, C. Li, B. Kang, G. Gao, C. Li, and Z. Zhang, "Gepia: a web server for cancer and normal gene expression profiling and interactive analyses," *Nucleic acids research*, vol. 45, no. W1, pp. W98–W102, 2017.

[34] H. R. Ali, L. Chlon, P. D. Pharoah, F. Markowetz, and C. Caldas, "Patterns of immune infiltration in breast cancer and their clinical implications:

a gene-expression-based retrospective study," *PLoS medicine*, vol. 13, no. 12, p. e1002194, 2016.

[35] P. Danaher, S. Warren, L. Dennis, L. D'Amico, A. White, M. L. Disis, M. A. Geller, K. Odunsi, J. Beechem, and S. P. Fling, "Gene expression markers of tumor infiltrating leukocytes," *Journal for immunotherapy of cancer*, vol. 5, pp. 1–15, 2017.

[36] Y.-L. Huang, G. Ning, L.-B. Chen, Y.-F. Lian, Y.-R. Gu, J.-L. Wang, D.-M. Chen, H. Wei, and Y.-H. Huang, "Promising diagnostic and prognostic value of e2fs in human hepatocellular carcinoma," *Cancer Management and Research*, pp. 1725–1740, 2019.

[37] M. Hallek, B. D. Cheson, D. Catovsky, F. Caligaris-Cappio, G. Dighiero, H. Döhner, P. Hillmen, M. Keating, E. Montserrat, N. Chiorazzi, *et al.*, "iwcll guidelines for diagnosis, indications for treatment, response assessment, and supportive management of cll," *Blood, The Journal of the American Society of Hematology*, vol. 131, no. 25, pp. 2745–2760, 2018.

[38] M. Pichler, E. Winter, A. L. Ress, T. Bauernhofer, A. Gerger, T. Kiesslich, S. Lax, H. Samonigg, and G. Hoefler, "mir-181a is associated with poor clinical outcome in patients with colorectal cancer treated with egfr inhibitor," *Journal of clinical pathology*, vol. 67, no. 3, pp. 198–203, 2014.

[39] M. Dietel, K. Jöhrens, M. Laffert, M. Hummel, H. Bläker, B. Pfitzner, A. Lehmann, C. Denkert, S. Darb-Esfahani, D. Lenze, *et al.*, "A 2015 update on predictive molecular pathology and its role in targeted cancer therapy: a review focussing on clinical relevance," *Cancer gene therapy*, vol. 22, no. 9, pp. 417–430, 2015.

[40] R. S. A. Sattar, R. Verma, A. Kumar, G. M. Dar, A. K. Sharma, I. Kumari, E. Ahmad, A. Ali, B. Mahajan, S. S. Saluja, *et al.*, "Diagnostic and prognostic biomarkers in colorectal cancer and the potential role of exosomes in drug delivery," *Cellular Signalling*, p. 110413, 2022.

[41] C. Lugassy, P. B. Vermeulen, D. Ribatti, F. Pezzella, and R. L. Barnhill, "Vessel co-option and angiotropic extravascular migratory metastasis: a continuum of tumour growth and spread?," *British Journal of Cancer*, vol. 126, no. 7, pp. 973–980, 2022.

[42] W.-l. Cai, M. Cheng, Y. Wang, P.-h. Xu, X. Yang, Z.-w. Sun, and W.-j. Yan, "Prediction and related genes of cancer distant metastasis based on deep learning," *Computers in Biology and Medicine*, p. 107664, 2023.

[43] C. Wang and D. Luo, "The metabolic adaptation mechanism of metastatic organotropism," *Experimental Hematology & Oncology*, vol. 10, no. 1, pp. 1–16, 2021.

[44] J. Fares, M. Y. Fares, H. H. Khachfe, H. A. Salhab, and Y. Fares, "Molecular principles of metastasis: a hallmark of cancer revisited," *Signal transduction and targeted therapy*, vol. 5, no. 1, p. 28, 2020.

[45] F. Vidal-Vanaclocha, O. Crende, C. G. de Durango, A. Herreros-Pomares, S. López-Doménech, Á. González, E. Ruiz-Casares, T. Vilboux, R. Caruso, H. Durán, *et al.*, "Liver prometastatic reaction: Stimulating factors and responsive cancer phenotypes," in *Seminars in Cancer Biology*, vol. 71, pp. 122–133, Elsevier, 2021.

[46] R. Y. Lee, C. W. Ng, M. P. Rajapakse, N. Ang, J. P. S. Yeong, and M. C. Lau, "The promise and challenge of spatial omics in dissecting tumour microenvironment and the role of ai," *Frontiers in Oncology*, vol. 13, p. 1172314, 2023.

[47] K. E. de Visser and J. A. Joyce, "The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth," *Cancer Cell*, vol. 41, no. 3, pp. 374–403, 2023.

[48] C. Lugassy, H. K. Kleinman, P. B. Vermeulen, and R. L. Barnhill, "Angiotropism, pericytic mimicry and extravascular migratory metastasis: an embryogenesis-derived program of tumor spread," *Angiogenesis*, vol. 23, pp. 27–41, 2020.

[49] Z. Baumann, P. Auf der Maur, and M. Bentires-Alj, "Feed-forward loops between metastatic cancer cells and their microenvironment—the stage of escalation," *EMBO Molecular Medicine*, vol. 14, no. 6, p. e14283, 2022.

[50] G. M. Stella, S. Kolling, S. Benvenuti, and C. Bortolotto, "Lung-seeking metastases," *Cancers*, vol. 11, no. 7, p. 1010, 2019.

[51] K. Mortezaee, "Organ tropism in solid tumor metastasis: an updated review," *Future Oncology*, vol. 17, no. 15, pp. 1943–1961, 2021.

[52] Q. Ji, L. Zhou, H. Sui, L. Yang, X. Wu, Q. Song, R. Jia, R. Li, J. Sun, Z. Wang, *et al.*, "Primary tumors release itgbl1-rich extracellular vesicles to promote distal metastatic tumor growth through fibroblast-niche formation," *Nature communications*, vol. 11, no. 1, p. 1211, 2020.

[53] C. Kim, M. Hannouf, S. Sarma, G. Rodrigues, P. Rogan, S. Mahmud, E. Winquist, M. Brackstone, and G. Zaric, "Survival outcome differences based on treatments used and knowledge of the primary tumour site for patients with cancer of unknown and known primary in ontario," *Current Oncology*, vol. 25, no. 5, pp. 307–316, 2018.

[54] H. Chen, V. Chengalvala, H. Hu, and D. Sun, "Tumor-derived exosomes: Nanovesicles made by cancer cells to promote cancer metastasis," *Acta Pharmaceutica Sinica B*, vol. 11, no. 8, pp. 2136–2149, 2021.

[55] D. Bilder, K. Ong, T.-C. Hsi, K. Adiga, and J. Kim, "Tumour–host interactions through the lens of drosophila," *Nature Reviews Cancer*, vol. 21, no. 11, pp. 687–700, 2021.

[56] P. Gu, M. Sun, L. Li, Y. Yang, Z. Jiang, Y. Ge, W. Wang, W. Mu, and H. Wang, "Breast tumor-derived exosomal microrna-200b-3p promotes specific organ metastasis through regulating ccl2 expression in lung epithelial cells," *Frontiers in Cell and Developmental Biology*, vol. 9, p. 657158, 2021.

[57] M. Riihimäki, H. Thomsen, A. Hemminki, K. Sundquist, and K. Hemminki, "Comparison of survival of patients with metastases from known versus unknown primaries: survival in metastatic cancer," *BMC cancer*, vol. 13, pp. 1–8, 2013.

[58] S. A. Nawaz, D. M. Khan, and S. Qadri, "Brain tumor classification based on hybrid optimized multi-features analysis using magnetic resonance imaging dataset," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2031824, 2022.

[59] T. Olivier, E. Fernandez, I. Labidi-Galy, P.-Y. Dietrich, V. Rodriguez-Bravo, G. Baciarello, K. Fizazi, and A. Patrikidou, "Redefining cancer of unknown primary: Is precision medicine really shifting the paradigm?," *Cancer treatment reviews*, vol. 97, p. 102204, 2021.

[60] E. Rassy and N. Pavlidis, "The currently declining incidence of cancer of unknown primary," *Cancer epidemiology*, vol. 61, pp. 139–141, 2019.

[61] K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," *Genome Medicine*, vol. 13, no. 1, pp. 1–17, 2021.

[62] Y. Liang, H. Wang, J. Yang, X. Li, C. Dai, P. Shao, G. Tian, B. Wang, and Y. Wang, "A deep learning framework to predict tumor tissue-of-origin based on copy number alteration," *Frontiers in bioengineering and biotechnology*, vol. 8, p. 701, 2020.

[63] S. Jeong, D. Lee, H. R. Park, J. Kang, Y. Yu, J. J. Hwang, and Y. H. Kim, "Deepcia: a novel deep-learning model for cancer type identification using class activation map via transcription factor expression," *American Journal of Cancer Research*, vol. 12, no. 12, p. 5631, 2022.

[64] M. Divate, A. Tyagi, D. J. Richard, P. A. Prasad, H. Gowda, and S. H. Nagaraj, "Deep learning-based pan-cancer classification model reveals tissue-of-origin specific gene expression signatures," *Cancers*, vol. 14, no. 5, p. 1185, 2022.

[65] K. Guo, M. Wu, Z. Soo, Y. Yang, Y. Zhang, Q. Zhang, H. Lin, M. Grosser, D. Venter, G. Zhang, *et al.*, "Artificial intelligence-driven biomedical genomics," *Knowledge-Based Systems*, p. 110937, 2023.

[66] T. Yoshida, Y. Yatabe, K. Kato, G. Ishii, A. Hamada, H. Mano, K. Sunami, N. Yamamoto, and T. Kohno, "The evolution of cancer genomic medicine in japan and the role of the national cancer center japan," *Cancer Biology & Medicine*, 2023.

[67] A. La Ferlita, S. Alaimo, A. Ferro, and A. Pulvirenti, "Pathway analysis for cancer research and precision oncology applications," in *Computational Methods for Precision Oncology*, pp. 143–161, Springer, 2022.

[68] A. Beg and R. Parveen, "Role of bioinformatics in cancer research and drug development," in *Translational bioinformatics in healthcare and medicine*, pp. 141–148, Elsevier, 2021.

[69] X. Pan, X. Lin, D. Cao, X. Zeng, P. S. Yu, L. He, R. Nussinov, and F. Cheng, "Deep learning for drug repurposing: Methods, databases, and applications," *Wiley interdisciplinary reviews: Computational molecular science*, vol. 12, no. 4, p. e1597, 2022.

[70] A. Pulvirenti, G. F. Privitera, S. Alaimo, G. Micale, L. Giaimi, M. Mare, E. Martorana, R. Villa, A. Ferro, and S. Forte, "Oncoreport: a system for integrative ngs analysis in precision medicine," 2023.

[71] L. Wang, Z. Wu, C. Xu, and H. Ye, "Ferroptosis-related genes prognostic signature for pancreatic cancer and immune infiltration: potential biomarkers for predicting overall survival," *Journal of Cancer Research and Clinical Oncology*, pp. 1–16, 2023.

[72] P. Mirabelli, L. Coppola, and M. Salvatore, "Cancer cell lines are useful model systems for medical research," *Cancers*, vol. 11, no. 8, p. 1098, 2019.

[73] S. Rao, B. Pitel, A. H. Wagner, S. M. Boca, M. McCoy, I. King, S. Gupta, B. H. Park, J. L. Warner, J. Chen, *et al.*, "Collaborative, multidisciplinary evaluation of cancer variants through virtual molecular tumor boards informs local clinical practices," *JCO clinical cancer informatics*, vol. 4, pp. 602–613, 2020.

[74] T. E. Tavolara, Z. Su, M. N. Gurcan, and M. K. K. Niazi, "One label is all you need: Interpretable ai-enhanced histopathology for oncology," in *Seminars in Cancer Biology*, Elsevier, 2023.

[75] N. El-Sayes, A. Vito, and K. Mossman, "Tumor heterogeneity: a great barrier in the age of cancer immunotherapy," *Cancers*, vol. 13, no. 4, p. 806, 2021.

[76] M. L. Rosano-Gonzalez, V. T. Sreedharan, A. Hanns, D. J. Stekhoven, and F. Singer, "Civicutils: Matching and downstream processing of clinical annotations from civic," *F1000Research*, vol. 12, p. 1304, 2023.

[77] E. Tasci, S. Jagasia, Y. Zhuge, K. Camphausen, and A. V. Krauze, "Gradwise: A novel application of a rank-based weighted hybrid filter and embedded feature selection method for glioma grading with clinical and molecular characteristics," *Cancers*, vol. 15, no. 18, p. 4628, 2023.

[78] J. M. Burgener, J. Zou, Z. Zhao, Y. Zheng, S. Y. Shen, S. H. Huang, S. Keshavarzi, W. Xu, F.-F. Liu, G. Liu, *et al.*, "Tumor-naïve multimodal profiling of circulating tumor dna in head and neck squamous cell carcinoma," *Clinical Cancer Research*, vol. 27, no. 15, pp. 4230–4244, 2021.

[79] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, p. baaa010, 2020.

[80] M. Frank, D. Drikakis, and V. Charissis, "Machine-learning methods for computational science and engineering," *Computation*, vol. 8, no. 1, p. 15, 2020.

[81] M. J. Iqbal, Z. Javed, H. Sadia, I. A. Qureshi, A. Irshad, R. Ahmed, K. Malik, S. Raza, A. Abbas, R. Pezzani, *et al.*, "Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future," *Cancer cell international*, vol. 21, no. 1, pp. 1–11, 2021.

[82] Z. Amiri, A. Heidari, N. J. Navimipour, M. Unal, and A. Mousavi, "Adventures in data analysis: A systematic review of deep learning techniques for pattern recognition in cyber-physical-social systems," *Multimedia Tools and Applications*, pp. 1–65, 2023.

[83] C. Sarkar, B. Das, V. S. Rawat, J. B. Wahlang, A. Nongpiur, I. Tiewsoh, N. M. Lyngdoh, D. Das, M. Bidarolli, and H. T. Sony, "Artificial intelligence and machine learning technology driven modern drug discovery and development," *International Journal of Molecular Sciences*, vol. 24, no. 3, p. 2026, 2023.

[84] L. Yang and A. Shami, "Iot data analytics in dynamic environments: From an automated machine learning perspective," *Engineering Applications of Artificial Intelligence*, vol. 116, p. 105366, 2022.

[85] U. M. R. Paturi, S. T. Palakurthy, and N. Reddy, "The role of machine learning in tribology: a systematic review," *Archives of Computational Methods in Engineering*, vol. 30, no. 2, pp. 1345–1397, 2023.

[86] J. H. Harrison Jr, J. R. Gilbertson, M. G. Hanna, N. H. Olson, J. N. Seheult, J. M. Sorace, and M. N. Stram, "Introduction to artificial intelligence and machine learning for pathology," *Archives of pathology & laboratory medicine*, vol. 145, no. 10, pp. 1228–1254, 2021.

[87] P. Abdollahiyan, F. Oroojalian, B. Baradaran, M. de la Guardia, and A. Mokhtarzadeh, "Advanced mechanotherapy: Biotensegrity for governing metastatic tumor cell fate via modulating the extracellular matrix," *Journal of Controlled Release*, vol. 335, pp. 596–618, 2021.

[88] M. Farshbafnadi and N. Rezaei, "The metabolism of cancer cells during metastasis," in *Handbook of Cancer and Immunology*, pp. 1–21, Springer, 2023.

[89] P. E. Saw and E. Song, "Distal onco-sphere: The origin and overview of cancer metastasis," in *Tumor Ecosystem: An Ecological View of Cancer Growth and Survival*, pp. 289–305, Springer, 2023.

[90] Z. Chen, M. Jobayer, M. R. Hasan, K. A. Ahmed, and M. Z. Hossain, "Mutfusvae: Mutational fusion variational autoencoder for predicting primary sites of cancer," *Procedia Computer Science*, vol. 222, pp. 272–283, 2023.

[91] V. Zelli, A. Manno, C. Compagnoni, R. O. Ibraheem, F. Zazzeroni, E. Alesse, F. Rossi, C. Arbib, and A. Tessitore, "Classification of tumor types using xgboost machine learning model: a vector space transformation of genomic alterations," *Journal of Translational Medicine*, vol. 21, no. 1, p. 836, 2023.

[92] S. A. Bhat, S. M. Ahmad, P. T. Mumtaz, A. A. Malik, M. A. Dar, U. Urwat, R. A. Shah, and N. A. Ganai, "Long non-coding rnas: Mechanism of action

and functional utility," *Non-coding RNA research*, vol. 1, no. 1, pp. 43–50, 2016.

[93] K. Katsushima, G. Jallo, C. G. Eberhart, and R. J. Perera, "Long non-coding rnas in brain tumors," *NAR cancer*, vol. 3, no. 1, p. zcaa041, 2021.

[94] A. Misawa, K.-i. Takayama, and S. Inoue, "Long non-coding rnas and prostate cancer," *Cancer science*, vol. 108, no. 11, pp. 2107–2114, 2017.

[95] A. B. Shinagare, N. H. Ramaiya, J. P. Jagannathan, F. M. Fennessy, M.-E. Taplin, and A. D. Van den Abbeele, "Metastatic pattern of bladder cancer: correlation with the characteristics of the primary tumor," *American Journal of Roentgenology*, vol. 196, no. 1, pp. 117–122, 2011.

[96] L. Zhao, X. Liang, L. Wang, and X. Zhang, "The role of mirna in ovarian cancer: an overview," *Reproductive Sciences*, pp. 1–8, 2022.

[97] K. S. Siveen, K. S. Prabhu, I. W. Achkar, S. Kuttikrishnan, S. Shyam, A. Q. Khan, M. Merhi, S. Dermime, and S. Uddin, "Role of non receptor tyrosine kinases in hematological malignances and its targeting by natural products," *Molecular cancer*, vol. 17, no. 1, pp. 1–21, 2018.

[98] C. Yue, M. Niu, Q. Q. Shan, T. Zhou, Y. Tu, P. Xie, L. Hua, R. Yu, and X. Liu, "High expression of bruton's tyrosine kinase (btk) is required for egfr-induced nf-κb activation and predicts poor prognosis in human glioma," *Journal of Experimental & Clinical Cancer Research*, vol. 36, no. 1, pp. 1–11, 2017.

[99] A. K. Nakamura-García and J. Espinal-Enríquez, "Pseudogenes in cancer: State of the art," *Cancers*, vol. 15, no. 16, p. 4024, 2023.

[100] S. Wang, Y. Zheng, F. Yang, L. Zhu, X.-Q. Zhu, Z.-F. Wang, X.-L. Wu, C.-H. Zhou, J.-Y. Yan, B.-Y. Hu, *et al.*, "The molecular biology of pancreatic adenocarcinoma: Translational challenges and clinical perspectives," *Signal transduction and targeted therapy*, vol. 6, no. 1, p. 249, 2021.

[101] U. C. S. W. Group *et al.*, "Us cancer statistics data visualizations tool, based on 2022 submission data (1999–2020): Us department of health and human services, centers for disease control and prevention and national cancer institute," *Centers for Disease Control and Prevention and National Cancer Institute*, 2023.

[102] K. Xu, J. Cui, V. Olman, Q. Yang, D. Puett, and Y. Xu, "A comparative analysis of gene-expression data of multiple cancer types," *PloS one*, vol. 5, no. 10, p. e13696, 2010.

[103] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "Genbank," *Nucleic acids research*, vol. 41, no. D1, pp. D36–D42, 2012.

[104] L. Jin, B. Han, E. Siegel, Y. Cui, A. Giuliano, and X. Cui, "Breast cancer lung metastasis: Molecular biology and therapeutic implications," *Cancer biology & therapy*, vol. 19, no. 10, pp. 858–868, 2018.

[105] F.-W. Lin, M.-H. Yeh, C.-L. Lin, and J. C.-C. Wei, "Association between breast cancer and second primary lung cancer among the female population in taiwan: A nationwide population-based cohort study," *Cancers*, vol. 14, no. 12, p. 2977, 2022.

[106] J. S. Damrauer, K. A. Hoadley, D. D. Chism, C. Fan, C. J. Tiganelli, S. E. Wobker, J. J. Yeh, M. I. Milowsky, G. Iyer, J. S. Parker, *et al.*, "Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology," *Proceedings of the national academy of sciences*, vol. 111, no. 8, pp. 3110–3115, 2014.

[107] M. T. Chang, A. Penson, N. B. Desai, N. D. Socci, R. Shen, V. E. Seshan, R. Kundra, A. Abeshouse, A. Viale, E. K. Cha, *et al.*, "Small-cell carcinomas of the bladder and lung are characterized by a convergent but distinct pathogenesis," *Clinical Cancer Research*, vol. 24, no. 8, pp. 1965–1973, 2018.

[108] W. Yan, I. I. Wistuba, M. R. Emmert-Buck, and H. S. Erickson, "Squamous cell carcinoma–similarities and differences among anatomical sites," *American journal of cancer research*, vol. 1, no. 3, p. 275, 2011.

[109] D. Denisko and M. M. Hoffman, "Classification and interaction in random forests," *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1690–1692, 2018.

[110] B. Farajollahi, M. Mehmannavaz, H. Mehrjoo, F. Moghbeli, and M. J. Sayadi, "Diabetes diagnosis using machine learning," *Frontiers in Health Informatics*, vol. 10, no. 1, p. 65, 2021.

[111] R. Chandra, M. Kapil, and A. Sharma, "Comparative analysis of machine learning techniques with principal component analysis on kidney and heart disease," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1965–1973, IEEE, 2021.

[112] J. Hong, L. D. Hachem, and M. G. Fehlings, "A deep learning model to classify neoplastic state and tissue origin from transcriptomic data," *Scientific reports*, vol. 12, no. 1, p. 9669, 2022.

[113] F. Brasó-Maristany, L. Paré, N. Chic, O. Martínez-Sáez, T. Pascual, M. Mallafré-Larrosa, F. Schettini, B. González-Farré, E. Sanfeliu, D. Martínez, *et al.*, "Gene expression profiles of breast cancer metastasis according to organ site," *Molecular oncology*, vol. 16, no. 1, pp. 69–87, 2022.

[114] N. Laprovitera, M. Riefolo, E. Ambrosini, C. Klec, M. Pichler, and M. Ferracin, "Cancer of unknown primary: challenges and progress in clinical management," *Cancers*, vol. 13, no. 3, p. 451, 2021.

[115] M. S. Lee and H. K. Sanoff, "Cancer of unknown primary," *Bmj*, vol. 371, 2020.

[116] A. Papanicolau-Sengos and K. Aldape, "Dna methylation profiling: an emerging paradigm for cancer diagnosis," *Annual Review of Pathology: Mechanisms of Disease*, vol. 17, pp. 295–321, 2022.

[117] K. Galbraith, V. Vasudevaraja, J. Serrano, G. Shen, I. Tran, N. Abdallat, M. Wen, S. Patel, M. Movahed-Ezazi, A. Faustin, *et al.*, "Clinical utility of whole-genome dna methylation profiling as a primary molecular diagnostic assay for central nervous system tumors—a prospective study and guidelines for clinical testing," *Neuro-oncology advances*, vol. 5, no. 1, p. vdad076, 2023.

[118] S. Zhang, S. He, X. Zhu, Y. Wang, Q. Xie, X. Song, C. Xu, W. Wang, L. Xing, C. Xia, *et al.*, "Dna methylation profiling to determine the primary sites of metastatic cancers using formalin-fixed paraffin-embedded tissues," *Nature Communications*, vol. 14, no. 1, p. 5686, 2023.