People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

M'hamed BOUGARA University of Boumerdes

Faculty of technology

Electrical Systems Engineering Department

Final Year Project Report Presented in Partial Fulfilment of the Requirements for the Degree of

Master

In Biomedical Engineering

Option: Biomedical Instrumentation

Title:

Machine learning classifiers for predicting the presence of cancer using gene expression data from CTCs/CTMs

Presented by:

BOUDALI Maya

Supervisor:

Dr. AMMAR Mohammed

2023/2024

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Allah.

His boundless blessings have endowed me with the strength, patience, and perseverance to complete this journey. Without His grace and blessings, none of this would have been possible.

I am profoundly grateful to my supervisor, Dr. AMMAR.

Over the past two years, your unwavering support and guidance have been invaluable. Thank you for believing in my potential from the beginning and allowing me to pursue this thesis project. Your insightful feedback, patience, and encouragement have guided me through every stage of this work. I am forever grateful for your kindness, expertise, and the profound impact you've had on my academic journey.

I extend my heartfelt thanks to all faculty members and staff who have contributed to my academic journey. Their support in various forms has been pivotal in my achievements.

Dedication

To my beloved parents, your guidance and support have been the cornerstone of my journey. You have been my source of strength, encouragement, and belief when challenges seem insurmountable. This thesis is dedicated to you with profound gratitude for your sacrifices and for instilling in me values of resilience and determination. You have shaped me into the person I am today, and I am forever grateful for the foundation you have provided.

To the best siblings in this world Ines, Nadia, Yacine, and Nawel, your unwavering love, encouragement, and sacrifices have been my pillar of strength throughout this journey. You believed in me when I doubted myself, supported me through every challenge, and celebrated my achievements with boundless joy. This work is dedicated to you as a token of my deepest gratitude for your endless support and unwavering belief in my dreams. You are truly one of Allah's greatest blessings in my life.

To my beloved nephews Aziz, Raouf, Wassim, and nieces Alaa, Rym, Ania, Yasmine, Noursine, and Taj, your laughter, curiosity, and boundless energy have illuminated my life. This thesis is dedicated to you, with love and pride. May you continue to shine brightly, lighting up the world with your unique brilliance.

To my dearest friends, Nadia, Khaoula, Ikram, Khadidja, Bouchra, Lina and Sarah, who are more like sisters, your companionship, laughter, and shared moments have woven a tapestry of love and support in my life. Through ups and downs, you've stood by me, offering encouragement, understanding, and warmth. Thank you for everything.

To all who shaped my journey, This work is dedicated to you All.

Table of contents

Chapter I: An Overview of Cancer

Chapter II: Artificial Intelligence

Chapter III: The experimental part

List des figures

List of Tables

Abstract:

The complexity and heterogeneity of cancer make early detection and effective treatment particularly difficult. Traditional diagnostic methods often involve invasive tissue biopsies, which can be uncomfortable and pose risks to patients. Recent advancements in non-invasive liquid biopsies, specifically through the analysis of circulating tumor cells (CTCs) and circulating tumor microemboli (CTMs), offer a promising alternative, by examining the gene expression profiles of CTCs, researchers can identify potential cancer biomarkers, facilitating early detection and characterization of various cancer types.

This thesis focuses on developing and evaluating machine learning classifiers for predicting the presence of seven types of cancer using gene expression data from CTCs and CTMs. The cancers investigated include liver cancer, breast cancer, colorectal cancer, non-small cell lung cancer, pancreatic cancer, prostate cancer, and melanoma. The study involves building binary classifiers to distinguish each cancer type from others and multiclass classifiers to predict all seven cancer types. The goal is to compare these approaches and identify the most effective model for accurate cancer prediction.

The findings demonstrate the significant potential of machine learning models in enhancing cancer diagnostics using minimally invasive methods.Among the models evaluated, the Random Forest multi-classifier emerged as the most reliable and effective, making it highly recommended for practical use in cancer diagnosis.

Keywords: Cancer prediction, Machine learning, Deep learning, Gene expression data, Circulating tumor cells (CTCs), Circulating tumor microemboli (CTMs), Liquid biopsy, Binary classification, Multiclass classification, Cancer biomarkers, Non-invasive diagnostics.

Résumé :

La complexité et l'hétérogénéité du cancer rendent la détection précoce et le traitement efficace particulièrement difficiles. Les méthodes de diagnostic traditionnelles impliquent souvent des biopsies tissulaires invasives, qui peuvent être inconfortables et présenter des risques pour les patients. Les progrès récents dans les biopsies liquides non invasives, notamment grâce à l'analyse des cellules tumorales circulantes (CTC) et des micro-emboles tumoraux circulants (CTM), offrent une alternative prometteuse. En examinant les profils d'expression génétique des CTC, les chercheurs peuvent identifier des bio-marqueurs potentiels du cancer, facilitant la détection précoce et la caractérisation de divers types de cancer.

Cette thèse porte sur le développement et l'évaluation de classificateurs d'apprentissage automatique permettant de prédire la présence de sept types de cancer à l'aide de données d'expression génétique provenant de CTC et de CTM. Les cancers étudiés comprennent le cancer du foie, le cancer du sein, le cancer colorectal, le cancer du poumon non à petites cellules, le cancer du pancréas, le cancer de la prostate et le mélanome. L'étude consiste à créer des classificateurs binaires pour distinguer chaque type de cancer des autres et des classificateurs multi-classes pour prédire les sept types de cancer. L'objectif est de comparer ces approches et d'identifier le modèle le plus efficace pour une prédiction précise du cancer.

Mots clés: Prédiction du cancer, Apprentissage automatique, Apprentissage profond, Données d'expression génétique, Cellules tumorales circulantes (CTC), Micro-emboles tumorales circulantes (CTM), Biopsie liquide, Classification binaire, Classification multi-classe, Bio-marqueurs du cancer, Diagnostics non invasifs.

Introduction

Introduction:

Cancer remains one of the leading causes of morbidity and mortality worldwide, with millions of new cases diagnosed each year. The complexity and heterogeneity of cancer pose significant challenges to early detection and effective treatment. Traditional diagnostic methods often rely on invasive tissue biopsies, which can be painful and carry risks for patients. In contrast, recent advancements in non-invasive liquid biopsies offer a promising alternative by analyzing circulating tumor cells (CTCs) and circulating tumor microemboli (CTMs) in the bloodstream.

CTCs are cancer cells that detach from the primary tumor and enter the bloodstream, serving as key indicators of metastasis and disease progression. By examining the gene expression profiles of CTCs, researchers can identify potential cancer biomarkers, aiding in the early detection and characterization of various cancer types.

This thesis explores the development of machine learning classifiers for predicting the presence of seven types of cancer using gene expression data from CTCs and CTMs. The seven cancer types investigated in this study are liver cancer, breast cancer, colorectal cancer, non-small cell lung cancer, pancreatic cancer, prostate cancer, and melanoma. Each of these cancers presents unique challenges in terms of early detection and accurate diagnosis. By leveraging gene expression data from CTCs and CTMs, I aim to improve the predictive accuracy of cancer diagnostics, potentially leading to earlier and more precise interventions.

The project involves building and evaluating machine learning models through two distinct approaches. Initially, four binary classifiers were developed, each trained to distinguish a specific cancer type from the combined presence of other cancers. Subsequently, the same models were adapted to function as multi-classifiers, capable of predicting the presence of any of the seven cancer types. This comparative analysis aims to identify the best-performing model, providing insights into the most effective method for cancer prediction using CTC and CTM gene expression data.

The findings of this study have the potential to enhance the field of cancer diagnostics, offering a data-driven approach to detecting multiple cancer types from minimally invasive samples.

Chapter I

An Overview of Cancer

Introduction:

Today, one-third of male deaths and one-quarter of female deaths are due to cancer. In industrialized countries, cancer ranks second among the leading causes of mortality after cardiovascular diseases, and according to epidemiological data, this trend is emerging in less developed countries **[1]**.

The improvement of cancer prevention techniques and early detection enables the majority of patients to be cured after adjuvant treatment. However, this disease remains fatal in the majority of cases **[2]**.

I- Cancer:

I-1- A historical perspective on cancer research:

Cancer has afflicted people for several centuries. The oldest documented case of cancer dates back to 3000 b.c in ancient Egypt. The details, recorded on a papyrus, described 8 cases of tumors occurring on the breast. Historical evidence suggests that ancient Egyptians possessed the ability to distinguish between malignant and benign tumors. According to inscriptions, surface tumors were surgically removed in a similar manner as they are removed today but there was no acknowledged treatment for the condition **[3].** Other types of cancers were also described, in 1500 BC, including stomach, uterus, rectum, and skin cancers. At that time, to explain this phenomenon, Egyptians used the term “incurable disease” or “the curse of God” **[4].** The actual name of the disease was given by Hippocrates in 460-370 BC who used the Greek term carcinos to describe the crab-like lesions **[3].** He thought that the tumor resembles the crab in the way breast cancer spreads to the skin. The physician Celsus, later translated this word into cancer, the Latin word for crab **[4].** Hippocrates was also the first who explained cancer from a scientific point of view, considering it occurs as a result of remarkable presence or increase in the quantity of black bile in the body **[4]**. He only had an experience on external tumors since in ancient Greece (as in ancient Egypt) corpses could not be used for medical analysis **[3].** Ancients back then considered that once cancer has spread, there is no curable treatment, and the intervention could be more harmful. Significant progress in cancer surgeries started in the 19th and early 20th centuries, coinciding with a deeper scientific understanding of cancer. This investigation and explanation of cancer went through a long path up to a century ago, when Boveri’s referred to cancer as a genetic disease.

With the development of research, new visions are highlighted, especially nowadays with genetics analysis. These analyses keep open-ended questions that could be resolved only by research and more developed science **[4].**

Figure 01: Oldest evidence of human cancer osteocarcinoma. According to National geographic reports **[1].**

I-2- Cancer Nowadays:

According to the World Health Organization (WHO), cancer is considered a leading cause of death worldwide, after cardiovascular disease **[4].**

In 2020, global cancer statistics from the International Agency for Research on Cancer (IARC) revealed an estimated 19.3 million new cancer cases (excluding non-melanoma skin cancer) and nearly 10.0 million cancer-related deaths (excluding non-melanoma skin cancer) worldwide. Among the most frequently diagnosed cancers were female breast cancer (2.26 million cases), lung cancer (2.21 million cases), and prostate cancer (1.41 million cases). The leading causes of cancer-related mortality were lung cancer (1.79 million deaths), liver cancer (830,000 deaths), and stomach cancer (769,000 deaths) **[5].**

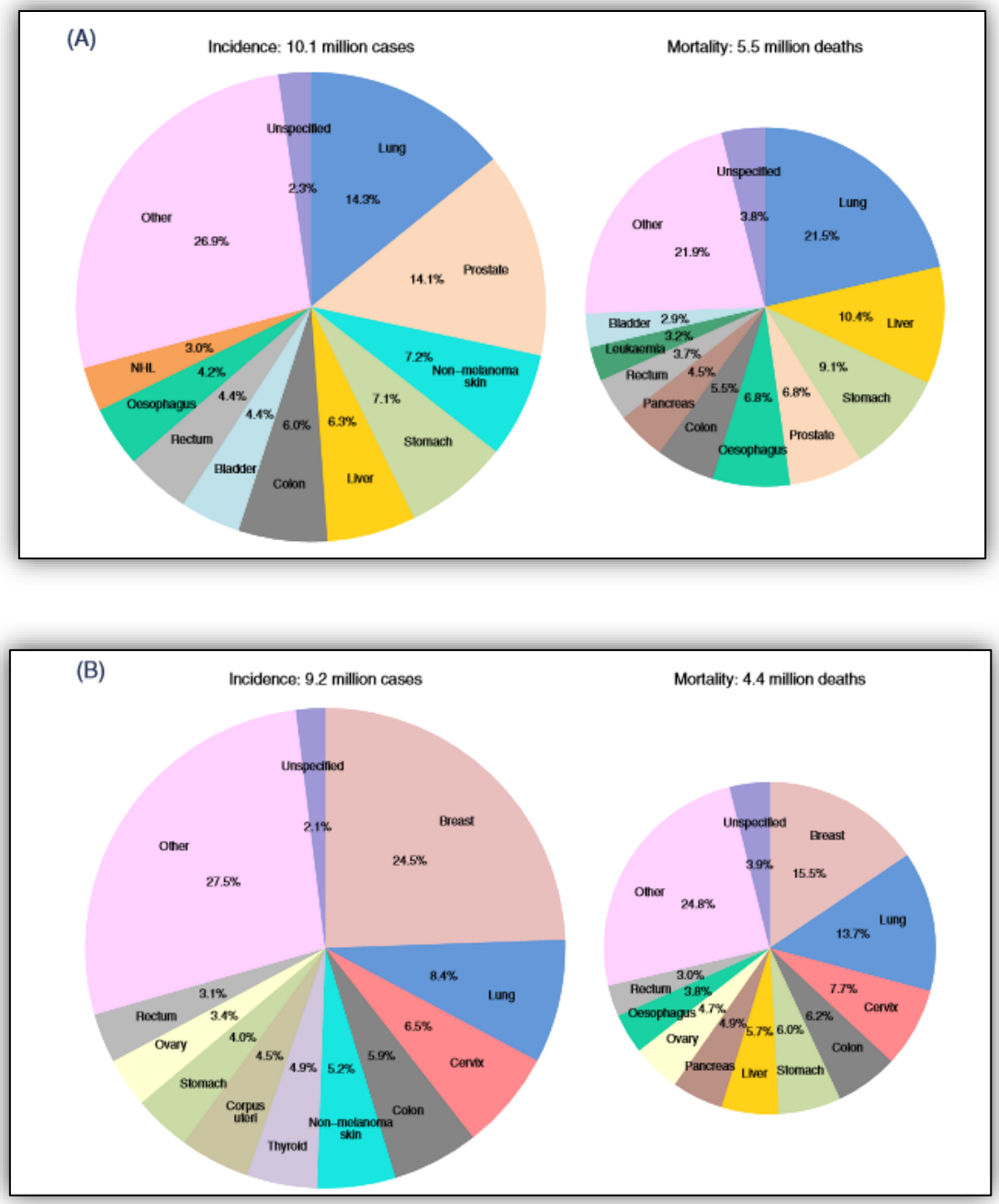

Figure 02: Distribution of the estimated new cases and deaths for the 10 most common cancers in 2020 in males (A) and females (B). For each sex, the area of the pie chart reflects the proportion of the total number of cases or deaths. NHL, non-Hodgkin lymphoma **[5].**

I-3- Definition of Cancer:

Cancer is a category of diseases characterized by the uncontrolled growth and spread of cells from the primary site to other parts of the body. With over 200 known types, cancers are classified based on the initially affected cell type. The unique characteristics of each cancer are determined by its tissue of origin. Approximately 85% of cancers, known as

carcinomas, occur in epithelial cells. Sarcomas originate from mesodermal cells such as bone and muscle, while adenocarcinomas arise from glandular tissues **[6] [7].**

Fundamentally, cancer is a disease of the genome, with a wide range of genomic alterations like point mutations, copy number changes, and rearrangements leading to its development. For many cancers, only 5-10% of cases are due to inherited genes, with BRCA1 and BRCA2 being examples for breast cancer. The majority of cases result from external factors like tobacco and radiation, which directly damage DNA. This damage can lead to various gene mutations, including the activation of oncogenes (cancer-causing genes) and the inhibition of tumor suppressor gene functions, ultimately leading to uncontrolled cell growth **[6].**

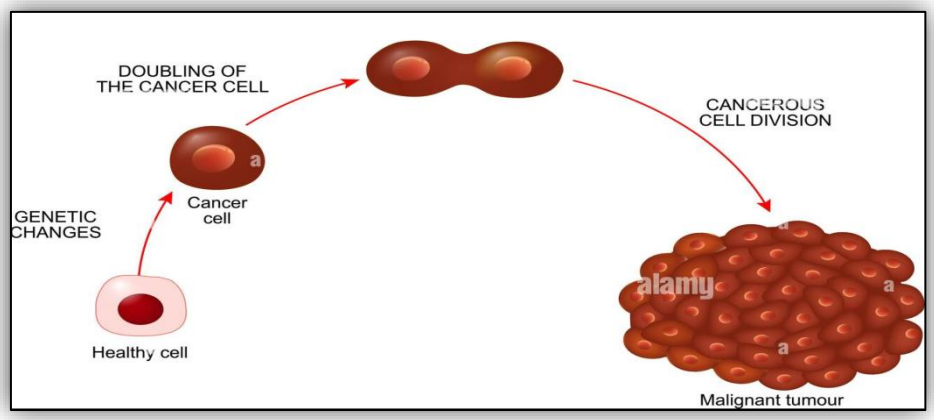

Figure 03: cancer cells.

I-4- Types of cancer:

Each type of cancer possessing a unique genetic profile. This genetic diversity among cancer types underscores the relationship between genetics and cancer. However, the scope of this thesis will be purposefully narrowed to focus on the metastasis of seven specific types of cancer. We will delve into the genetic underpinnings of these seven cancers, exploring the specific genes associated with them.

I-4-1- Breast cancer: a malignant tumor that originates from the epithelium of the breast tissue, has no single known cause **[8].** It is more prevalent in women but can also occur in men. The disease typically begins in the cells lining the milk ducts or the glands that produce milk **[9] [10] [11].**

Breast cancer is a complex disease with a significant genetic component **[12].** Its development and progression, both primary and metastatic, are driven by mutations in a multitude of genes **[13].** These include BRCA1/2, TP53, STK11, PTEN, CDH1, NF1, NBN, ATM, CHEK2, PALB2, RAD50, CCND1, ERBB2, CDK12, ADGRA2, ZNF703, FGFR1, KAT6A, POLB, COL1A1, AXIN2, ZNF217, GNAS, FGF3, FGF4, FGF19, Nf1, and Trps1 **[14] [15].**

In addition to these, mutations in mitochondrial tRNA genes, such as tRNA Val G1606A, tRNA Ile A4300G, tRNA Ser (UCN) T7505C, tRNA Glu A14693G, and tRNA Thr G15927A, have been associated with breast cancer development due to their impact on mitochondrial functions **[16].** Somatic mutations in PIK3CA, particularly in exons 9 and 20, have been identified in a significant percentage of breast cancer cases and are associated with negative lymph node status **[17].**

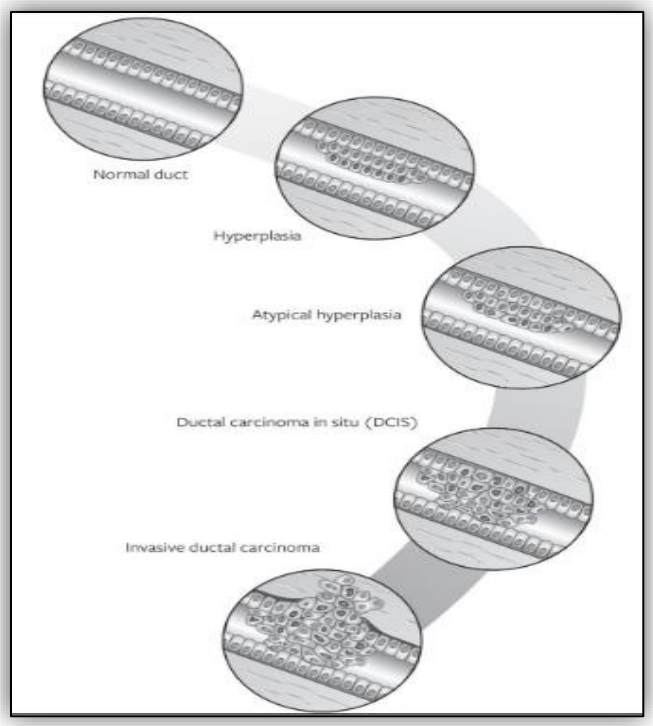

Figure 04: the spectrum of changes from normal to cancer in the ducts of the breast [**18].**

I-4-2- Colorectal cancer: which originates in the colon or rectum—parts of the lower digestive system—is the third most common cancer worldwide, with a cumulative risk of onset at 2.27% **[19].** Despite the availability of screening methods, most cases are only diagnosed after the onset of symptoms **[20].** The disease's leading risk factors include a family history, pre-cancerous conditions, physical inactivity, and dietary factors.

The development of colorectal cancer (CRC) is often associated with various genetic mutations. Research has identified several key mutations linked to CRC progression, including those in genes such as KRAS, TP53, APC, SMAD4, and FBXW7, which play pivotal roles in colorectal tumorigenesis **[21] [22] [23].**

Additionally, mutations in genes like CTNND1, AXIN1, TCF3, TGFBR1, RASGRF1, RASA1, and RAF1 have been identified as drivers in CRC, impacting pathways such as Wnt, TGF-β/BMP, and MAP kinase **[24] [25].** DNA methylation changes, particularly the hypermethylation of FIGN, HTRA3, BDNF, HCN4, and STAC2, have been associated with a poor prognosis in colorectal cancer patients **[26].**

Comparisons of the expression profiles of primary colorectal cancer and their metastatic lesions have led to the identification of early driver genes and metastasis-specific genes **[27].** Genetic biomarkers such as KRAS, NRAS, BRAF, HER2, and microsatellite instability have practical implications for the treatment of metastatic colorectal cancer **[28].**

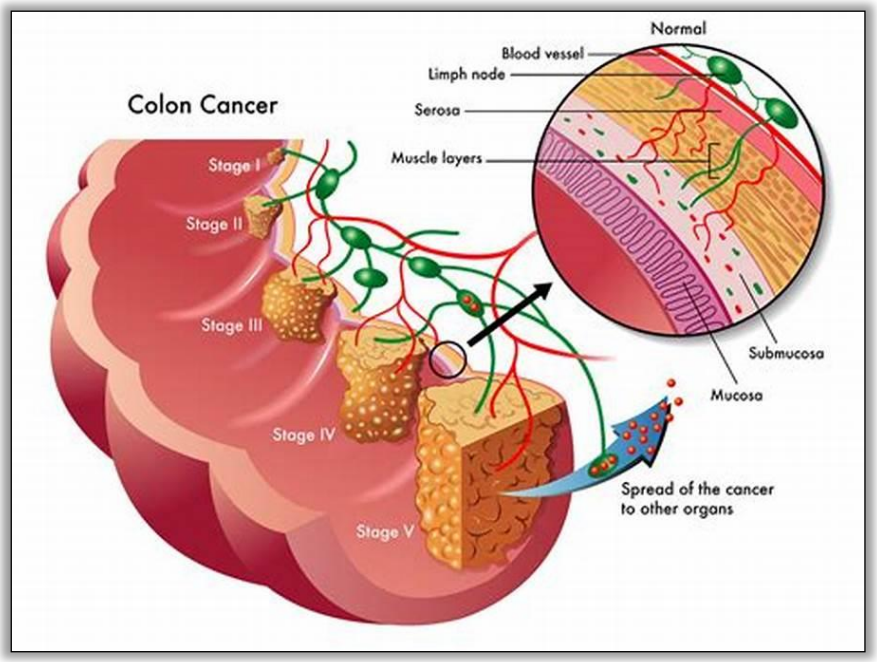

Figure 05: Colorectal cancer (CRC) stages and development **[29].**

I-4-3- Melanoma : a highly aggressive form of skin cancer, originates from melanocytes, the cells responsible for pigment formation **[30].** Its varied presentations make early detection vital for improved outcomes **[31].** Risk factors for melanoma include UV radiation exposure, light skin type, atypical nevi, and a family history of the disease **[31].**

Genetic research on melanoma has pinpointed several key genes and pathways implicated in the disease. Notably, mutations in the BRAF gene, especially the p.V600E and p.V600K mutations, are common in melanoma cases. The p.V600E mutation alone accounts for up to 95% of BRAF-mutant melanomas **[32].**

In the advanced stages of melanoma, mutations in NRAS, CDKN2A, TP53, PTEN, and the TERT promoter are frequently observed, contributing to tumor progression **[33].** TERT promoter mutations, such as the ATG start site −124C>T and −146C>T, are associated with increased TERT mRNA expression, heightened telomerase activity, and a poorer prognosis in melanoma patients **[34].**

Figure 06: Melanoma skin cancer **[35].**

I-4-4- Non-small cell lung cancer (NSCLC):

Is a complex disease characterized by genetic and cellular heterogeneity **[36].** As the most prevalent subtype of lung cancer, it carries a high mortality rate **[37].** Tobacco smoking is the primary risk factor for NSCLC, but radon exposure and air pollution also contribute **[38].**

A variety of genetic alterations in NSCLC have been identified in studies, including point mutations, indels, and gene fusions **[39].** Certain genes, such as TSPAN14, SLC2A13, and PHF20, are associated with the promotion of NSCLC, while others like CYP4Z1, KIR, and RDH10 are linked to its progression **[40].**

Deleterious mutations in the ATM gene are frequently found in NSCLC **[40] [41].** These mutations characterize a unique subset of NSCLC with distinct clinicopathologic,

genomic, and immunophenotypic features. These include associations with female sex, smoking history, non-squamous histology, and a higher tumor mutational burden **[42] [43].**

Furthermore, co-occurring mutations in KRAS, STK11, and ARID2 are significantly enriched in ATM-mutant NSCLCs. In contrast, TP53 and EGFR mutations are more prevalent in ATM wild-type cases **[44].**

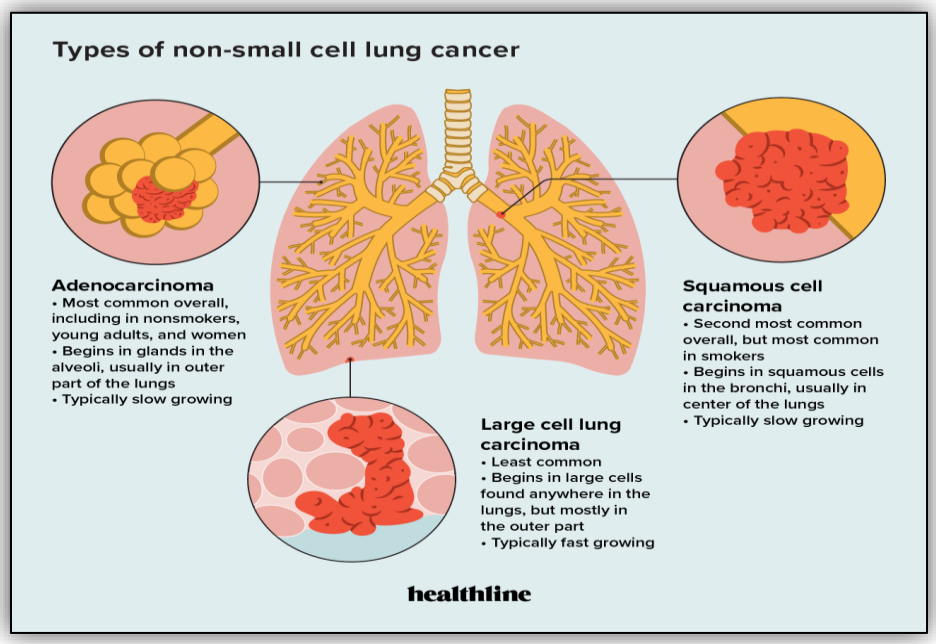

Figure 07: Types of non-small cell lung cancer **[45].**

I-4-5- Pancreatic cancer:

A highly lethal disease, often remains asymptomatic until it reaches advanced stages, making it difficult to detect **[46].** This disease is more prevalent in men and older individuals, and its incidence is on the rise. It spreads rapidly, with adenocarcinomas accounting for 90% of cases **[47].** Risk factors for pancreatic cancer include hereditary mutations, smoking, alcohol consumption, and a diet low in fruits and vegetables **[48].**

The development of pancreatic cancer is associated with several common genetic mutations, including those in the KRAS, TP53, SMAD4, CDKN2A, ARID1A, TENM4, TTN, RNF43, FLG, and GAS6 genes **[49] [50] [51].** Furthermore, mutations in genes such as BRCA1/2, PALB2, MLL3, TGFBR2, and SF3B1 have been linked to the most common form of pancreatic cancer **[52] [53].**

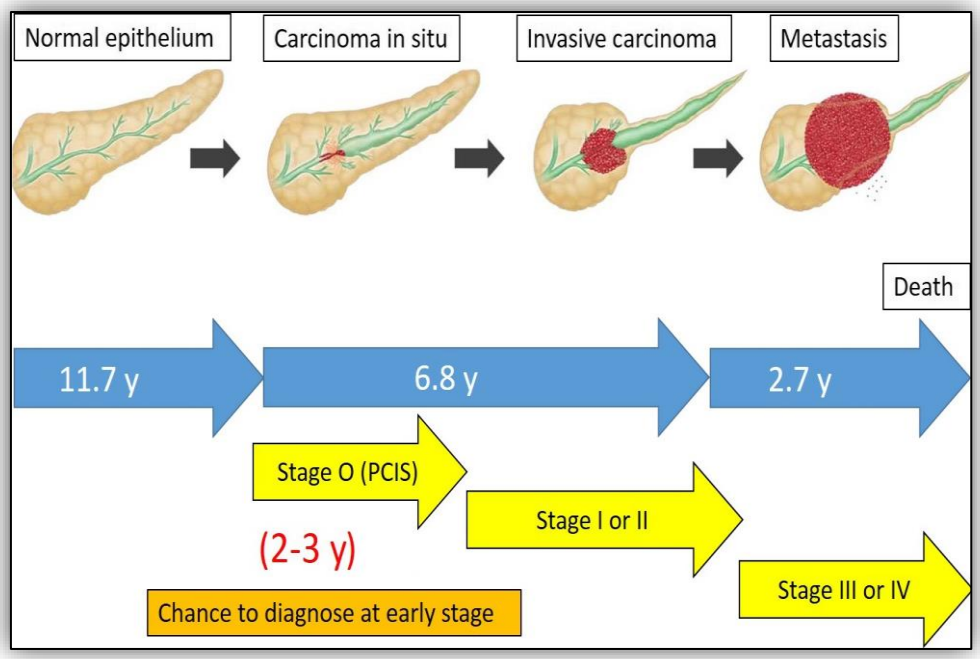

Figure 08: Progression model and stage of pancreatic cancer **[53].**

I-4-6- Prostate cancer:

The most common malignancy in elderly men, often presents with asymptomatic prostate nodules or bone pain **[54].** It has a high chance of cure when diagnosed early, even in metastatic cases **[55].** The genetics of prostate cancer have been extensively studied, with various genetic changes identified as potential contributors to the disease. Zhang identified several regions on chromosomes that may harbor genes predisposing individuals to prostate cancer **[56].** Both Roylance and Brothman discussed the consistent genetic changes and specific genes implicated in the development and progression of prostate tumors **[57][58].** Edwards further explored the existence of a high-risk gene in families with familial prostate cancer and the contribution of common lower penetrance genes, such as single nucleotide polymorphisms, to the disease **[59].**

Common genetic mutations associated with the development of prostate cancer include alterations in genes such as TERT, PTEN, BRCA1, BRCA2, and genes involved in the homologous recombination repair (HRR) pathway. Studies have shown that mutations in TERT are linked to prostate tumorigenesis and severity, with specific variants associated with aggressive prostate cancer **[60].** Additionally, mutations in PTEN have been identified as independent predictors of increased overall mortality among patients with prostate cancer,

regardless of smoking history **[61].** Germline mutations in BRCA1 and BRCA2 genes, known for their association with breast cancer, also increase the risk of developing prostate cancer **[62].** Mutations in HRR genes, such as TP53 and KRAS, have been detected in prostate cancer patients, suggesting potential targets for targeted treatment **[63].**

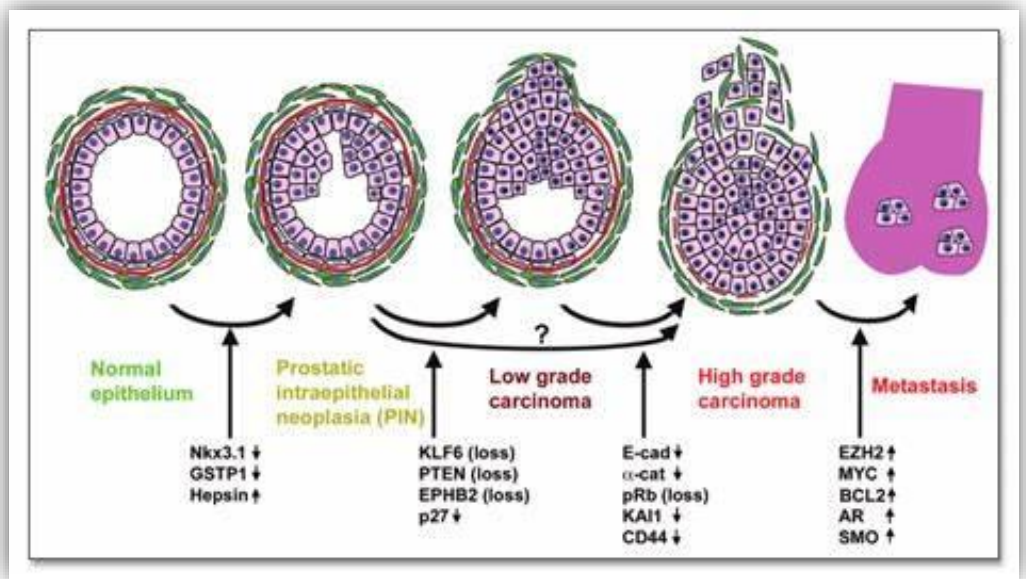

Figure 09: Model of prostate cancer progression. Morphologic features of different stages of prostate cancer progression correlate with specific genetic and epigenetic events. **[64].**

I-4-7- Liver cancer:

A complex and heterogeneous disease, is a significant global health concern due to its rapid progression and low life expectancy **[65].**This disease is characterized by a range of tumors, with hepatocellular carcinoma being the most prevalent and often having a poor prognosis due to late detection. Hepatoblastoma, on the other hand, is the type most commonly found in children**[66].**The incidence of liver cancer is on the rise, with over 900,000 new cases annually and a projected increase to over 1.4 million by 2040 **[65].**

Various genetic mutations are associated with liver cancer, including pathogenic alleles, non-coding mutations, single-nucleotide polymorphisms (SNPs), and alterations in genes such as TP53, catenin beta 1, Axin1, p16INK4, insulin-like growth factor 2 receptor, RB transcriptional corepressor 1, and cyclin D1 **[67] [68].** Specific mutations, including rs17401966 and BRAF V600E, have been implicated in the development of liver cancer. **[69] [70]**. Additionally, mutations in genes like EZH2 and CCND1 have been identified as

potential biomarkers for liver cancer, with a high tumor mutation burden (TMB) serving as a prognosis indicator. **[71].**

Hepatocellular carcinoma (HCC) development is influenced by a complex interplay of genetic and environmental factors. Researchers such as Ozen and Zhang have highlighted the role of genetic mutations and epigenetic aberrations in HCC development. Ozen, in particular, has identified mutated genes such as TP53, CTNNB1, AXIN1, and CDKN2A. These mutations, along with imprints of mutagenic exposure, contribute to the genomic landscape of liver cancer **[72] [73].**

The genetic basis for susceptibility to hepatocarcinogenesis has been further explored by Dragani, who has identified six regions on chromosomes 2, 5, 7, 8, 12, and 19 linked to hepatocellular tumor development in mice **[74].** Ding has further emphasized the importance of genetic factors in liver cancer development by discussing the role of loss of heterozygosity (LOH) in liver tumors **[75].**

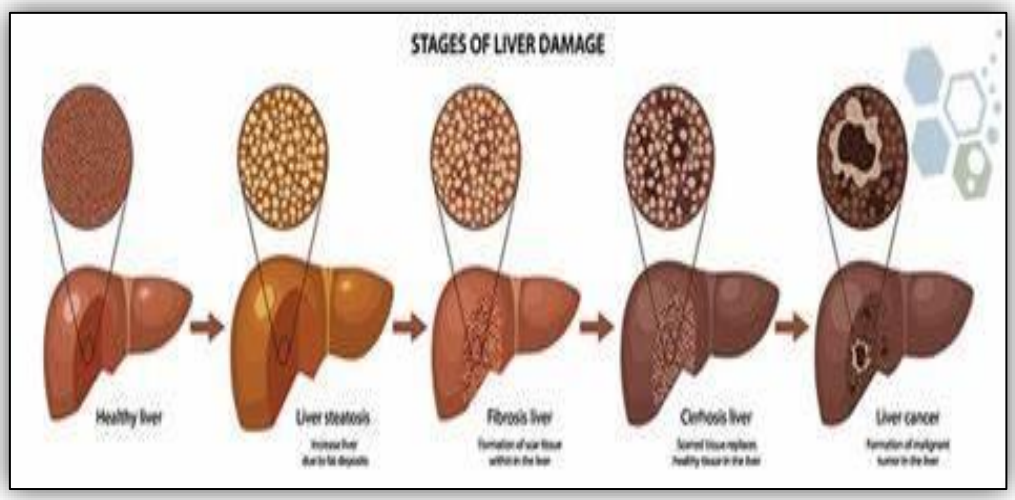

Figure 10: stages of liver damege **[76].**

I-5- Cellular Hallmarks of Cancer:

The hallmarks of cancer, a pivotal concept in oncology research, represent a set of fundamental characteristics that are shared by virtually all cancer cells. These hallmarks refer to the acquired biological capabilities that cancer cells develop during tumor development and progression. The number of cancer hallmarks has evolved over time, In 2000, researchers Douglas Hanahan and Robert Weinberg identified six key hallmarks shared by cancer cells,These hallmarks include sustaining proliferative signaling, evading growth suppressors,

resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis.This was later expanded to include two emerging hallmarks: reprogramming of energy metabolism and evading immune destruction **[77] [78] [79] [80].** Fast forward to 2021, Sasi S. Senga and Richard P. Grose proposed four additional hallmarks that shed light on cancer complexity. The new hallmarks were dedifferentiation and transdifferentiation, epigenetic dysregulation, altered microbiome and altered neuronal signaling. These new dimensions provide potential targets for therapeutic interventions, offering hope for improved cancer diagnosis and treatments **[4].**

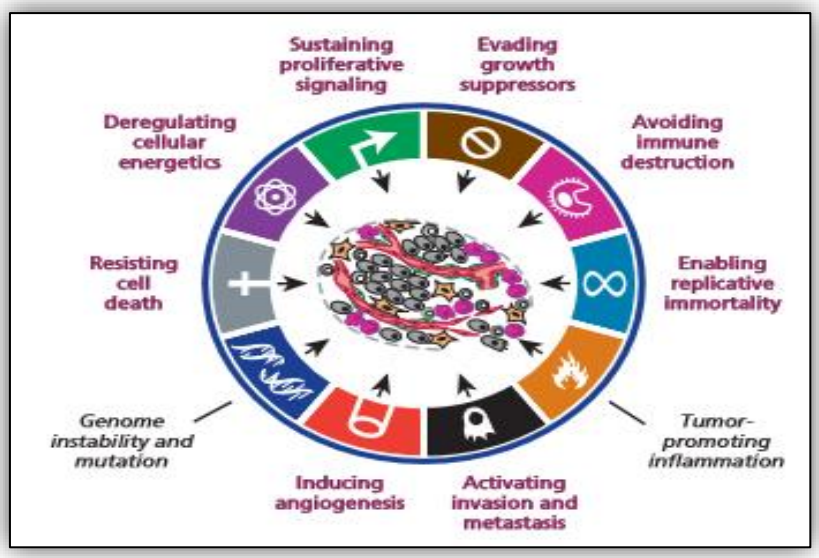

Figure 11: the Hallmarks of Cancer, circa 2022 **[79].**

- **Sustaining Proliferative Signaling:**

Cancer cells continuously receive growth-promoting signals, leading to uncontrolled cell proliferation. This capability is driven by activating mutated driver oncogenes, such as EGFR and KRAS, which sustain proliferative signaling pathways in cancer cells. These signals promote cell cycle progression and drive the uncontrolled growth of tumors, a fundamental aspect of cancer development and progression **[77].**

- **Evading Growth Suppressors:**

Cancer cells bypassing mechanisms that normally inhibit cell growth, such as tumor suppressor genes like p53 and Rb. These genes act as brakes on cell division, but in cancer cells, mutations or alterations allow them to override these growth-inhibitory signals, leading to uncontrolled proliferation **[81].**This hallmark is crucial for cancer cells to evade the natural

barriers that limit cell proliferation and maintain tissue homeostasis. Genetic profiling studies have shown that a majority of human tumors contain defects in the functions of tumor suppressor pathways, highlighting the importance of evading growth suppressors in cancer development and progression **[78].**

- **Resisting Cell Death:**

This hallmark is characterized by cancer cells evading programmed cell death (apoptosis) that would normally eliminate abnormal or damaged cells. Cancer cells can develop mechanisms to resist cell death signals, allowing them to survive and proliferate uncontrollably. This capability is crucial for cancer cells to evade one of the body's natural defense mechanisms against the development of tumors.Cancer cells can resist cell death through various mechanisms, such as mutations in apoptotic pathways, overexpression of anti-apoptotic proteins, or alterations in pro-survival signaling pathways. By avoiding cell death, cancer cells can continue to grow and accumulate genetic changes that drive tumor progression **[78].**

- **Enabling Replicative Immortality**:

Cancer cells maintain the ability to divide indefinitely by overcoming the natural limit on cell division imposed by telomere shortening. Cancer cells achieve this by activating mechanisms for telomere maintenance and extension, such as the expression of the telomere-extending enzyme telomerase or alternative recombination-based mechanisms.By acquiring the capability to maintain their telomeres, cancer cells can avoid the barrier of shortened telomeres and achieve cellular immortality, allowing for the continuous expansion of cancer cell populations **[78].**

- **Inducing Angiogenesis:**

Tumors stimulate the formation of new blood vessels to ensure a steady supply of oxygen, nutrients, and waste removal, supporting their growth and survival. This process involves the activation of angiogenic factors, such as vascular endothelial growth factor (VEGF), which promote the growth of blood vessels from pre-existing vasculature towards the tumor.The tumor-associated vasculature in cancer is often abnormal, with vessels that are tortuous, dilated, and leaky, leading to erratic blood flow patterns. Angiogenesis is crucial for tumor progression, enabling cancer cells to access the nutrients and oxygen necessary for their

survival and proliferation. While some tumors may co-opt existing blood vessels, most rely on chronic angiogenesis to sustain their growth **[78].**

- **Activating Invasion and Metastasis:**

Cancer cells acquire the ability to invade surrounding tissues and migrate to distant sites, leading to the formation of secondary tumors (metastases). This capability enables cancer cells to intravasate into blood or lymphatic vessels, travel to distant organs, and establish new tumor colonies. The process of invasion and metastasis is complex and involves both cell-intrinsic programs, such as the epithelialmesenchymal transition (EMT), and interactions with the tumor microenvironment, including hypoxia-induced signaling pathways.The regulation of invasion and metastasis is crucial for the spread of cancer cells and involves intricate molecular mechanisms that facilitate migration, invasion, survival in the bloodstream, and colonization of distant tissues **[78].**

- **Deregulating Cellular Energetics and Metabolism**:

Deregulating Cellular Energetics and Metabolism is a hallmark of cancer characterized by alterations in how cancer cells utilize energy sources and metabolize nutrients to support their rapid proliferation. This hallmark was initially described by Otto Warburg almost 90 years ago, highlighting the observation that cancer cells exhibit increased glucose uptake and glycolysis even in the presence of oxygen, known as the Warburg effect.Cancer cells often rely on glycolysis for energy production, which is less efficient than oxidative phosphorylation but provides essential building blocks for cell growth and division. Additionally, cancer cells may exhibit altered metabolism of other nutrients like glutamine to support their energetic and biosynthetic needs **[78].**

- **Avoiding Immune Destruction:**

Cancer cells develop strategies to evade detection and elimination by the immune system, allowing them to survive and proliferate. This hallmark involves mechanisms that enable cancer cells to escape immune surveillance, such as downregulating antigen presentation, inducing immune tolerance, or expressing immune checkpoint molecules that inhibit immune responses **[78].**

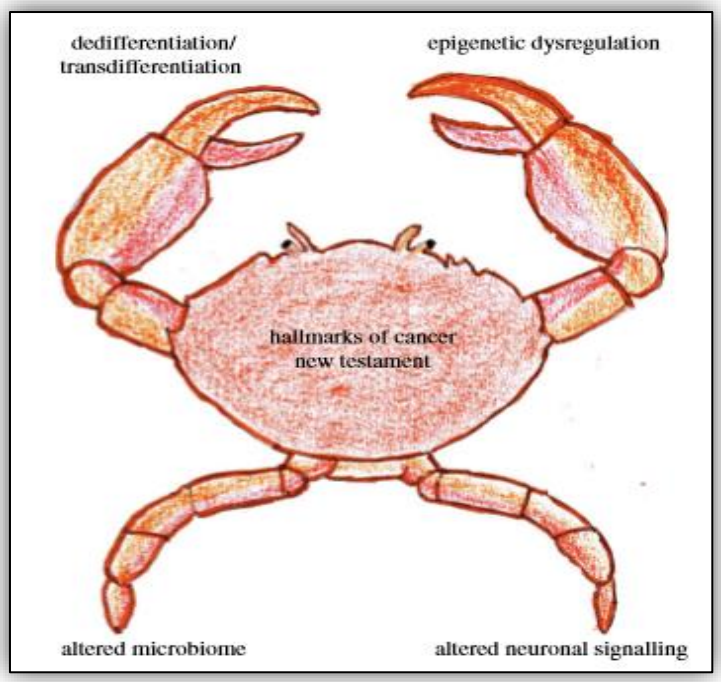

Figure 12: Novel hallmarks of cancer **[4].**

- **Dedifferentiation and transdifferentiation:**

Dedifferentiation and transdifferentiation are two key processes in cancer biology that are proposed as new hallmarks of cancer. Dedifferentiation refers to the ability of non-cancer stem cells to revert to a less specialized, stem cell-like state, allowing them to acquire stem cell-like features. This process enables cells to regain pluripotency and plasticity, contributing to tumor heterogeneity and progression. On the other hand, transdifferentiation involves the conversion of one differentiated cell type into another, bypassing the stem cell state. In cancer, transdifferentiation can lead to the emergence of different cell lineages within the tumor, contributing to its complexity and adaptability **[4].**

- **Epigenetic dysregulation:**

Refers to the abnormal changes in gene expression patterns that occur without alterations in the underlying DNA sequence. This hallmark involves modifications to the structure of DNA and histones, such as DNA methylation and histone modifications, which can lead to changes in gene activity and cellular function. Epigenetic dysregulation plays a crucial role in cancer development by influencing key processes like cell differentiation, proliferation, and metastasis. It can lead to the activation of oncogenes or the silencing of tumor suppressor genes, contributing to tumorigenesis and progression **[4].**

- **The altered microbiome:**

Refers to changes in the composition and function of microbial communities within the human body, particularly the gut microbiota. These changes can influence cancer development and progression by interacting with the host immune system, affecting metabolism, and modulating the tumor microenvironment. Dysbiosis in the microbiome has been associated with various cancers and can impact the efficacy of cancer treatments **[4].**

- **Altered neuronal signaling:**

Is an enabling hallmark of cancer that provides tumors with a means of interacting with their microenvironment to facilitate metastatic progression. Tumors can use nerves to establish blood vessels and garner proliferative cues, and cancer cells recruit numerous nerves, which can be intercepted for pain management. The modulation of GluN2B expression has shown that NMDAR signaling is critical for the proliferation of breast cancer cells in the brain **[4].**

I-6- Understanding metastasis:

a- Cancer metastasis :

Metastatic cancer is mostly incurable, resulting in more than 90% of cancer-related deaths. Patients with metastatic cancer have much lower 5-year survival rates than those with localized cancer **[82].**

Metastasis is an organ-selective process that is started by escape of tumor cells from the primary tumor and ended with colonizing secondary tumors in the distant sites **[83].** This process involves several key steps that disseminated tumor cells (DTCs) must overcome to colonize distant organs. These steps include the detachment of tumor cells from the primary site, invasion into surrounding tissues, intravasation into blood or lymphatic vessels, circulation, stasis, extravasation into a new tissue, and ultimately, proliferation **[84].**

The initiation of metastasis is facilitated by cancer cells that migrate using extracellular components and metastatic traits conferred through epigenetic regulation. The interface between the tumor and stromal cells plays a crucial role in this migration. Interactions between these cells promote the progression of metastasis **[85][86]**

Metastasis can occur through various pathways, including hematogenous and lymphogenous routes, or by seeding into body cavities **[87].**

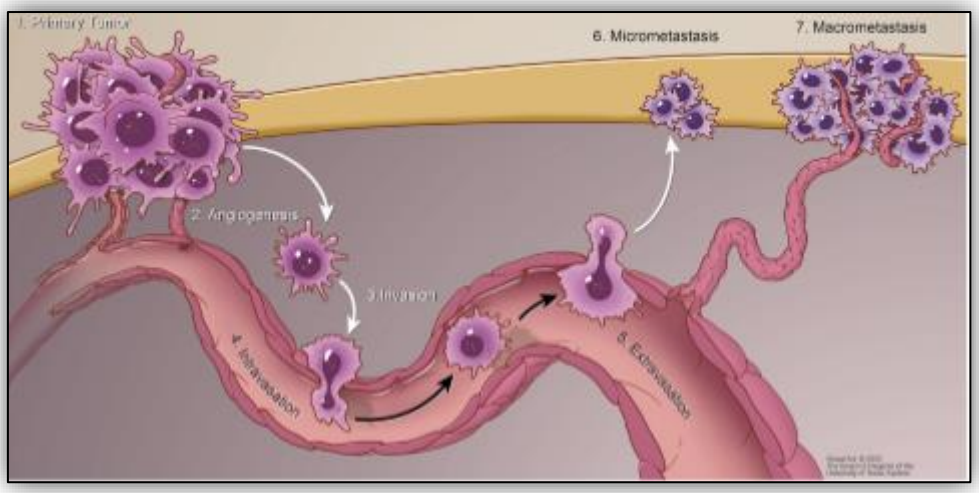

Figure 13: Progression of cancer metastasis. Illustration of the stages of progression from primary tumor formation to the establishment of a metastatic tumor **[82].**

b- Steps of metastasis:

Metastasis, the complex process of cancer spread, involves several distinct steps. These steps include local invasion of tumor cells, intravasation into the vasculature, survival in circulation as circulating tumor cells (CTCs), extravasation into secondary organs, and colonization leading to metastatic outgrowth. The process also encompasses interactions with stromal cells, adaptation to new environments, evasion of immune responses, and establishment of a metastatic niche. Different types of cell movements, such as collective, mesenchymal, and amoeboid migration, play crucial roles in metastasis **[84] [88] [89].** Additionally, factors like epithelial-mesenchymal transition (EMT), genetic changes facilitating migration, and establishment of a vascular network within secondary sites are essential components of the metastatic cascade **[85] [90].**

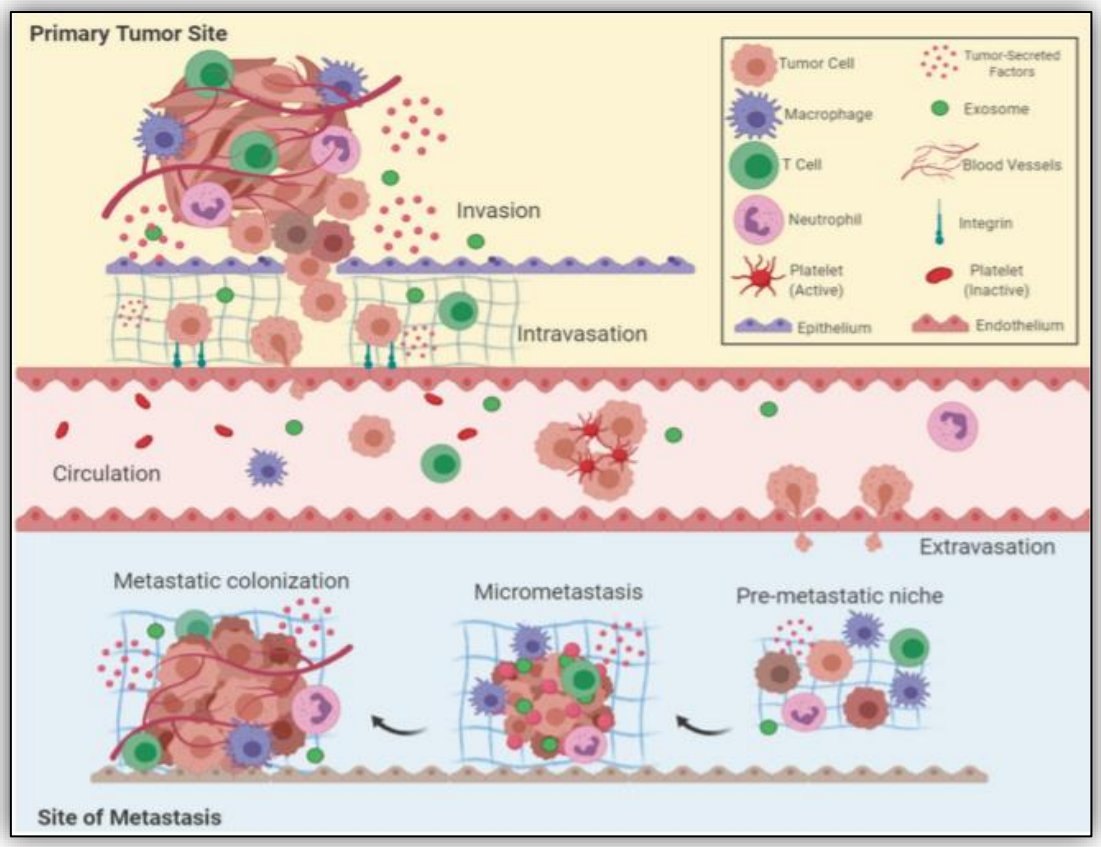

Figure 14: Overview of the metastatic cascade: The five key steps of metastasis include invasion, intravasation, circulation, extravasation, and colonization **[91].**

b-1- Invasion:

Tumor cell invasion refers to the directed migration of cancer cells into surrounding tissues, marking the initial step towards metastasis **[92].** This invasive process involves complex mechanisms such as alterations in the extracellular matrix through the secretion of matrix-degrading enzymes, excessive cell proliferation, and migration **[93].** Cancer cells exhibit different migration patterns, including collective and individual cell migration, each with distinct morphological and molecular genetic characteristics **[94].** The invasion of tumor cells is facilitated by factors like epithelial-mesenchymal transitions, cell adhesion molecules, and actin cytoskeleton proteins **[95].** Circulating tumor cells play a crucial role in invasion and metastasis, relying on elements like epithelial-mesenchymal transition, cancer stem cell properties, and selective metastasis **[96].**

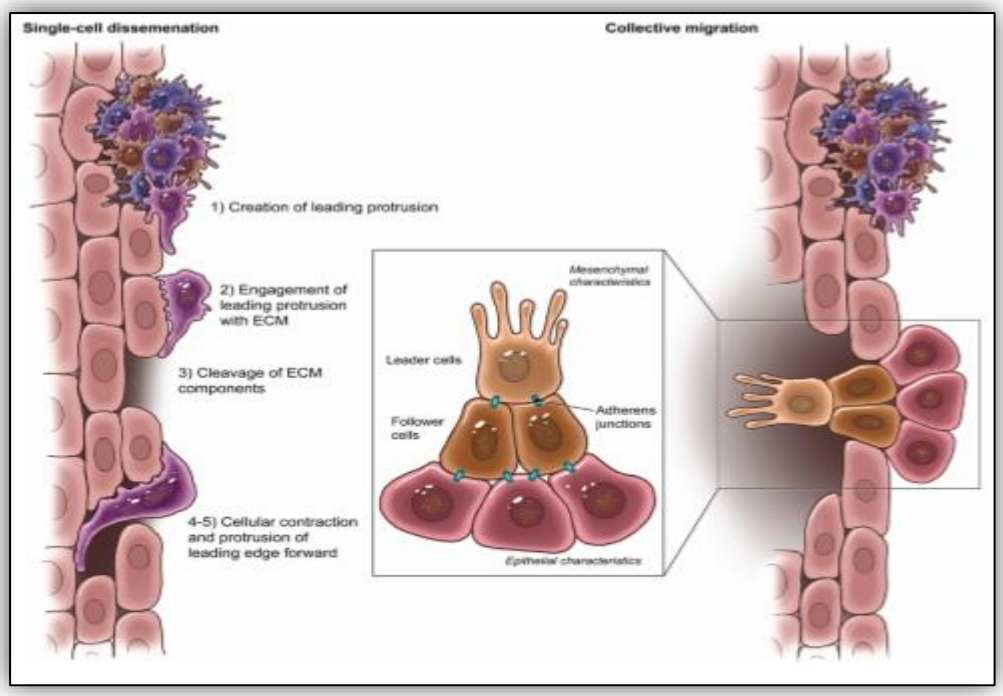

Figure 15: Types of invasion during cancer progression **[82].**

b-2- Intravasation:

It is the process where invasive tumor cells penetrate the vessel wall, entering the circulation as circulating tumor cells, and potentially seeding metastases. This process plays a crucial role in the progression of cancer metastasis by enabling tumor cells to migrate across the endothelium **[97] [98].** Studies have shown that tumor cells disrupt the vessel endothelium through cell division, facilitating their detachment into circulation, a process mediated by mitosis **[99].** Chemo-mechanical signaling between tumor cells and endothelial cells is essential during intravasation, triggering changes in cell morphology and behavior necessary to breach the vessel wall **[99].** Additionally, the tumor microenvironment plays a significant role in intravasation, with factors like macrophages influencing the process through paracrine signaling pathways and TMEM-mediated mechanisms **[98].**

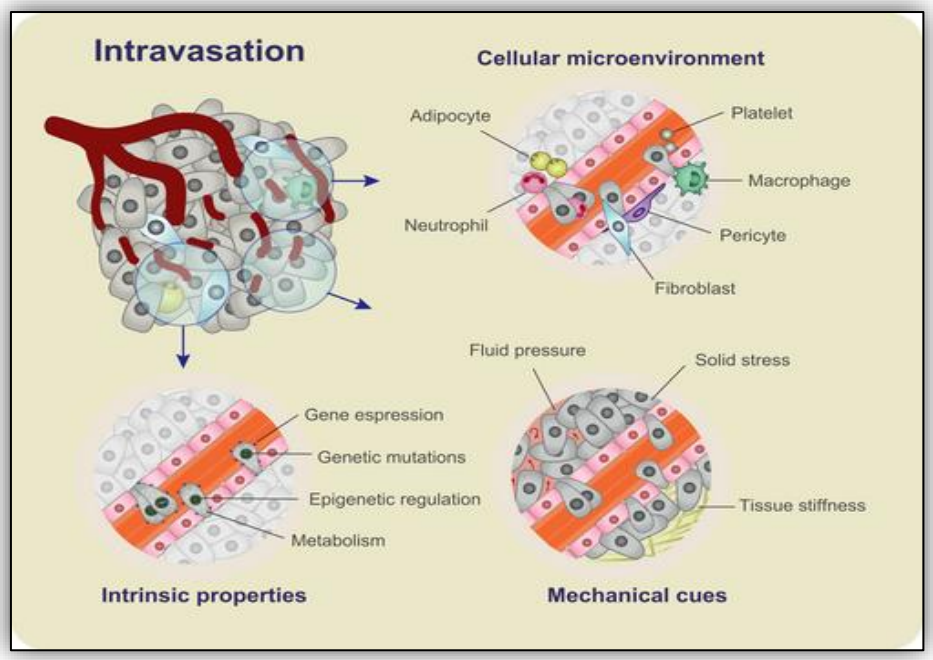

Figure 16: The intravasation process is regulated by intrinsic, microenvironmental, and mechanical factors. **[101].**

b-3- Circulation:

CTCs and CTMs survive in circulation during metastasis through various mechanisms. CTCs adapt to biomechanical constriction forces in the microcirculation, interact with blood components like immune cells and platelets, and undergo molecular adaptations to promote metastasis formation **[102]**. CX3CR1, a chemokine receptor, plays a crucial role in CTC reseeding to multiple organs, affecting tumor growth and numerical expansion; targeting this receptor can prolong CTC permanence in the blood, leading to apoptosis and counteracting metastatic reseeding **[103].** CTC survival is influenced by interactions with the physical environment, intrinsic biophysical characteristics, and the heterogeneous nature of CTCs due to intratumoral heterogeneity and molecular plasticity **[104].** Additionally, CTCs must withstand shear stress, avoid coagulation, and survive immune attacks to successfully form distant metastases **[105].** Cell-cell interactions within CTMs enhance metastatic properties, including improved cell survival, immune evasion, and effective extravasation into distant organs **[106].**

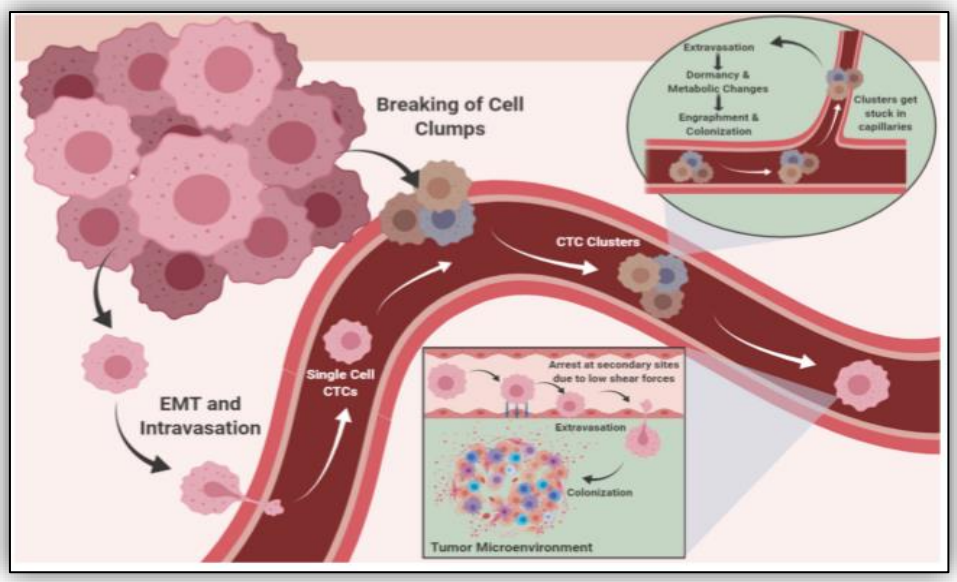

Figure 17: Cancer cells circulate as single units or in clusters **[91].**

b-4- Extravasation:

It is the step where cancer cells exit the bloodstream at distant sites to establish new colonies **[107] [108].** This process involves cancer cells adhering to vascular endothelial cells, crossing vessel walls, and interacting with circulating platelets, leukocytes, and the local tissue microenvironment **[109].** Furthermore, the formation of pre-metastatic niches through extracellular vesicles (EVs) has been identified as a key step in creating a favorable environment for metastatic cell engraftment and growth **[110].**

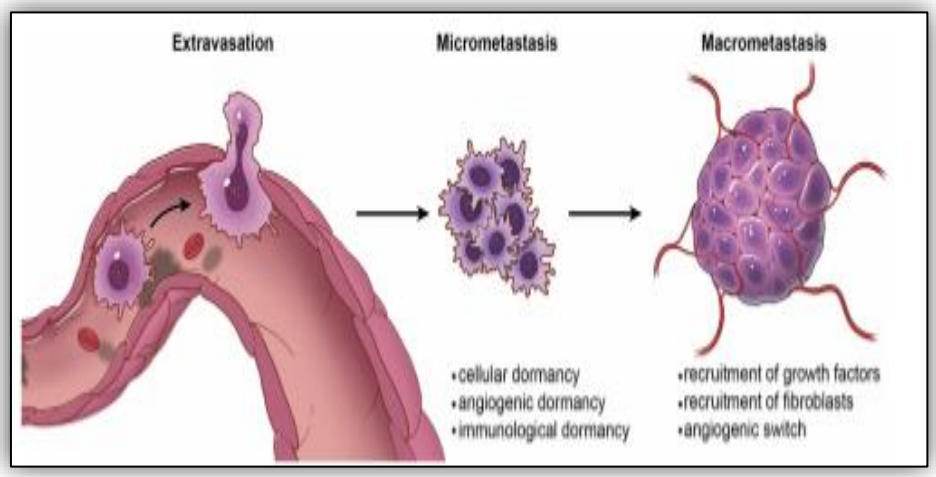

Figure 18: Extravasation to micro- and macrometastases **[82].**

b-5- Colonization:

Metastatic colonization is a critical step in the metastatic cascade, where cancer cells must adapt to the new environment by reprogramming their metabolic states. Studies have shown that successful colonization involves interactions with resident cells in the target organ, such as hepatocytes in the liver **[111] [112].**

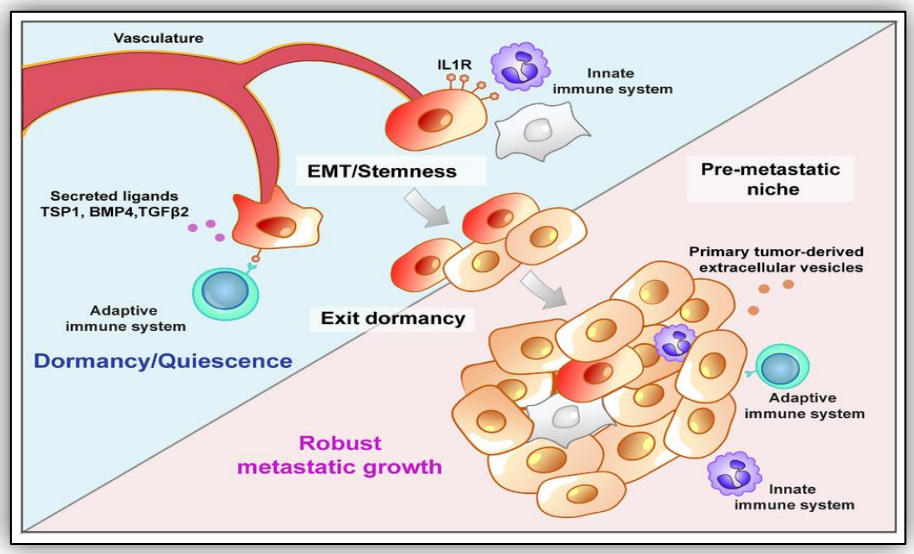

Figure 19: Metastatic colonization **[113].**

II- CTCs/CTMs in metastasis:

Circulating tumor cells (CTCs) are single cancer cells that detach from a solid tumor lesion and enter the bloodstream, leading to the spread of cancer to distant organs. Despite their rarity in peripheral blood, CTCs are crucial for disease progression as they contain a population of metastatic precursors.

CTCs face numerous challenges once they leave the tumor bulk and enter blood vessels. These challenges include mechanical stress, a foreign microenvironment, and immunosurveillance. These factors lead to a short half-life and a high rate of apoptosis and anoikis.

In addition to CTCs, there are other tumor derivatives in circulation. Among these, circulating tumor microemboli (CTMs) are of particular interest. CTMs are clusters of tumor cells or a mix of tumor and non-tumor cells. They can be classified into homotypic CTMs

(clusters of only cancer cells) and heterotypic CTMs (clusters of cancer and non-cancer cells such as immune cells, platelets, or stromal cells).

Although CTMs are less frequent than single CTCs, they have a significantly higher metastatic potential. It has been estimated that the metastatic potential of CTMs is 23 to 50 times greater than that of single CTCs. Furthermore, the presence and size of CTC clusters are associated with worse clinical outcomes in multiple cancer types. This highlights the importance of studying both CTCs and CTMs in cancer research **[114] [115] [116].**

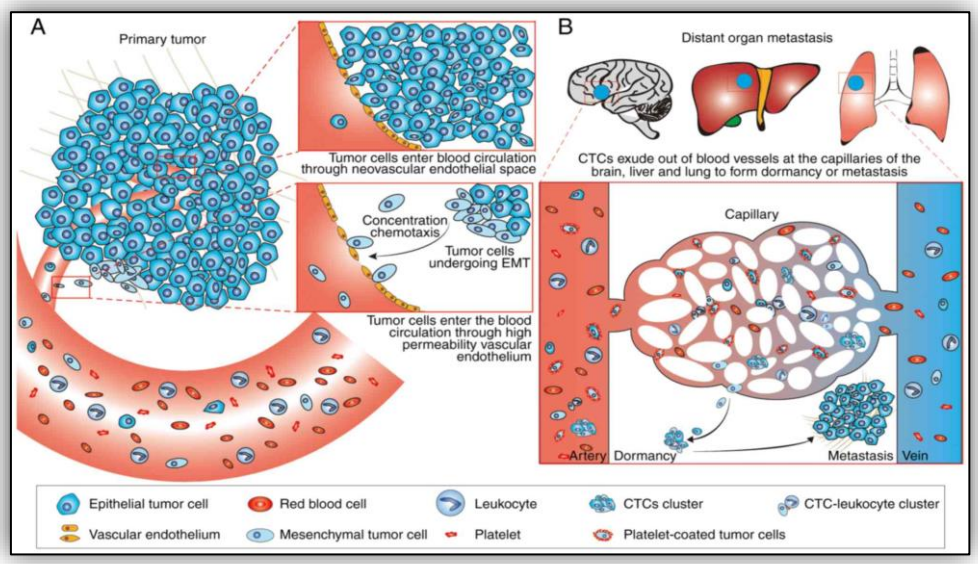

Figure 20: Biology of CTCs **[117].**

II-1- CTCs/CTMS isolation methods:

Various methods have been developed for the isolation of circulating tumor cells (CTCs), each with its own advantages and limitations.

II-1-1- CTC-iChip:

The CTC-iChip is a microfluidic device that uses a combination of techniques, including deterministic lateral displacement, inertial focusing, and magnetophoresis, to isolate and enrich circulating tumor cells (CTCs) from blood samples **[118].** This technology has been further developed to include single-cell immunoblotting, allowing for direct protein analysis of CTCs **[119].** The CTC-iChip has been used to successfully isolate CTCs from patients with various types of cancer, including breast, prostate, lung, and melanoma **[120].** The ability to isolate viable CTCs using the CTC-iChip has significant implications for

personalized cancer therapy, as it allows for the genetic and functional characterization of these cells, which can then be used to guide treatment decisions **[121].**

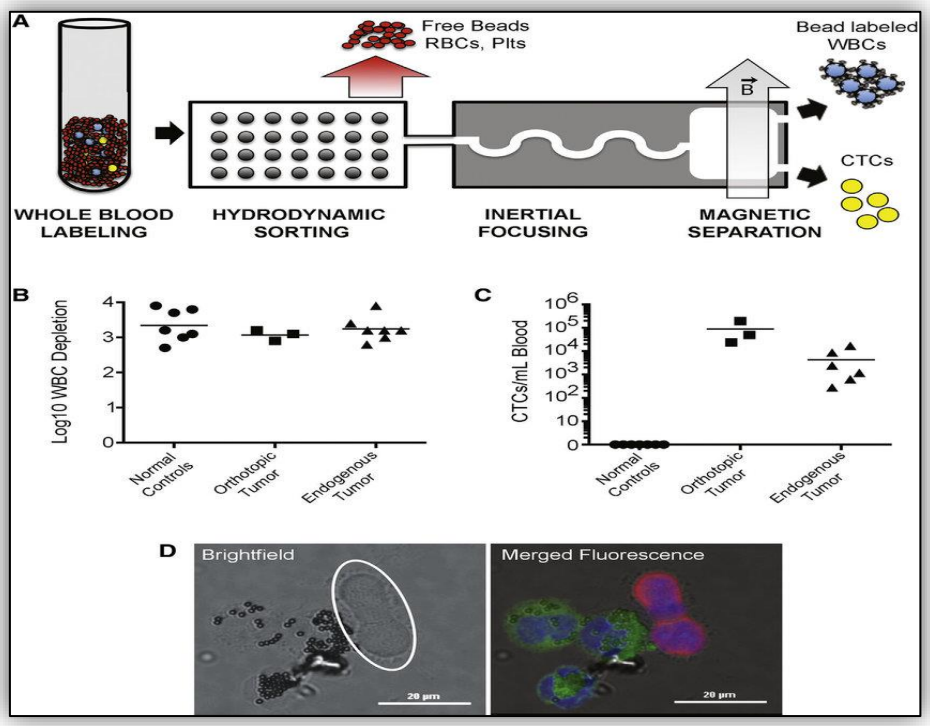

Figure 21: CTC Single-Cell Isolation **[123].**

II-1-2- The advantages and limitations:

The CTC-iChip technique offers advantages such as high efficiency in isolating circulating tumor cells (CTCs) from blood samples, enabling genetic analysis, cancer progression prediction, drug development, and treatment evaluation **[123].** However, challenges persist in CTC isolation due to their rarity and heterogeneity, hindering wider clinical use. Microfluidic-based technologies like the CTC-iChip aim to address these limitations by balancing recovery and purity of isolated CTCs **[124].** Despite advancements, current techniques still face issues with CTC damage during isolation, emphasizing the need for further improvements in maintaining CTC viability for downstream applications like drug screening and modeling metastatic processes **[125].**

a- Micromanipulation:

Micromanipulation, a technique used in the isolation of circulating tumor cells (CTCs), involves the use of microfluidic technology, which offers high precision and sensitivity **[126].** This technique can also be used to isolate platelet-covered CTCs, a

challenging task due to the masking of surface markers, by depleting unbound platelets and then capturing the CTCs using a herringbone micromixing device **[127].** Another approach to CTC isolation is the use of lithographic microfilters, which can be systematically studied and optimized for CTC capture **[128].** Additionally, an immunomicromachine-based approach has been proposed for the direct detection of CTCs without sample pre-processing **[129].**

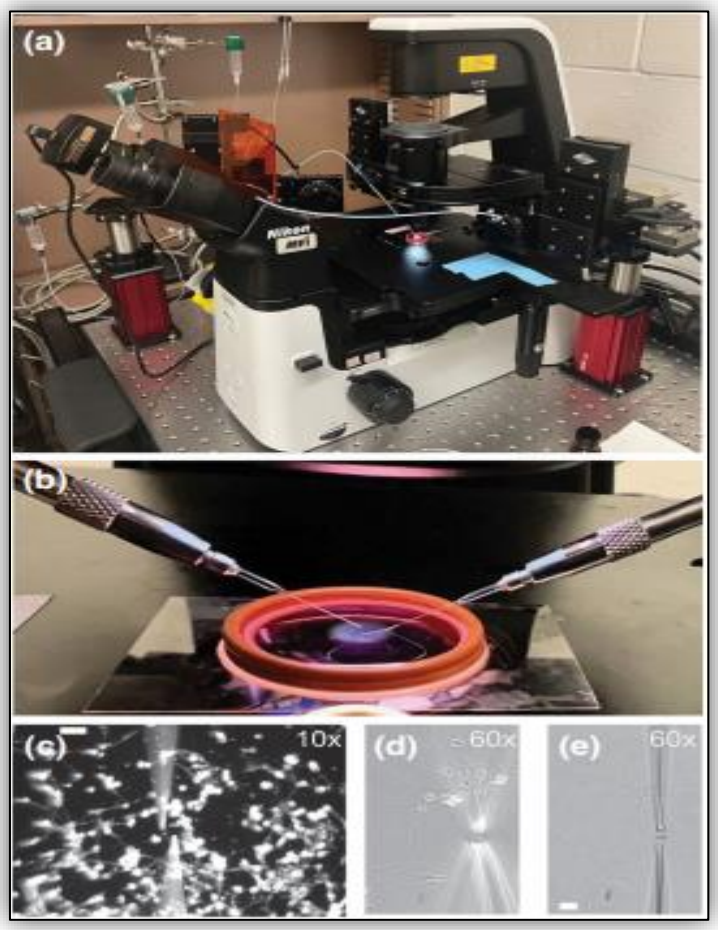

Figure22: Micromanipulation apparatus setup is adaptable to any inverted microscope **[130].**

- **Advantages and limitations:**

Micromanipulation techniques, particularly those utilizing microfluidics, offer several advantages in the isolation of circulating tumor cells (CTCs). These techniques are highly sensitive, allowing for the detection of rare CTCs, and can be automated for increased precision and accuracy **[126].** They also enable label-free isolation, preserving the phenotypic and genotypic characteristics of the isolated cells. However, these techniques may have limitations in terms of throughput and purity, with biology-based techniques often exhibiting

high purity but low throughput **[127].** Despite these limitations, microfluidics technology continues to advance, offering potential for improved CTC isolation and analysis **[131].**

b- Immunofluorescence:

Immunofluorescence is a technique used in the isolation of circulating tumor cells (CTCs) from blood samples. This method involves the use of immunomagnetic enrichment and fluorescence-activated cell sorting (IE/FACS) to isolate CTCs with little to no contamination of normal blood cells **[132].** The technique has been further advanced with the integration of microfluidic chips, which enable the isolation and subsequent biological analysis of CTCs **[133].** The use of ultra-high throughput microfluidic Vortex technology has also been shown to improve the capture efficiency and purity of CTCs **[134].** Additionally, the CTC-iChip technology, which is tumor antigen-independent, has been developed for the isolation of rare CTCs from blood samples **[118].**

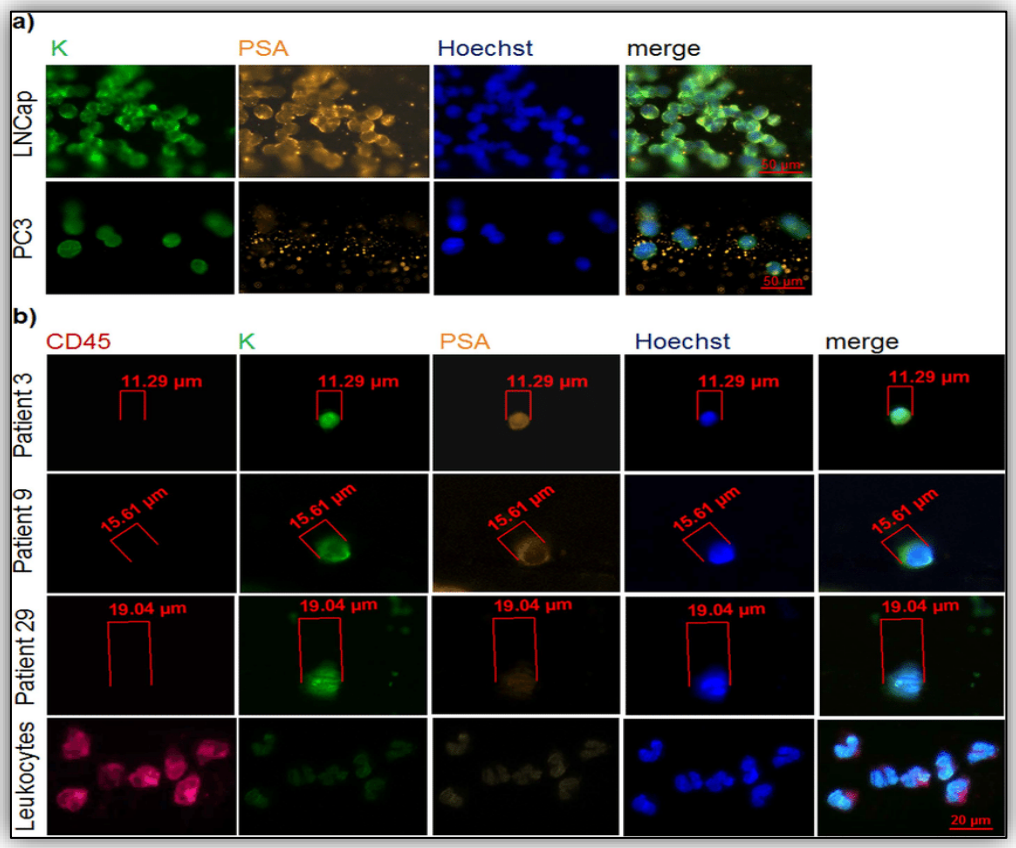

Figure 23: Immunofluorescence staining of established tumor cell lines or CTCs captured by the CellCollector **[135].**

❖ **Advantages and limitations:**

Immunofluorescence has both advantages and limitations. On the positive side, it allows for the identification and characterization of CTCs, which can provide valuable information for cancer prognosis and treatment **[136].** However, the technique is complex and can result in low yield and purity, making it less efficient for CTC isolation **[136] [137].** Despite these limitations, immunofluorescence remains a valuable tool in the study of CTCs, particularly when combined with other isolation methods such as immunomagnetic enrichment and fluorescence-activated cell sorting **[132].**

c- Microfluidics:

Microfluidics, a cutting-edge technology, has revolutionized the isolation of circulating tumor cells (CTCs) from blood samples. Ajanth and Descamps both highlight the potential of microfluidic devices in CTC isolation, emphasizing their label-free nature and ability to balance recovery and purity **[119] [131].** Macaraniag further underscores the role of microfluidics in studying CTC clusters, which are associated with poor prognosis and high metastatic potential **[132].** The use of 3D printed microfluidic devices, as demonstrated by Chen, has significantly improved the capture efficiency of CTCs, paving the way for potential clinical applications **[133].**

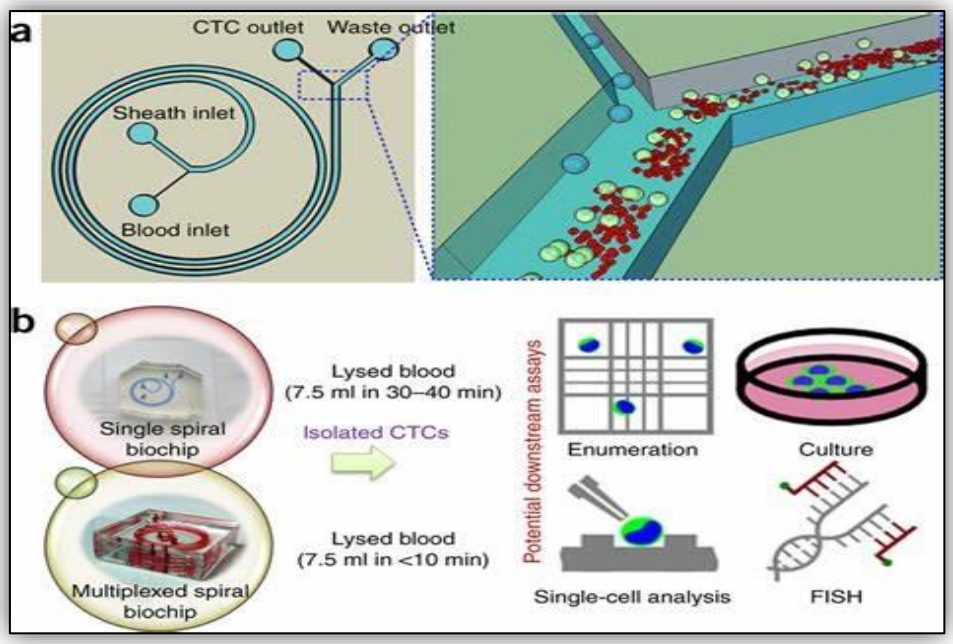

Figure 24: Overview of CTC isolation using spiral microfluidics **[141].**

❖ **Advantages and limitations:**

Microfluidics, a promising technology for isolating circulating tumor cells (CTCs), offers several advantages. It is cost-effective, has short reaction times, high throughput, and is easy to use **[142].** The technology also allows for the preservation of the phenotypic and genotypic characteristics of isolated cells, making it a valuable tool for cancer diagnosis and treatment **[126].** Furthermore, microfluidic devices can be integrated with nanotechnologies and enable in situ or sequential analysis of captured CTCs, enhancing their potential for clinical impact **[143].** However, there are limitations to microfluidics, including the need for further research to improve capture efficiency, purity, and sensitivity **[126].**

d- Immunomagnetic:

Immunomagnetic separation, a technique used to isolate circulating tumor cells (CTCs), has been significantly advanced in recent years. Magbanua developed a protocol combining immunomagnetic enrichment and fluorescence-activated cell sorting (IE/FACS) to isolate highly pure CTCs for molecular profiling **[132].** This was further improved by Chen who integrated advanced technologies with microfluidic chips, enhancing the isolation and subsequent analysis of CTCs **[133].** Sun proposed a strategy for isolating tumor initiating cells (TICs) using immunomagnetic separation, which involved screening two surface markers **[144].** Tang applied a chip-assisted immunomagnetic separation system to efficiently capture and identify CTCs, demonstrating its potential for clinical use **[145].**

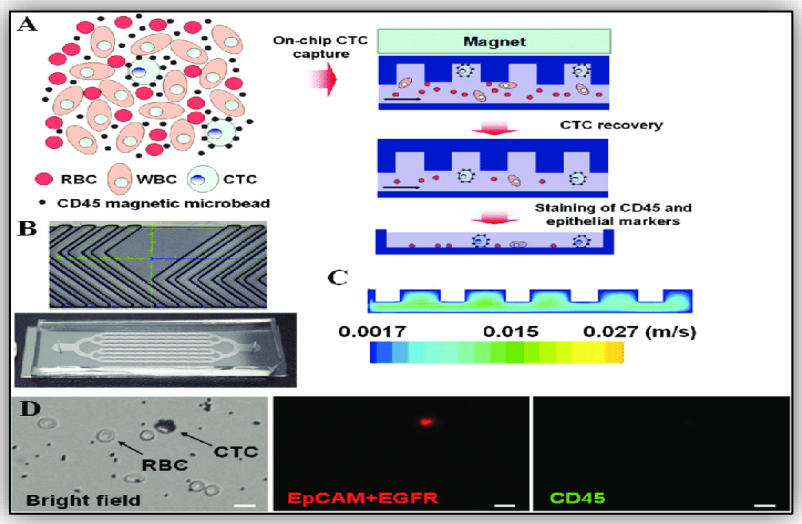

Figure 25: Microfluidics-based immunomagentic isolation of CTCs from whole blood of lung adenocarcinoma patients **[146].**

❖ **Advantages and limitations:**

Immunomagnetic isolation of circulating tumor cells (CTCs) offers several advantages, including high efficiency capture and release of tumor cells from whole blood **[147],** enhanced isolation efficiency in spiked blood samples and robustness in downstream cell sequencing **[148],** and efficient capture and in situ identification of CTCs with negligible influence on cell viability **[145].** Furthermore, the combination of immunomagnetic enrichment and fluorescence-activated cell sorting (IE/FACS) allows for the isolation of highly pure CTCs for molecular profiling **[126].** However, this technique also has limitations, such as nonspecific adsorption of biomolecules by immunomagnetic nanoparticles **[148]** and the need for expensive instrumentation for enrichment and characterization **[147].**

e- Immunomagnetic purification:

Immunomagnetic purification, a technique used in the isolation of circulating tumor cells (CTCs), involves the use of magnetic beads conjugated to monoclonal antibodies against specific cell markers, such as epithelial cell adhesion marker (EpCAM), to enrich for tumor cells. This method has been shown to produce highly pure CTCs, making it a valuable tool for downstream molecular analyses **[132].** The technique has also been applied to the purification of T cell subpopulations, resulting in high purity and reproducibility **[149].** A CTC microseparator based on lateral magnetophoresis and immunomagnetic nanobeads has been developed, demonstrating high isolation efficiency and purity **[150].** The use of immunomagnetic separation in microfluidic chips has further advanced the isolation of CTCs, with potential applications in cancer research and diagnosis **[140].**

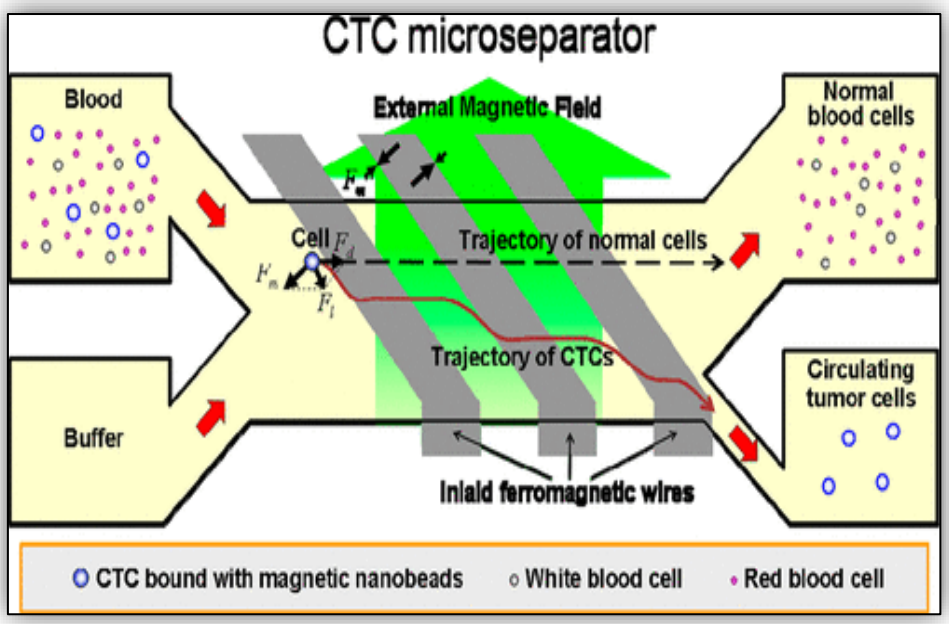

Figure 26: Circulating tumor cell microseparator based on lateral magnetophoresis and immunomagnetic nanobeads **[150].**

❖ **Advantages and limitations:**

Immunomagnetic purification, a technique used for the isolation of circulating tumor cells (CTCs), offers several advantages. It allows for the efficient capture and in situ identification of CTCs, with high purity and minimal contamination of normal blood cells **[132].** However, this method also has limitations, including time-consuming protocols and low efficiency. Recent advances in microfluidic devices have shown promise in addressing these limitations, offering greater efficiency, sensitivity, selectivity, and accuracy in capturing and isolating CTCs **[151]**. These devices, when integrated with immunomagnetic isolation, have the potential to significantly improve the process of CTC isolation and subsequent biological analysis **[130].**

f- Fluorescence-activated cell sorting (FACS):

FACS is a powerful technique used to isolate specific cell populations based on their fluorescence and size, and is particularly useful when high purity is required **[152] [153].** This technique has been applied in various fields, including biological research and cancer detection **[152] [126].** In the context of cancer, FACS can be used to isolate circulating tumor cells (CTCs) from the bloodstream, which can serve as biomarkers for cancer detection and provide insights into cancer metastasis **[126].**

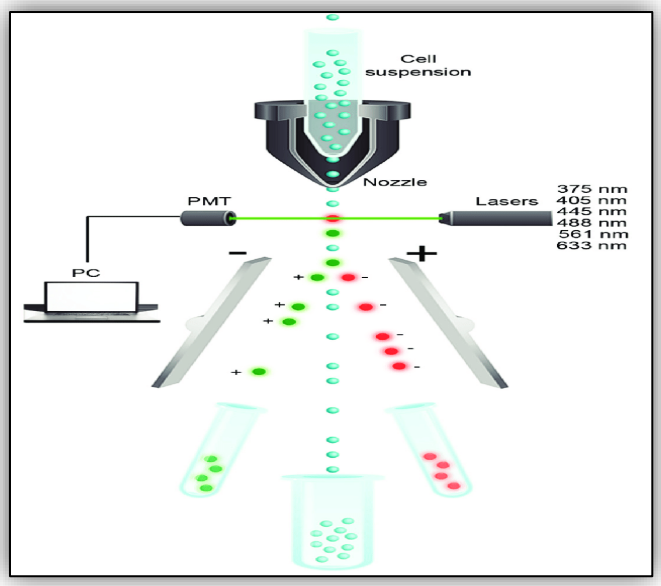

Figure 27: A fluorescence-activated cell sorter (FACS) **[154]**.

❖ **Advantages and limitations:**

FACS, or fluorescence-activated cell sorting, is a powerful technique for isolating circulating tumor cells (CTCs) due to its ability to label and sort cells based on their specific characteristics. However, it has some limitations, including the potential for cell damage during the labeling process and the need for specialized equipment and expertise. Other techniques, such as microfluidic technology, offer advantages such as automation, high precision, and portability, but also have limitations, such as the use of chemicals for labeling that can interfere with downstream assays **[126].** Size-based isolation methods, such as FAST, provide a clog-free, highly sensitive, and rapid isolation of CTCs, but may have lower recovery rates and purity **[150].** An integrated microfluidic device for CTC isolation and single-cell analysis has been developed, which offers high throughput and efficiency with less cell damage **[155].** Despite these advancements, the isolation and detection of CTCs remain challenging due to their rarity and variability, and further research is needed to address these limitations **[131].**

III- RNA, or Ribonucleic Acid:

III-1- Definition:

Is a macromolecule that plays diverse roles in various biological processes. It is a single-stranded nucleic acid that is essential for protein synthesis and gene expression in living organisms **[156].**

RNA structure is composed of ribose sugar based on nucleic acids and primarily four bases: adenine, cytosine, guanine, and uridine. It consists of three parts: a phosphate group, a 5-carbon sugar, and a nitrogenous base. RNA can form primary, secondary, and tertiary structures. The primary structure refers to the linear RNA sequence of nucleotides linked by phosphodiester bonds. The secondary structure is formed by Watson-Crick canonical base pairs, containing helices, loops, bulges, and junctions. The tertiary structure is the three-dimensional arrangement of RNA building blocks with both canonical and noncanonical base pair interactions **[157].**

III-2- ARN types:

ARN types include various forms such as messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and non-coding RNAs like small regulatory RNAs:

- ✓ **Messenger RNA (mRNA):** this type carries genetic information from the DNA in the nucleus to the ribosomes in the cytoplasm for protein synthesis.
- ✓ **Transfer RNA (tRNA):** it Transfers amino acids to the ribosome during protein synthesis.
- ✓ **Ribosomal RNA (rRNA)**: it Forms a major part of ribosomes, where protein synthesis occurs.
- ✓ **Small nuclear RNA (snRNA):** it is Involved in RNA splicing processes.
- ✓ **Small nucleolar RNA (snoRNA):** this RNA type Guides chemical modifications of other RNAs.
- ✓ **MicroRNA (miRNA):** it Regulates gene expression by targeting specific mRNAs for degradation or inhibiting translation.
- ✓ **Long non-coding RNA (lncRNA**): it Regulates gene expression at various levels without being translated into proteins **[158].**

Conclusion:

In this chapter, I explored the multifaceted landscape of cancer, including its historical context, modern understanding, types, genetic aspects, hallmarks, metastasis, and the significance of circulating tumor cells (CTCs) and microemboli (CTMs). I also delved into the role of ribonucleic acid (RNA) in cancer biology. This foundational knowledge forms the basis for our investigation into predictive modeling and machine learning techniques for cancer prediction using gene expression data from CTCs and CTMs.

Chapter II

Artificial Intelligence

Introduction:

Artificial intelligence (AI) is one of the most transformative technologies of the 21st century, making huge impacts across all fields, including healthcare. It has transformed traditional healthcare practices, resulting in more accurate treatment and diagnoses for various diseases, including cancer. This chapter gives a comprehensive overview of artificial intelligence and of basic concepts in both machine learning and deep learning.

1- Definition:

Artificial intelligence (AI) is a branch of computer science that aims to emulate intelligent behavior. It can be defined as the capacity of a computer system or machine to mimic and carry out operations like learning, problem-solving, and logical reasoning that would typically require human intelligence. The foundation of artificial intelligence lies in the application of machine learning algorithms and technologies, which enable machines to apply specific cognitive capacities and carry out tasks autonomously or semi-autonomously. It is the degree of autonomy or cognitive capacity that sets artificial intelligence apart. Its capacity can be superlative, general, weak, or limited. It can be reactive, deliberate, cognitive, or completely autonomous because of its autonomy. Many processes are becoming more efficient as artificial intelligence advances, and tasks that seem difficult now will be completed more quickly and accurately **[159].**

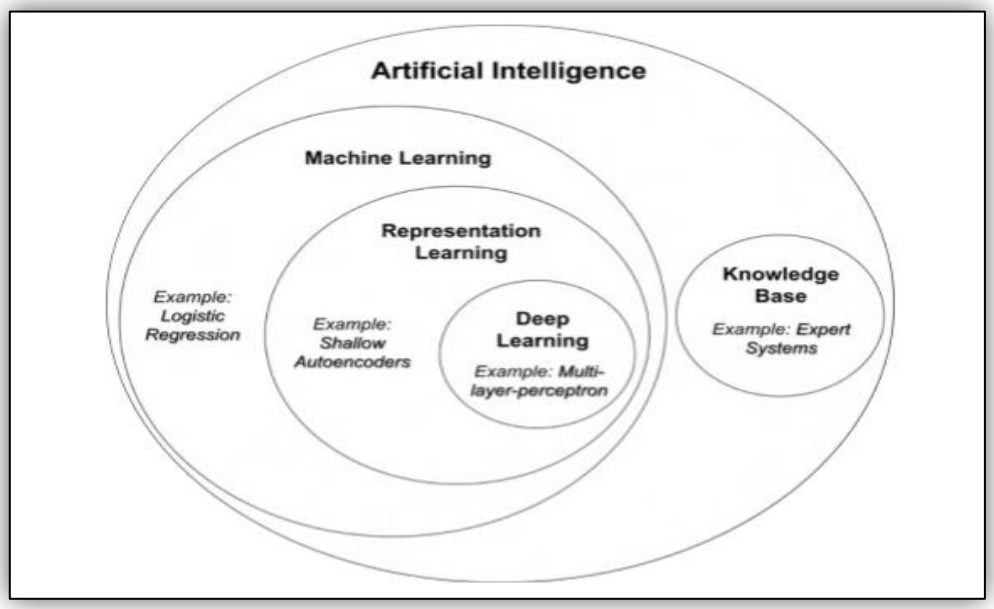

Figure 28: Representation of the different AI disciplines **[160].**

2- Brief History of AI:

Many people are surprised to know that AI is nothing new, AI technologies have existed for decades **[161].** The roots of AI trace back to early computer development, which initially focused on automating calculations. Over time, computers evolved to handle more complex tasks during the Industrial Revolution **[162].**

The dream of AI can be linked to Ada Lovelace's invention of the first programming language in 1820 and Descartes' conceptualization of the world as a giant machine. These foundational ideas set the stage for integrating logic processing and human-like intelligence into computers **[163].**

In 1956, American computer scientist John McCarthy coined the term "artificial intelligence." Around the same time, Frank Rosenblatt introduced the perceptron (the first machine capable of learning) inspired by Hebb's cognitive theory. Although limited by its single layer, the perceptron was revolutionary.

However, after Marvin Minsky and Seymour Papert highlighted the perceptron's limitations, faith in this approach waned. Funding for AI research decreased in 1969, leading to an "AI winter" during the 1970s.

Despite setbacks, progress continued. Kunihiko Fukushima proposed the Cognitron (1975) and Neocognitron (1981), drawing inspiration from the cat's visual cortex. The Neocognitron, with its four layers, alternated simple and complex cells, culminating in a specialized classification layer **[160].**

In 1986, backward error propagation and the Hierarchical Learning Machine (HLM) advanced multilayer artificial neural networks (ANNs). Yann Le Cun's Convolutional Neural Network (CNN) emerged in 1988, mimicking the visual cortex structure **[160].** The 1990s saw advancements in computer architecture, enabling real-time intelligent applications. The Support Vector Machine (SVM) gained prominence between 1992 and 1995 **[164].**

Despite another gloomy period for ANNs, persistence paid off in 2012. Deep learning, fueled by GPUs and large datasets, revolutionized AI. Today, deep learning permeates our daily lives, demonstrating its value across various tasks **[160].**

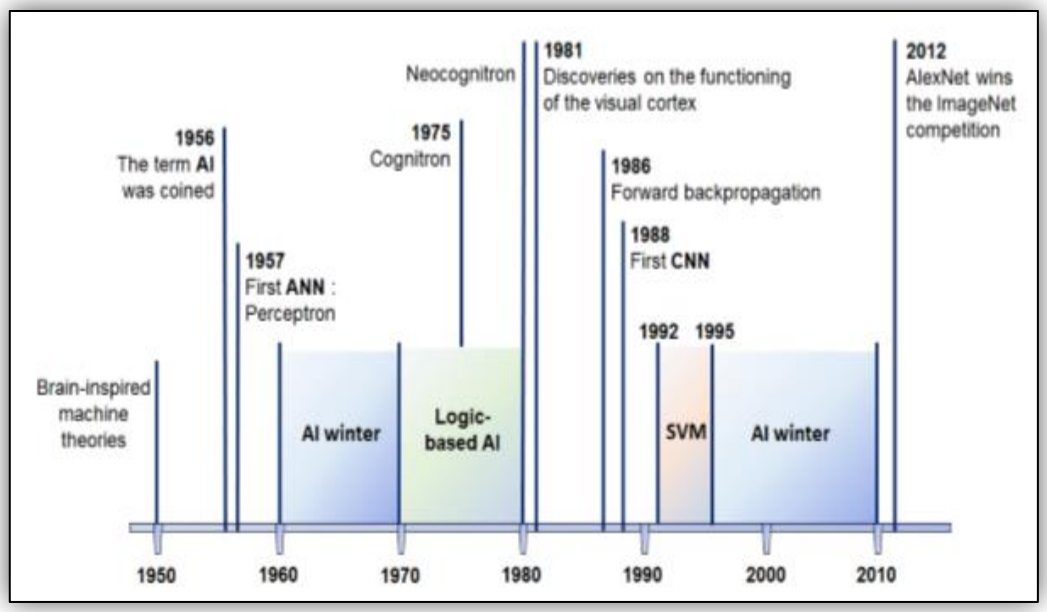

Figure 29: Most significant events in the history of artificial intelligence **[160].**

3- Machine Learning:

Machine learning is a branch of artificial intelligence (AI) that focuses on creating models and algorithms that let computers learn from data and make decisions or predictions without explicit programming. It involves the study of statistical and mathematical models and methods that let computers automatically identify relationships and patterns in vast amounts of data and use those insights to predict or make decisions.

A labeled dataset, which contains input data accompanied by matching desired outputs or labels, is typically used to train a model in a machine learning process. This training set of data helps the model learn by pointing out trends and connections that connect the input data to the intended results.

Machine learning finds wide-ranging applications in diverse fields, such as natural language processing, recommender systems, medical diagnosis, image and speech recognition, and many more. It is an effective tool for handling challenging issues and drawing conclusions from large amounts of information due to its capacity to automatically learn from data and adapt to changing circumstances.

In order to handle increasingly complicated and varied datasets, improve performance, address ethical issues, and improve interpretability, researchers in the field of machine learning are continuously creating new models, techniques, and algorithms **[165].**

3-1- Types of machine learning:

ML algorithms are classified as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning **[166].**

a- Supervised algorithms:

The first category of ML algorithms is called supervised algorithms; each sample in this class has a classer label attached to it, and they require labeled data. The term "labelled data" refers to data items that have a known description attached to them. For instance, test findings for cancer can indicate various symptoms in twenty different persons. We are able to tag or label each patient based on the test results, indicating whether they are positive or negative for cancer. Therefore, the tagged data gives the output a shape. Thus, the machine will identify patterns and categorize data as a result of the supervised learning process. The unseen data can be located using the same techniques. Two types of supervised learning are distinguished: **[160] [166]**

- Classification: classification algorithmsclassify the input data from pre-defined classes, they are used to predict and classify discrete values such as Cat or Dog, Spam or No Spam, etc.
- Regression: Regression algorithmsfind the relationship between the variables/features, they are used to predict continuous values such as the value of a company's stock, house prices, etc. **[166] [167]**

b- Unsupervised Learning:

Unsupervised learning is the second category. In this case, the machine only receives the input values or data; the system does not receive a fixed output. By identifying the hidden pattern in the input data, it predicts the result. An unlabeled input dataset is given to the machine during its unsupervised learning process. An unsupervised learning algorithm uses this data to make independent hypotheses about patterns within the data. The pattern is used to group instances of datapoints. Data that match a similar group and pattern is expected to be the result. It can be used for segmentation, anomaly detection, etc **[166].**

- ✓ **Clustering (Partitional or flat):** This method of unsupervised learning relies on making clusters from the input data. The datapoints that have similarities will make

clusters, and using those clusters, we will be able to make predictions. as do k-means and self-organizing maps.

- ✓ Association (Hierarchical): The second method of unsupervised learning is association, in which the algorithms find the rules from the input data and make prediction from the data. Such as AGNES (AGglomerative NESting) or DIANA (DIvisive ANAlysis Clustering) **[160] [166].**

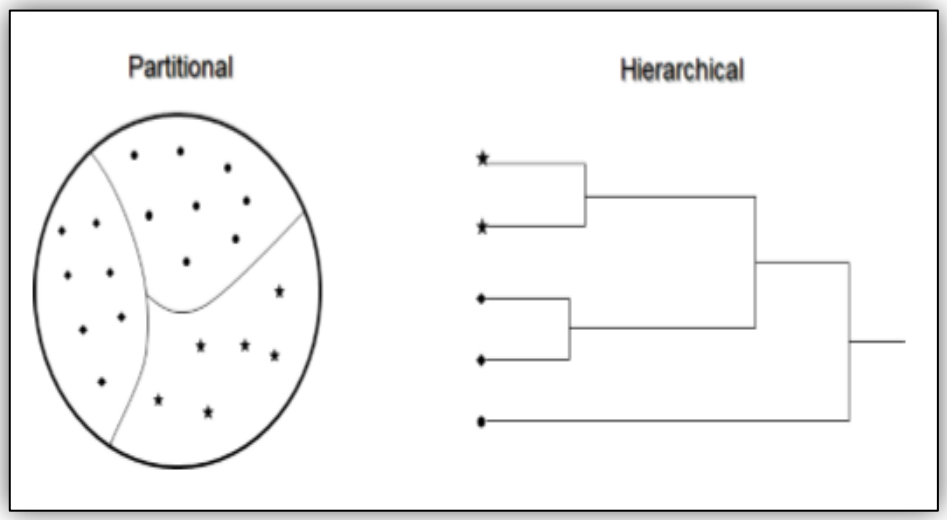

Figure 30: Representation of the two categories of unsupervised learning algorithms **[160].**

c- Semi-Supervised Learning:

This category is a combination of supervised learning and unsupervised learning; it uses both labeled and unlabeled data to build a prediction model. The difference between the above algorithms and the semi-supervised method is that the unlabeled data is larger in number than the labeled data **[166].**

d- Reinforcement Learning:

Reinforcement learning, which focuses on determining the best way to take in a situation in order to maximize the correct outcome, is the last class of machine learning algorithms. Sequential decisions are made. The algorithm may produce a positive or negative result at each stage along the way to the final results. Thus, the total of all of the path's favourable and unfavourable outcomes is the overall result. Finding the optimal path to maximize the outcome is the algorithm's main objective. Generally, reinforcement learning is used to solve various challenges, including video games, telecommunications, and robot control **[166] [167].**

3-2- Machine learning algorithms:

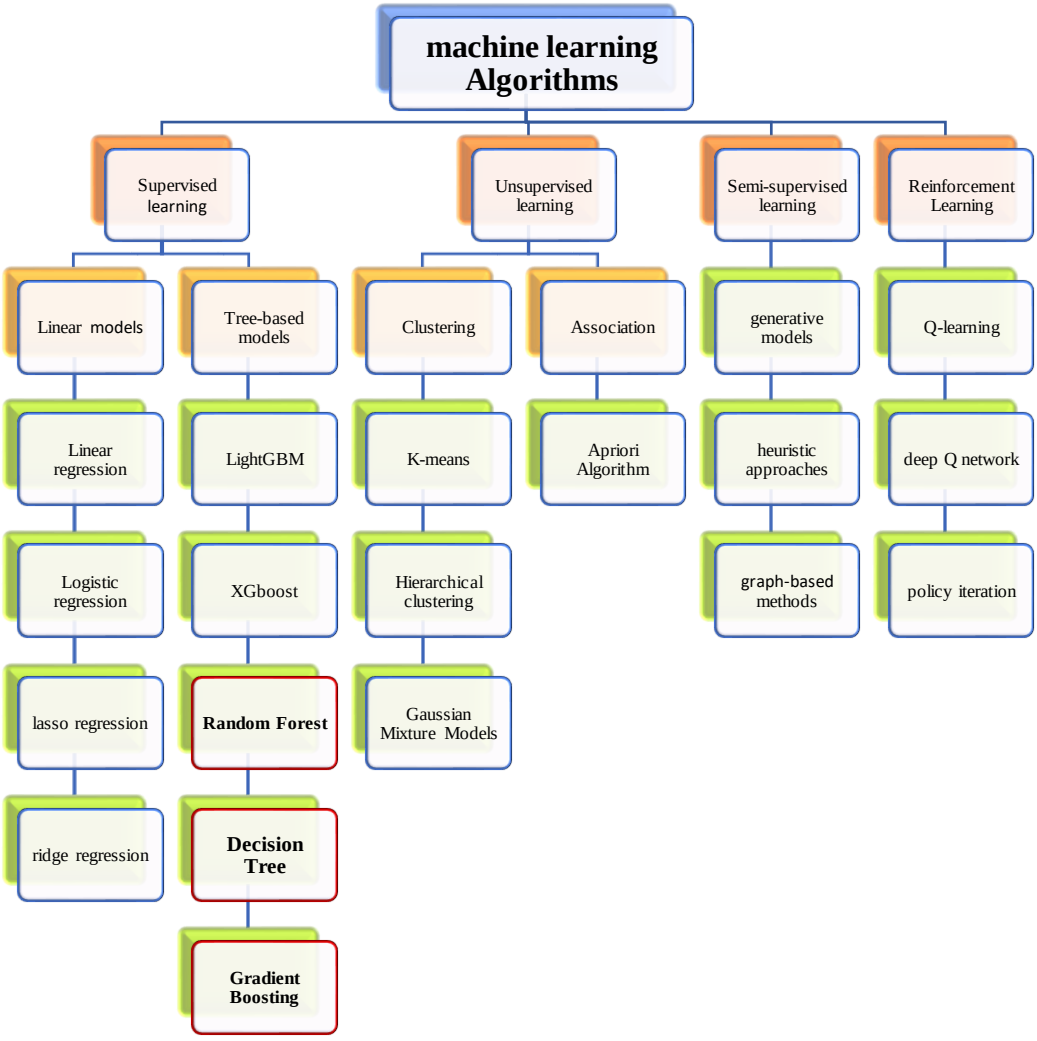

Figure 31: Machine Learning Algorithms **[168].**

Considering the effectiveness of the supervised algorithms such us Decision Tree, Random Forest and gradient boosting in various fields, including genomics, this thesis will focus only on these three machine learning models.

a- Decision Tree:

A decision tree classifier is a popular data modeling technique used in various real-world supervised learning problems, known for its ease of implementation, understanding, and ability to handle non-linear relationships effectively. It operates as a top-down recursive algorithm, dividing datasets until all instances within a sub-dataset belong to the same class value.

Decision trees are suitable for both classification and regression tasks and traditionally use metrics like information gain, gain ratio, and Gini values to select the best node for splitting **[169] [170].**

Decision tree learners are not always the most competitive learners in terms of accuracy because of their high adaptivity to the learning sample. A small perturbation of the learning sample may result in very different decision trees.There are several existing approaches to improving the accuracy of decision trees by reducing their variance, such as pruning, which removes some of their test nodes so as to find the best compromise between bias and variance. However, a more efficient way to improve the accuracy of decision trees is to aggregate the predictions given by several trees. Various techniques have been proposed in the literature to generate different decision trees from the same learning sample for aggregation, and they often give very impressive improvements in accuracy **[171].**

b- Random Forest:

Random Forest (RF) is a powerful machine learning method widely used for classification and regression tasks due to its high predictive accuracy, low variance, and ease of optimization **[172] [173].** The original RF algorithm was invented by Breimann in 2001 and is based on the assumption that a collection of weak classifiers outperforms a single weak classifier, namely a weak decision tree. Decision trees are very attractive classifiers; however, they suffer from the “high variance” problem, meaning they risk overfitting the training data. Breimann solved this problem by combining baggingwith random variable selection at each node. Bagging stands for bootstrap aggregating, meaning each individual decision tree in the forest will learn a different classification model based on a bootstrap sample of the original training set. One bootstrap sample has approximately 63% of the original training data points, sampled with replacement, while the remaining 37% form the out-of-bag data. Each time a tree is added to the forest, the out-of-bag data is used as internal validation data for estimating classification error and variable importance. For classification tasks, each tree will learn a model only on $\sqrt{d}$ variables, where d is the dimension of any data point **[174].**

c- Gradient Boosting:

Gradient Boosting Machine is one of the most successful algorithms in supervised learning. This algorithm was developed by Friedman in 2001 for both problems in classification and regression. Gradient boosting is applied to combine different weak learners into strong ones so that each new model is created in a way that it is trained to minimize the

loss function of the previous model using gradient descent. On every iteration, the algorithm computes a gradient of the loss function relative to the predictions of the current ensemble and trains a new weak model to minimize this gradient. Then it adds the made predictions of this new model to the ensemble and repeats the process until the stopping criterion satisfies. This is one of the most successful ideas in Machine Learning, achieving very high practical performance with very little tuning **[175] [176] [177].**

3-3- Deep learning:

Deep learning (DL) is a subfield of machine learning that is fully supported by artificial neural networks. Since neural networks are designed to mimic the functioning of the human brain, deep learning is also a type of brain mimicking.

The methodology of deep learning involves the application of nonlinear transformations and high-level abstractions, enabling the system to learn from large datasets and continually improve its predictions and decisions.

Compared to traditional machine learning techniques, deep learning has shown remarkable success in a number of fields, including computer vision and linguistic communication processes, as it attempts to form a stronger analysis and can potentially learn enormous amounts of unlabeled knowledge **[178].**

4- Artificial Neural Networks:

Artificial neural networks (ANNs) are mathematical functions inspired by the biological neural networks that constitute animal brains. They are made up of interconnected processing elements, or "neurons," which process information based on their dynamic state response to external inputs **[167] [179].**

In the human brain, there are tens of billions of neurons, and each neuron can have thousands of billions of interconnections. A human neuron is composed of an axon that acts as a cable to transmit and receive information, a dendrite that receives information, a soma that processes information, and a synapse that joins the axon to the dendrite of another neuron. Artificial neurons in ANNs function similarly, where the soma becomes a node, the dendrites serve as input, the synapses act as weights, and the axons carry the output **[179].**

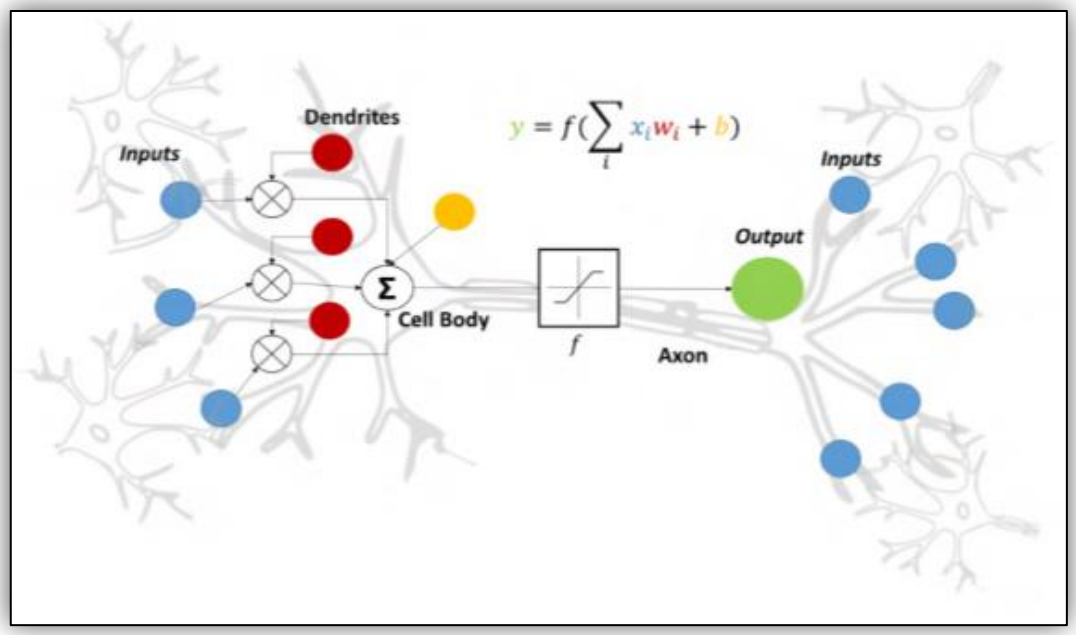

Figure 32: Comparison between a biological and an artificial neuron **[160].**

4-1- Feedforward neural network:

A feedforward neural network is a type of artificial neural network (ANN) where the connections between nodes do not form cycles, meaning the input signal (data) moves in one direction, from input nodes through hidden nodes to output nodes, without any feedback loops. These networks are designed to learn patterns in various types of data, such as structured, textual, speech, or visual data **[180] [181] [182].**

a- The perceptron:

The perceptron is a single-layer neural network and it represents the simplest and most basic form of a feedforward neural network. It is a mathematical model of a biological neuron used for supervised learning of binary classifiers, limited to linear decision boundaries.
the perceptron consists of four main parts : input values x, weights w, bias w_0, net sum, and an activation function f(detailed in Section 2.4.2.3) **[184] [185] [186].**

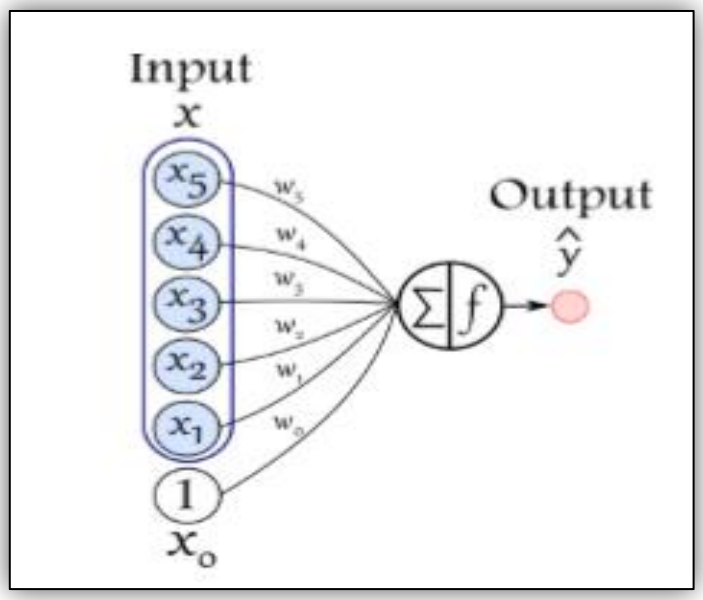

Figure 33: The perceptron network **[184].**

This is how the output value $\hat{y}$is generated for an input of I dimensions:

$$\hat{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{f}(\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{I}} \boldsymbol{w_i x_i}) \textbf{(2.1)}$$

The mapping function between the input label x and the target label $\hat{y}$ is learned by the Perceptron. It first calculates the weighted sum of its input vectors, $x$and w, and then outputs a scalar value, $\hat{y}$, by applying an activation function to the total. The Perceptron's weights are updated using a process known as backpropagation (detailed in Section 2.4.2.4).

Two additional factors need to be considered throughout the training process of the Perceptron (or any other ANN architecture): the batch size and the number of epochs. Epoch is a parameter that determines how many times the algorithm runs over the given set of data. The maximum amount of data that the algorithm can observe before adjusting its weights is defined by the batch size option **[166].**

b- MultiLayers Perceptron (MLP):

While the Perceptron represents a particular type of FNN, The term "feedforward neural network" generally refers to more complex architectures such as the MLP with hidden layers, which are capable of handling non-linear boundaries and solving more complex problems through deeper processing and learning mechanisms **[185] [186].**

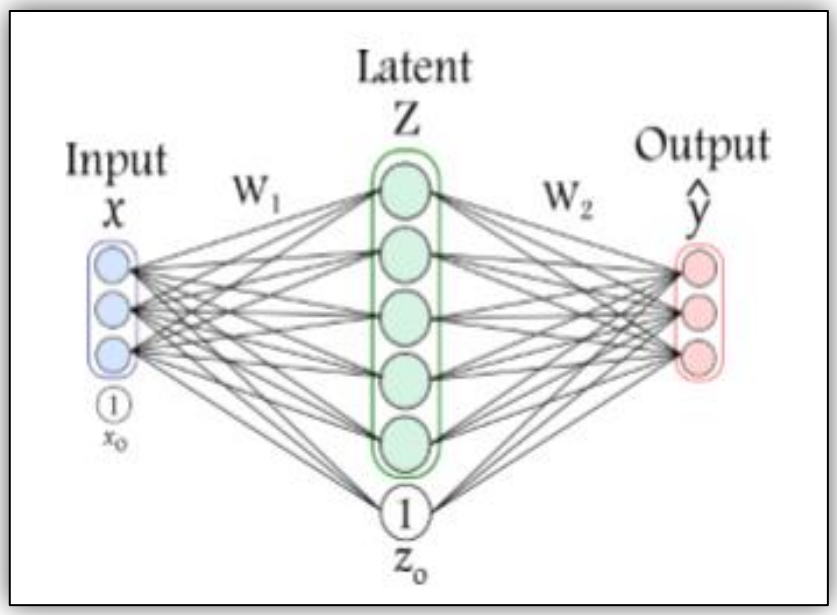

Figure 34: Multi-layer perceptron. Schematic representation of a MLP with single hidden layer **[183].**

The MLP can be thought of as an extension of the Perceptron since it is made up of multiple layers as opposed to just one, as the Perceptron is.

The input layer, the hidden layer, and the output layer are the three main layers of the MLP architecture. There can be one or many neurons in each one of them. The neural network receives its initial data from the input layer. All of the computing is done in the hidden layers, which are the layers that sit in between the input and output layers. The output layer generates the predictions for the supplied input.

An ANN is referred to as a Deep Neural Network (DNN) when it has a deep stack of hidden layers.

Like a Perceptron, an MLP learns the mapping function between input X and target label $\hat{y}$. However, each neuron in a multilayered network, cannot compute in the same way as a single neuron in a perceptron.

An input vector of size N, represented by$x_1,\ldots,x_N \in X$, is passed into the input layer of the MLP architecture, which outputs a list of values (each neuron *j* returns a scalar value y_j). The following equation describes the computation performed by each neuron:

$$\boldsymbol{\sigma_j^0 = \theta_j^0 + (\sum_{i=1}^{N} w_{ij}^0 x_i)} \textbf{(2.2)}$$

The connection between the input value $x_i$and the neuron j in the first layer is represented by the weight w_{ij}^0. g represents the activation function, and $\theta_j^0$is the bias that should be applied to neuron j.

The MLP architecture's hidden layer receives all of the size N_{l-1}neurons' outputs, represented asy_1^{l-1}, … , y_N^{l-1}, and outputs a new list of values (each neuron *j* returns a scalar value y_j^l, as computed in this equation:

$$\boldsymbol{\sigma_j^l = \theta_j^l + (\sum_{i=1}^{N_{l-1}} w_{ij}^l . y_i^{l-1},)} \textbf{(2.3)}$$

The connection between the output of neuron *i* at the preceding layer $\ell-1$ and the neuron *j* at layer ℓ is represented by the weight w_{ij}^l; the activation function is represented by g; and the bias to be applied to neuron *j* at layer ℓ is represented by θ_j^l[9].

c- Activation Function:

In order to address complicated nonlinear issues, artificial neural networks need to have an activation function during the learning phase. It is a scalar-to-scalar function that changes a neuron's inputs into an output signal known as the "activation level of the unit." Depending on how important a neuron's inputs are to the learning process, the latter will determine whether or not to activate a neuron. There are several activation functions (such as Sigmoid, Tanh, and ReLu), and their two main characteristics are differentiability and nonlinearity. In fact, more robust functions are required to simulate more complicated difficulties, given the constraints of linear functions. Thus, to add nonlinearity to models, non-linear activation functions are employed. Furthermore, taking into account the requirement to compute the loss function's gradient during the backpropagation stage.

For backpropagation to be possible, the activation function must be differentiable. Finally, another primary rationale for using activation functions is their capacity to limit the amplitude of the outputs by squashing them into a specific range, as we may obtain values in any range during the computation of weighted sums **[187].**

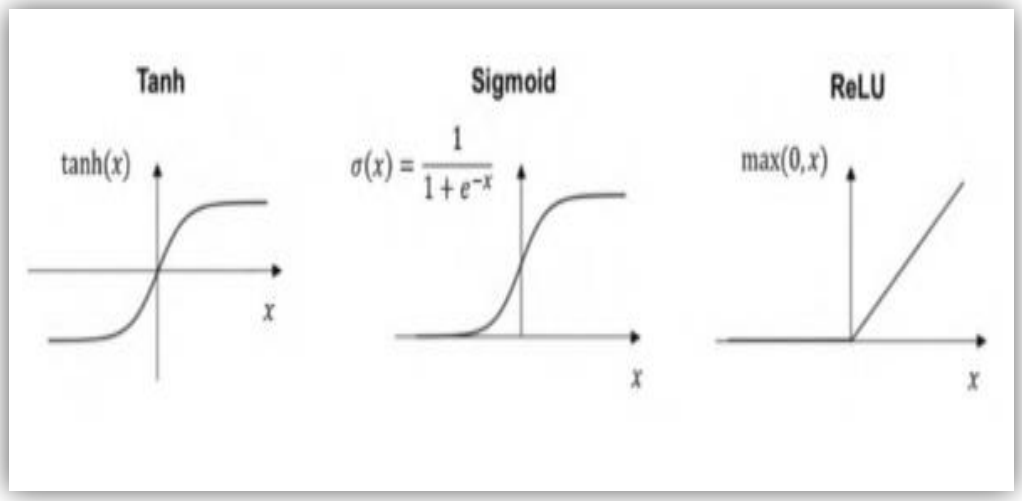

Figure 35: Examples of activation functions, used to introduce nonlinearity in feedforward MLPs **[160].**

d- Backpropagation:

Backpropagation is a fundamental algorithm in artificial neural networks (ANNs) that plays a crucial role in training neural networks by adjusting the model's weights to minimize the error between predicted and actual outputs. It involves a two-step process: Forward Propagation: During this step, input data is passed through the network to generate predictions.

backward propagation: the algorithm calculates the gradient of the loss function while respecting the network's weights, allowing efficient updates **[188] [189] [190].**

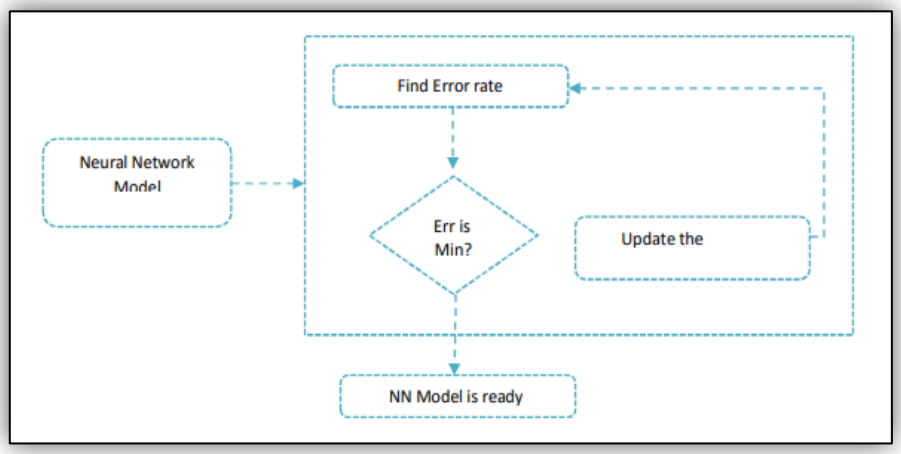

Figure 36: Backpropagation Algorithm and computational process **[188].**

The backpropagation algorithm relies on a loss function to quantify the network's error. For classification tasks, a binary cross-entropy function is commonly used, while mean squared error serves as the primary function for regression tasks. Specifically, the algorithm computes the gradient of the loss function respecting the model's parameters. This gradient guides the adjustment of connection weights and bias terms to minimize the loss error. Ensuring that the gradient is both computable and differentiable is essential.To update the weights effectively, we employ optimization algorithms such as gradient descent or other stochastic optimization methods. These algorithms apply the computed gradients iteratively until the network converges toward a solution.

The term "stochastic optimization" refers to the use of randomness in optimization techniques. Its primary goal is to find a global minimum while avoiding local minima traps. By reducing the probability of getting stuck in a local minimum, stochastic optimization algorithms increase the chances of finding the global minimum.

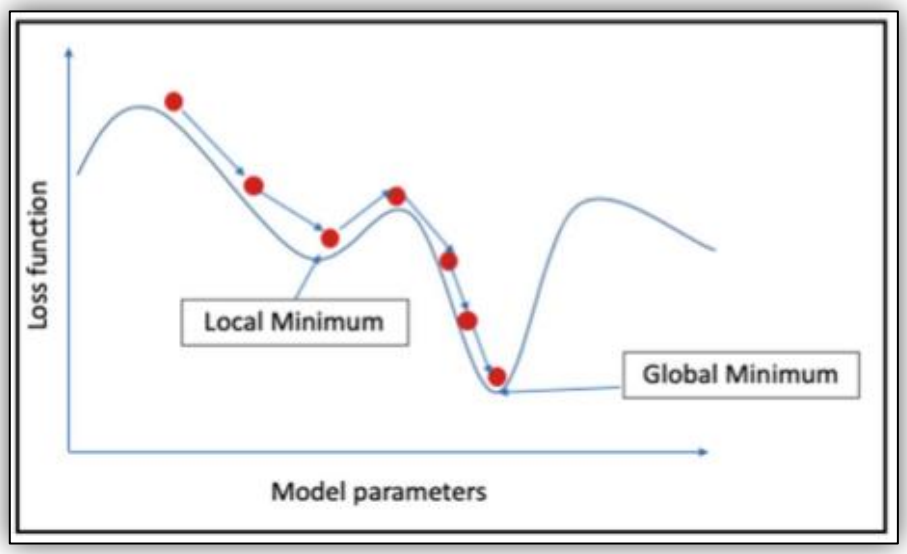

Figure 37: Illustration of local minimum and global minimum **[167].**

While local minimum may be acceptable solutions with interesting properties, they do not necessarily match the optimal solution. Our objective is to identify the global minimum. We define it using all available data from the training dataset, assuming that the global minimum from the training data aligns with the actual global minimum when sufficient data is present.

Convex functions naturally find global minimum because any local minimum for them is also a global minimum. However, non-convex functions pose a challenge, they require careful navigation to avoid getting trapped in local minimum.

To address this issue, various stochastic optimization algorithms (also known as optimizers) are available, including Adam, Root Mean Square Propagation (RMSprop), and Stochastic Gradient Descent (SGD) **[167].**

5- Regularization techniques:

In machine learning, one of the main problems is when a model works well with training data but performs poorly with new inputs (during the test phase, for example). This issue is known as overfitting, it arises when the model is excessively complex and has a high number of degrees of freedom, making it impossible for it to learn by predicting different concepts. Thankfully, an enormous amount of research has been performed to identify potential solutions for this problem. We refer to these methods as regularization techniques **[187].**

5-1- L1 and L2 Regularization:

Weight penalty regularization is a widely used model-training technique that is applied in L1 and L2. Models with higher weights are thought to be more complex than those with lower weights. Ensuring that the weights are either zero or extremely small is the purpose of the penalties. The weight penalty, sometimes referred to as weight decay, denotes the reduction of weights to zero or a smaller unit.

The so-called regularization term Ω is added to the neural network's J loss function during L1 and L2 regularization, which limits the model's capacity. The regularized objective function is given by:

$$\bar{J} = J + \lambda\Omega \quad (2.4)$$

where the regularization penalty term, Ω, is weighted in relation to the standard objective function J by a hyperparameter called λ. When λ is set to 0, no regularization occurs. Greater regularization is correlated with larger values of λ. Between L1 and L2, the regularization termΩ is different.

The sum of the absolute values of the weight parameters of the weight matrix is used as the regularization term Ωin the case of L1 regularization.

$$\Omega = \frac{1}{2}\|W\|_1 (2.5)$$

while in L2, Ω is defined as:

$$\Omega = \frac{1}{2}\|W\|_2^2 (2.6)$$

Unlike L2 regularization, which lowers the average magnitude of all weights, L1 regularization forces more weights to be zero, which introduces sparsity in the weights. Stated differently, L1 recommends that some features be excluded from training. However, the weight vector's directions (which don't really "contribute" much to the loss function) will be greatly affected by L2 regularization **[183].**

5-2- Data augmentation:

One method to prevent overfitting is to train the model on larger training datasets, which prevents the model from memorizing the network's parameters instead of allowing it to generalize to learn new concepts. The data augmentation technique is the name given to this approach **[187].**

For instance, the process of adding noise to a neural network's input can be regarded as data augmentation. Algorithms for unsupervised learning, like the denoising autoencoder, incorporate input noise injection. Hidden units can also be subjected to noise injection. This can be interpreted as a growth in the number of datasets at various abstraction levels **[183].** When considering object recognition, data augmentation allows us to replicate various variation factors from images. Even small translations of a few pixels can enhance performance. However, it’s crucial to avoid operations that could fundamentally change the nature of the augmented class (such as rotating “6” into “9).

Data augmentation needs to be aligned with each label's features to avoid bias introduction and to enhance performances **[160].**

5-3- Dropout:

Another effective regularization technique is dropout, which simulates the training of numerous neural networks with various architectures concurrently. A wide family of deep neural networks is regularized through dropout.

The dropout function's has a hyperparameter called the probability, p; this probability is used for choosing how many nodes to drop. The dropout layer experiences random unit termination (dropout) with probability p during training, which results in fewer neurons functioning in the forward process. As a result, the neural network's general structure was made simpler. However, we retain all the units during the test phase, so the values will be significantly higher than expected. Consequently, we have to reduce them by p.

The application of dropout forces all neurons to learn during the training phase, which eliminates the model's ability to rely on individual neurons (which may become muted in the process). As a result, a different "view" of the configured layer is used for each update of the dropout layer during training. The trained model is more robust because, conceptually, it approximates the training of numerous neural networks with various architectures in parallel **[183].**

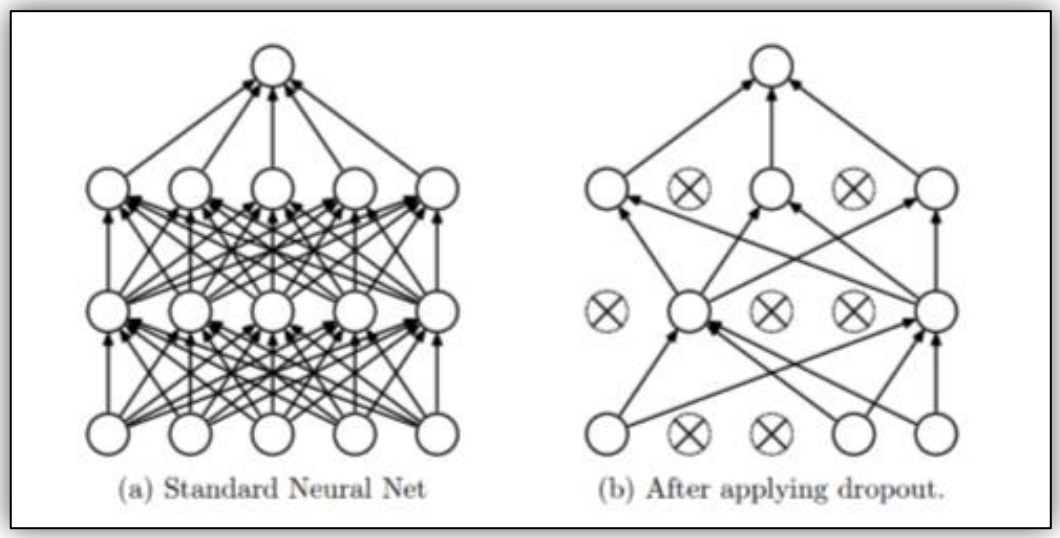

Figure 38: Neural network before (a) and after (b) applying dropout regularization technique **[187].**

5-4- Weightdecayapproach :

Weight decay, also known as weight regularization, is a well-known technique used to achieve faster convergence during model training, improve overall performance, and enhance model generalization to prevent overfitting. The weight decay strategy serves as a valuable alternative, especially in scenarios such as deep learning-based medical applications, where acquiring large training datasets is challenging. By incorporating a regularization term into the loss function, the weight decay strategy effectively regulates the growth of neural network weights **[187].**

5-5- Early stopping:

When training a deep network with a large learning capacity, the model attempts to gradually decrease the training data's loss function. But at some point in the training process, the model overfits the task and stops to generalize, instead learning the statistical noise present in the training dataset. It is difficult to train the network for a long enough time to produce a good input-to-output mapping without overfitting it with training data. The application of early stopping regularization is one method to solve this issue. The dataset can be divided into training, validation, and test sets to accomplish this. The algorithm is trained on the training set using early stopping, and the validation set is used to determine when training should end. It means that whenever the error in the validation set decreases during training, we save a copy of the model parameters. We switch to these parameters—which produce the least validation set error—instead of the most recent ones when the training algorithm finishes. After a predetermined number of iterations, the algorithm stops if no parameters have improved over the best-recorded validation error.

One of the most popular regularization techniques in deep learning is early stopping. Its simplicity and effectiveness are the reasons for its popularity. Since early stopping almost entirely avoids changing the objective function, the set of allowable parameter values, or the underlying training process, it can be regarded as implicit regularization. One can employ early stopping on its own or in combination with other regularization techniques. Because it lowers the training procedure's computational cost, early stopping is also beneficial.

A validation set is necessary for early stopping, which suggests that some data are not supplied to the model. After the first training is finished, more training can be carried out to fully utilize these additional data. The second training step contained all of the training data.

The second training procedure employed two fundamental strategies. Restarting the model and retraining it with all the data is one method. We trained the network for the same number of steps in this second training pass as we did in the first, using the early stopping strategy that was found to be ideal. However, there are some differences to this strategy that need to be considered.

Retaining the parameters from the initial training cycle and carrying on with the training while utilizing all the data is an additional approach to utilizing all the data. We no longer have a guideline for when to stop at this point. Alternatively, we can keep an eye on the validation set's average loss function and train until it drops below the training set objective's value, at which point the early stopping process is terminated. Although it is less well-behaved, this method avoids the high expense of retraining the model from scratch **[183].**

Conclusion:

In this chapter, I have covered foundational concepts related to artificial intelligence, machine learning, and deep learning. This includes a brief historical evolution of AI, defining some important concepts in machine learning, and describing the various ML algorithms. The discussion was extended to deep learning and artificial neural networks, including perceptrons, MLPs, activation functions, and backpropagation. I also highlighted a few regulation techniques, which play an essential role in improving the performance of the machine learning models.

Chapter III
The experimental part

Introduction:

In this chapter, I have worked on building machine learning models to predict the presence of seven cancer types (liver cancer, breast cancer, colorectal cancer, non-small cell lung cancer, pancreatic cancer, prostate cancer, and melanoma) using gene expression data of circulating tumor cells (CTCs) and microemboli (CTMs). First, I have built four binary classifiers. Each classifier was trained to distinguish one cancer class from other classes combined. Then I used the same models, but this time as multi-classifiers that could indicate the presence of these seven cancer types. This work aims to compare these two approaches and select the best-performing model for predicting cancer using this data type. The project was done using Python code, and the results are presented in detail in the following sections.

1- Evaluation metrics:

In the field of medicine, especially in cancer prediction using gene expression data, various evaluation metrics play an important role in assessing the performance of clinical tests. A basic tool is the confusion matrix (also known as the error matrix), which provides a comprehensive overview of how well a machine learning model performs on test data. The confusion matrix summarizes the predictions of the model by comparing them with the actual class labels for all data instances. It consists of four components: true-positives (TP), true-negatives (TN), false-positives (FP), and false-negatives (FN).

- **True Positives (TP)**: instances where the model accurately predicts a positive class.
- **True Negatives (TN)**: Instances where the model accurately predicts a negative class.
- **False Positives (FP)**: Instances where the model predicts a positive class incorrectly (type I errors).
- **False Negatives (FN)**: Instances where the model mispredicts a negative class (type II errors).

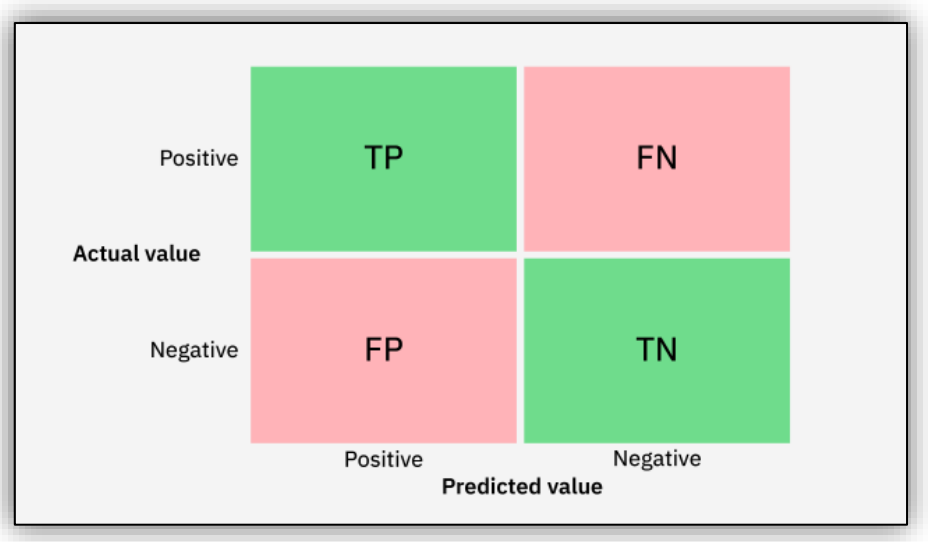

Figure 39: A standard confusion matrix template for a binary classifier **[194].**

These matrices are applicable to any classifier algorithm and allow the visualization of correct and incorrect predictions for all classes. In multiclass classification, the matrix extends to represent multiple classes **[191][192] [193] [194].**

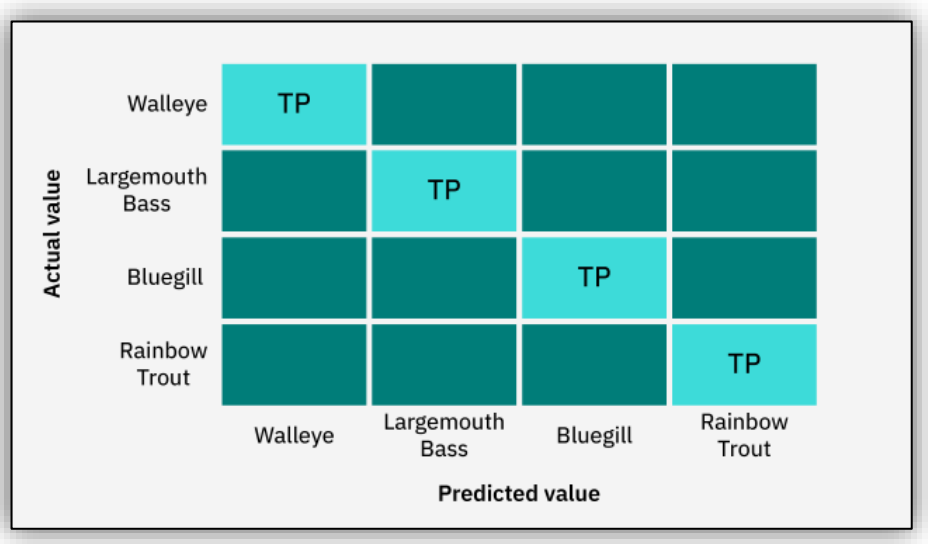

Figure 40: A confusion matrix for a multiclass classification problem **[194].**

Beyond basic accuracy metrics, the confusion matrix helps evaluate recall, precision, and overall model effectiveness **[194] [195] [196].**

- **Precision**: Measures the proportion of correctly predicted positive instances among all instances predicted as positive.

$$Precision = \frac{TP}{TP+FP}$$

✓ **Recall**: calculates the proportion of correctly predicted positive instances among all actual positive instances.

$$Recall = \frac{TP}{TP+FN}$$

✓ **F1-score**: the harmonic mean of precision and recall. It becomes particularly important in scenarios with unbalanced data and multi-class text classification.

$$F = \frac{2PR}{P+R}$$

✓ **Accuracy**: It determines the overall correctness of the predictions, but it alone is not always informative. For instance, a model with 99% accuracy may misclassify highly contagious diseases, posing a significant risk.

2- Tools:

The algorithm for the code was written in Python using Google Colab.

2-1- Python:

Python is an interpreter, object-oriented, and high-level programming language created by Guido van Rossum in the late 1980s. Its simplicity and readability have contributed to its

widespread adoption. Python's concise syntax makes it ideal for Rapid Application Development (RAD) and as a scripting language for gluing components together **[167] [197] [198].**

Python is a highly favored language for artificial intelligence (AI) projects due to its extensive library support and ease of use. These libraries cover a wide range of AI applications such as machine learning, deep learning, data science, and data visualization **[199] [200] [201].**

Figure 41: Python logo.

2-2- Google Colab:

Google Colab is a cloud-based service that provides a convenient environment for writing and running code using Jupyter Notebooks. These notebooks are widely used by data scientists and developers due to their interactive coding capabilities. Specifically, Colab is well-suited for machine learning and data analysis tasks, as it grants users free access to high-computing resources such as GPUs and TPUs. These resources are essential for efficiently training models, especially when dealing with large datasets. **[202]**

3- Experiment and results:

3-1- Gene expression Data:

Before constructing the models, I downloaded raw RNA-seq data from circulating tumor cells (CTCs) and circulating tumor microemboli (CTMs) across seven datasets available in ctcRbase **[201].** The downloaded data contains CTSs and CTMs samples; these samples represent seven cancer types: liver cancer (LIHC), breast cancer (BRCA), colorectal cancer (COAD), non-small cell lung cancer (LUSC), pancreatic cancer (PAAD), prostate cancer (PRAD), and melanoma (SKCM).

Cancer Type	Dataset name	Number of Samples	Use
liver cancer	**GSE117623**	**45**	**Training**
breast cancer	**GSE111065**	**69**	**Training**
colorectal cancer	**GSE74369**	**17**	**Training**
non-small cell lung cancer	**GSE74639**	**10**	**Training**
pancreatic cancer	**GSE40174**	**12**	**Training**
prostate cancer	**GSE104209**	**12**	**Training**
melanoma	**GSE38495**	**6**	**Training**
breast cancer	**GSE55807**	**6**	**Independent dataset**
pancreatic cancer	**GSE60407**	**7**	**Independent dataset**

Table 01: Gene expression data.

3-2 Coding Part:

- **Libraries and package validation:**

The first step in building the ML models was importing essential Python libraries and confirming their successful installation. These libraries collectively provided the necessary functionalities to preprocess the genomic data, develop machine learning models, evaluate their performance, and visualize the results.

a- **Pandas (pd):** An open-source, powerful data manipulation and analysis library in Python, Pandas offers a rich set of data structures and functions. It enables efficient operations on data, making it an indispensable tool for my project.

b- **NumPy (np):** NumPy is a Python library designed for working with arrays. It provides support for large, multi-dimensional arrays and matrices, performing operations up to 50 times faster than traditional Python lists. I leveraged NumPy extensively for efficient data handling.

c- **Scikit-learn:** As a comprehensive machine learning library in Python, Scikit-learn offers a wide range of algorithms and tools. I utilized it for data preprocessing, model training, and evaluation. Its user-friendly interface allowed me to experiment with various techniques seamlessly.

d- **TensorFlow (tf):** TensorFlow, an open-source software library, excels in high-performance numerical computation. It serves as a foundation for creating deep learning models directly or through wrapper libraries. My project benefited from TensorFlow's versatility and scalability.

e- **Seaborn (sns) and Matplotlib.pyplot (plt):** These Python libraries specialize in data visualization, providing customizable plotting functions. Seaborn enhances the aesthetics of plots, while Matplotlib.pyplot offers flexibility for visualizing patterns and relationships within the data.

```
import pandas as pd
import numpy as np
import sklearn
import tensorflow as tf
import seaborn as sns
import matplotlib.pyplot as plt
```

- **Uploading the datasets:**

The next step is to upload each dataset to google colab using this code:

```
#load data
GSE111065= pd.read_excel('/content/drive/MyDrive/PFE/database xlxs/BR_GSE111065_gene_FPKM_mat.xlsx')
```

- **The preprocessing steps:**

Raw gene expression data often contains missing values, noise, and inconsistencies, making preprocessing essential to ensuring data quality before classification. For this reason, before I started building the classifiers, I followed these preprocessing steps:

a- **Remove and impute Missing Values:** First, I dropped rows with missing values (NaN) from the dataset then I filled these NaN values with either the mean, median, or mode of the respective column.

```
# Data Cleaning Steps:
cleaned_GSE111065 = GSE111065.iloc[:, 1:]
cleaned_GSE111065.dropna(inplace=True)
cleaned_GSE111065.fillna(cleaned_GSE111065.mean(), inplace=True)
cleaned_GSE111065.fillna(cleaned_GSE111065.median(), inplace=True)
cleaned_GSE111065.fillna(cleaned_GSE111065.mode().iloc[0], inplace=True)
```

b- **Duplicate Removal:** second, I identified the duplicate rows and removed them.

```
# Duplicate Removal:
print(GSE111065.duplicated().sum())
GSE111065.drop_duplicates(inplace=True)
plt.figure(figsize=(12, 6))  # Adjust figure size if needed
sns.boxplot(data=GSE111065, palette='pastel')
plt.title('Grouped Boxplots for All Columns')
plt.xlabel('Columns')
plt.ylabel('Values')
plt.xticks(rotation=90)  # Rotate x-axis labels for readability
plt.show()
```

c- **Outlier Detection and Removal:** third, I filtered out outliers using the interquartile range (IQR) method, a measure of variability that tells us the range where the bulk of the values lie.

```
# Outlier Detection and Removal:
GSE111065 = GSE111065.iloc[:, 1:]
Q1 = GSE111065.quantile(0.25)
Q3 = GSE111065.quantile(0.75)
IQR = Q3 - Q1
GSE111065 = GSE111065[~((GSE111065 < (Q1 - 1.5 * IQR)) | (GSE111065 > (Q3 + 1.5 * IQR))).any(axis=1)]
```

d- Column Value Counts: fourth, I explored unique values in each column.

```
# Column Value Counts:
unique_values = {}
for column in GSE111065.columns[1:]:
    unique_values[column] = GSE111065[column].unique()

for column, values in unique_values.items():
    print(f"Unique values in '{column}': {values}")
```

e- Data Types Conversion: fifth, I converted the first column, which contains the gene names, to a float data type.

```
# Data Types Conversion:
column_index = 1  # Replace with the desired column index
GSE111065.iloc[:, column_index] = GSE111065.iloc[:, column_index].astype(float)
```

f- Robust Scaling: sixth, I Applied robust scaling to the cleaned dataset. Robust scaling is a technique used to standardize input variables in the presence of outliers. It involves ignoring the outliers from the calculation of the mean and standard deviation, then using the calculated values to scale the variable.

```
# Robust Scaling:
robust_scaler = RobustScaler()  # Instantiate RobustScaler
GSE111065_scaled = robust_scaler.fit_transform(GSE111065)  # Fit and transform
```

g- **Save Cleaned Dataset:** Finally, I saved the cleaned data to a new CSV file in order to work with it in further steps.

```
# Save Cleaned Dataset:
GSE111065.to_csv('GSE111065.csv', index=False)
```

- **Combining dataset:**

The data downloaded from ctcRbase was stored in multiple datasets, necessitating their combination.

For the binary classification, the idea was to merge at least two datasets, one specific to the cancer type I wanted to predict. Then, add a new column called the 'target column.' This column takes a value of 1 if the gene (raw data) corresponds to the target cancer and 0 otherwise.

```
# 1 for Breast cancer samples, 0 for melanoma cancer samples
GSE74369['target_cancer'] = 0
GSE111065['target_cancer'] = 1

# Concatenate the dataframes vertically
combined_Breast = pd.concat([GSE111065, GSE74369], ignore_index=True)

# Now combined_breast contains both melanoma and Breast cancer data
# with the 'target_cancer' column indicating the cancer type

# Shuffle the rows if needed
combined_Breast = combined_Breast.sample(frac=1).reset_index(drop=True)

# Save the combined dataset to a CSV file
combined_Breast.to_csv('combined_Breast.csv', index=False)


print(combined_Breast.head())
print(combined_Breast.shape)
```

For the multi-class classification, I combinedthe sevencancer's datasets, associated them with their specific cancer types, then, encoded the cancer types numerically.

```
datasets = [GSE117623,
            GSE74369,
            GSE74639,
            GSE38495,
            GSE40174,
            GSE111065,
            GSE104209
            ]
cancer_types = ['liver cancer',
                'colorectal cancer',
                'Non-small-cell-lung cancer',
                'Melanoma',
                'pancreatic cancer',
                'Breast cancer',
                'prostate cancer',
                ]

data_extend=[]
# Add 'Cancer_Type' column to each dataset
for df, cancer_type in zip(datasets, cancer_types):
    df['Cancer_Type'] = cancer_type
    print(df.head)
    data_extend.append(df)
```

```
le = LabelEncoder()
final_dataset['Encoded_Cancer_Type'] = le.fit_transform(final_dataset['Cancer_Type'])

final_dataset.drop('Cancer_Type',axis=1,inplace=True)

print(final_dataset.head)
```

- **handling NAN values :**

Due to the process of combining datasets, new missing values appeared in the combined data. To adjust this, I used the K-Nearest Neighbors (KNN) imputer technique to fill in missing values in the dataset.

```
# Initialize the KNNImputer with the desired number of neighbors (e.g., 2)
imputer = KNNImputer(n_neighbors=2)

# Create an empty DataFrame to store the imputed values
combined_Breast_imputed = pd.DataFrame(columns=combined_Breast.columns)

# Iterate over chunks of your dataset (adjust chunk size as needed)
chunk_size = 100
for i in range(0, len(combined_Breast), chunk_size):
    chunk = combined_Breast.iloc[i:i + chunk_size]
    imputed_chunk = imputer.fit_transform(chunk)
    combined_Breast_imputed = pd.concat([combined_Breast_imputed, pd.DataFrame(imputed_chunk, columns=chunk.columns)])

# Convert the imputed DataFrame back to the original shape
combined_Breast_imputed.reset_index(drop=True, inplace=True)
```

- **Using the SMOTE (Synthetic Minority Oversampling Technique):**

I used this step for multiclass classification with the combined dataset. SMOTE is an oversampling technique designed to address class imbalances. It generates synthetic samples for the minority class by interpolating between existing positive instances. Unlike simple duplication (which doesn't add new information), SMOTE creates new examples in the feature space. Given that my data was imbalanced (with some classes having significantly fewer examples than others), it was necessary to use this technique to prevent machine learning models from performing poorly on the minority class due to a lack of sufficient training data.

```
# SMOTE
smote = SMOTE(random_state=42)

# oversampling
X_res, y_res = smote.fit_resample(X, y)

# Printing Class Distribution after the oversampling
print("Répartition des classes après l'oversampling :", pd.Series(y_res).value_counts())
```

```
Répartition des classes après l'oversampling : Encoded_Cancer_Type
4    60
3    60
2    60
1    60
5    60
0    60
6    60
Name: count, dtype: int64
```

- **Building the classifiers:**

The selection of an appropriate classifier is crucial for the project, as each model possesses unique strengths and weaknesses. For this project, three machine learning models and one deep learning model were selected:

- ❖ Random Forest Classifier.
- ❖ Gradient Boosting Classifier.
- ❖ Decision Tree Classifier.
- ❖ Feedforward Neural Network.

Experiment 1: Binary Classification

In the first experiment, I used the selected Algorithms as binary classifiers, each classifier was trained seven times, with each iteration focusing on distinguishing one type of cancer from other types combined. The models' parameters were optimized using the breast cancer dataset. This particular dataset was chosen due to the larger number of CTCs and CTMs samples available compared to other cancer types. Once the optimal configuration for each model was determined, it was applied to the remaining cancer types.

The model	Parameters	Configuration
Random Forest	**n_estimators**	**600**
	random_state	**42**
Gradient Boosting	**n_estimators**	**1000**
	max_depth	**20**
	min_samples_split	**20**
	learning_rate	**0.09**
	loss	**squared_error**
Decision Tree	**criterion**	**gini**
	max_depth	**none**
	min_samples_split	**2**
	min_samples_leaf	**1**

Feed forward Neural Network	**Dense Layers**	**1st**	**neurons**	**256**
			function	**ReLU**
		2nd	**neurons**	**128**
			function	**ReLU**
	Output Layer	**neurons**	**1**	
		function	**sigmoid**	
	compilation	**optimizer**	**adam**	
		loss	**binary_crossentropy**	

Table 02: The models' configuration.

1- Splitting the data:

The last step before models' training is the division of data into training and testing sets. For the machine learning classifiers, the data was split into 80% for training and 20% for testing.

```
# Split into training and testing sets (80% train, 20% test)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

For the deep learning model, the data was divided into 70% for training, 15% for testing, and another 15% for validation.

```
train_df, temp_df = train_test_split(combined_Breast_imputed, test_size=0.3, random_state=42)
val_df, test_df = train_test_split(temp_df, test_size=0.5, random_state=42)
```

2- The results:

2-1- Random forest classifier:

The tables and matrices below show the results obtained after using the Random Forest model for each cancer type:

- ✓ **Liver cancer (LIHC):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.994**	**0.994**	**0.994**	**0.994**

Table 03: Training results of the Random Forest model for predicting liver cancer.

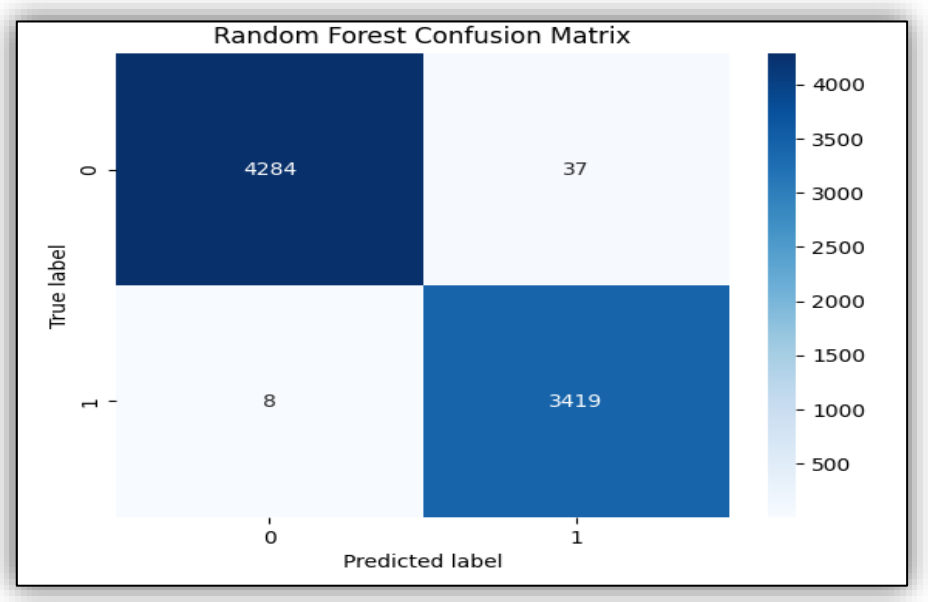

Figure 42: Random Forest confusion matrix (liver cancer).

True Negative (TN): 4284 (correctly predicted).

False Positive (FP): 37 (incorrectly predicted).

False Negative (FN): 8 (cases missed by the model).

True Positive (TP): 3419 (correctly predicted cases).

✓ **Breast cancer (BRCA):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.933**	**0.934**	**0.933**	**0.933**

Table 04: Training results of the Random Forest model for predicting breast cancer.

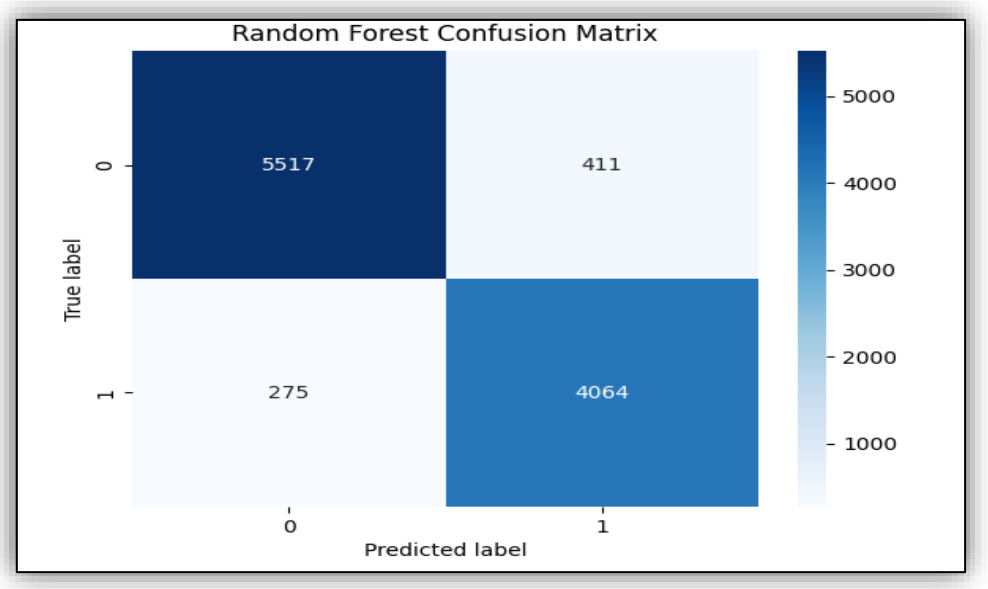

Figure 43: Random Forest confusion matrix (breast cancer).

True Negative (TN): 5517 (correctly predicted cases).

False Positive (FP): 411 (incorrectly predicted cases).

False Negative (FN): 275 (cases missed by the model).

True Positive (TP): 4064 (correctly predicted cases).

✓ **Colorectal cancer (COAD):**

	Accuracy	Precission	Recall	F1 score
Model Effectiveness	**0.977**	**0.977**	**0.977**	**0.977**

Table 05: Training results of the Random Forest model for predicting colorectal cancer.

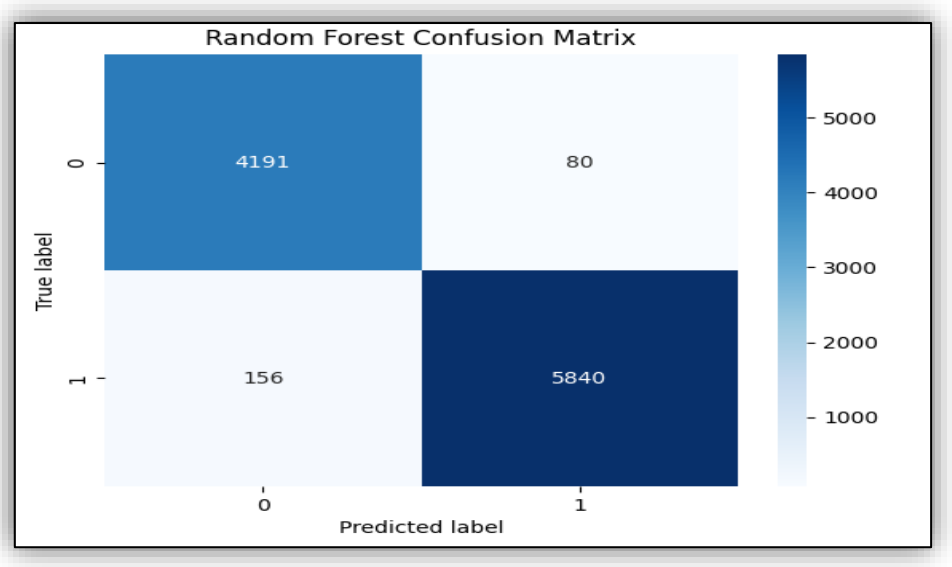

Figure 44: Random Forest confusion matrix (colorectal cancer).

True Negative (TN): 4191 (correctly predicted cases).

False Positive (FP): 80 (incorrectly predicted cases).

False Negative (FN): 156 (cases missed by the model).

True Positive (TP): 5840 (correctly predicted cases).

- ✓ **Non-small cell lung cancer (LUSC):**

	Accuracy	**Precission**	**Recall**	**F1 score**
Model Effectiveness	**0.942**	**0.946**	**0.942**	**0.942**

Table 06: Training results of the Random Forest model for predicting Non-small cell lung cancer.

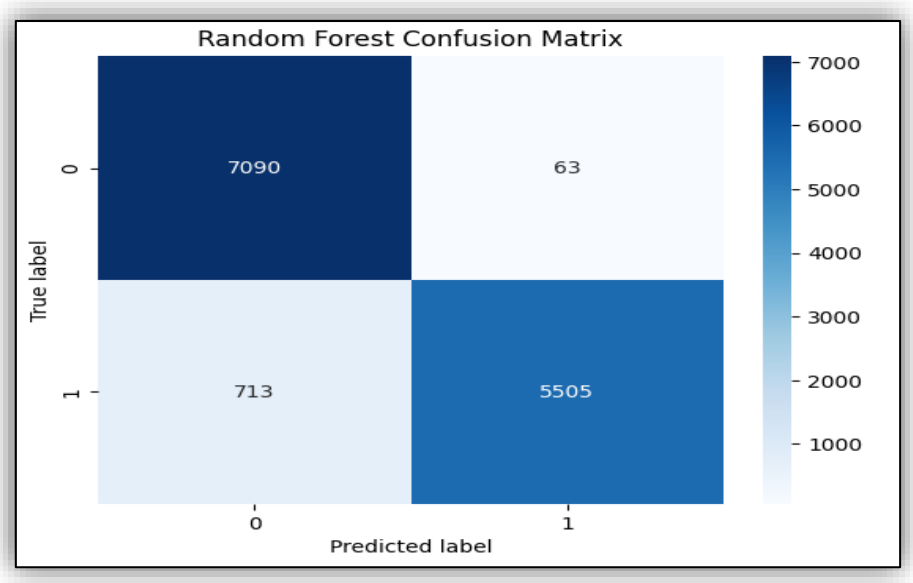

Figure 45: Random Forest confusion matrix (predicting Non-small cell lung cancer).

True Negatives (TN): 7090 (correctly predicted cases).

False Positives (FP): 63 (misclassified case).

False Negatives (FN): 713 (missed cases).

True Positives (TP): 5505 (correctly predicted cases).

✓ **Pancreatic cancer (PAAD):**

	Accuracy	**Precission**	**Recall**	**F1 score**
Model Effectiveness	**0.925**	**0.927**	**0.925**	**0.925**

Table 07: Training results of the Random Forest model for predicting pancreatic cancer.

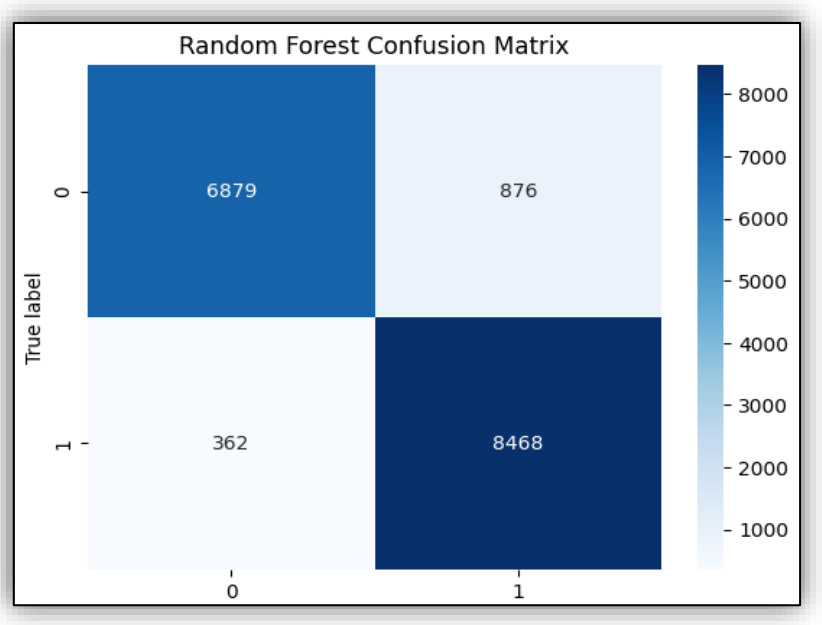

Figure 46: Random Forest confusion matrix (pancreatic cancer).

True Negatives (TN): 6879 (correctly predicted non-cancer cases)

False Positives (FP): 876 (non-cancer cases misclassified as cancer)

False Negatives (FN): 362 (missed cancer cases)

True Positives (TP): 8468 (correctly predicted cancer cases)

✓ **Prostate cancer (PRAD):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.944**	**0.948**	**0.944**	**0.944**

Table 08: Training results of the Random Forest model for predicting prostate cancer.

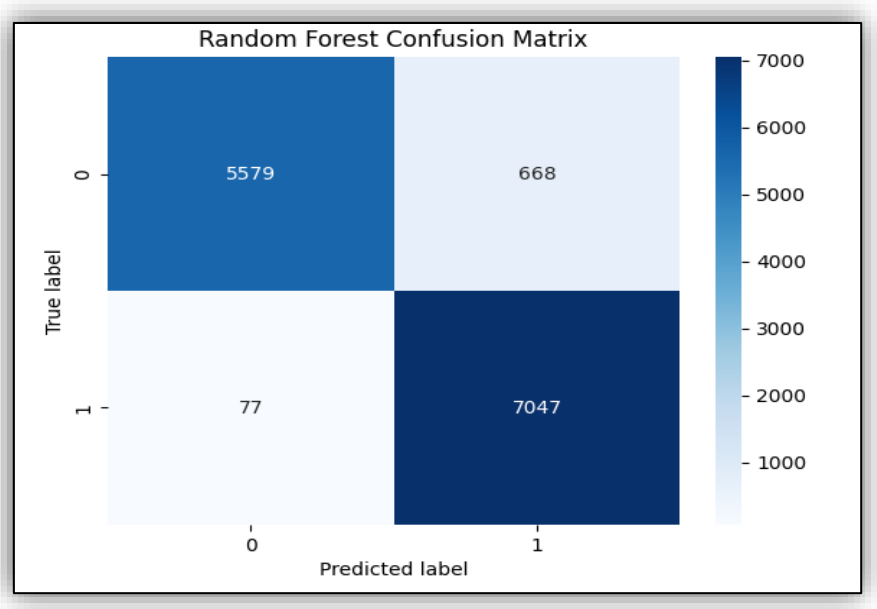

Figure 47: Random Forest confusion matrix (prostate cancer).

True Negatives (TN): 5579 (correctly predicted cases).

False Positives (FP): 668 (misclassified cases).

False Negatives (FN): 77 (missed cases).

True Positives (TP): 7047 (correctly predicted cases).

✓ **Melanoma (SKCM):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.926**	**0.927**	**0.926**	**0.926**

Table 09: Training results of the Random Forest model for predicting melanoma.

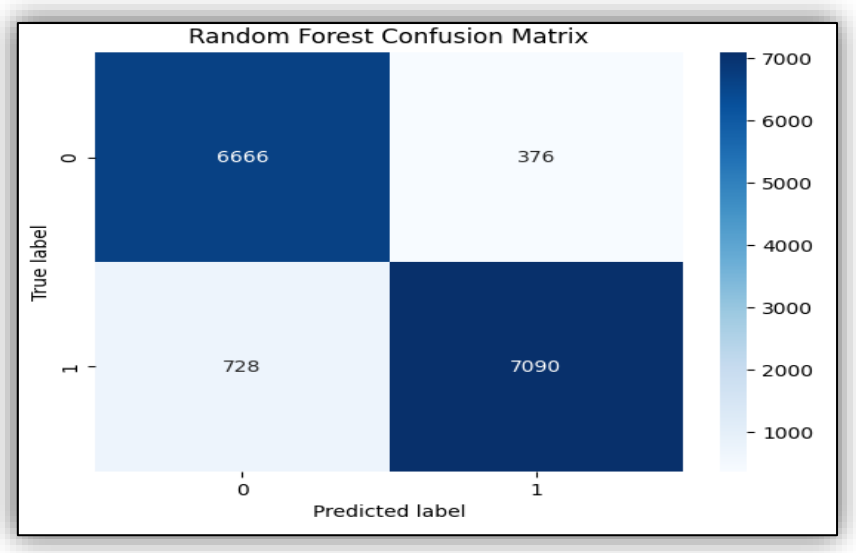

Figure 48: Random Forest confusion matrix (melanoma).

True Negatives (TN): 6666 (correctly predicted cases)

False Positives (FP): 376 (misclassified cases)

False Negatives (FN): 728 (missed cases)

True Positives (TP): 7090 (correctly predicted cases)

2-2- Gradient Boosting Classifier:

The tables and matrices below show the results obtained after using the Gradient Boosting Classifiermodel for each cancer type:

- ✓ **Liver cancer (LIHC):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.995**	**0.995**	**0.995**	**0.995**

Table 10: Training results of the gradient boosting model for predicting liver cancer.

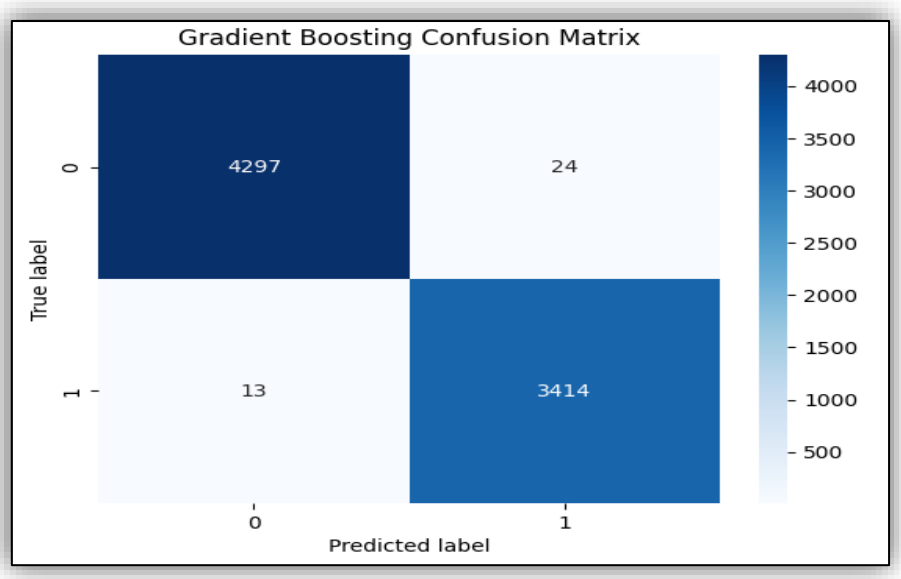

Figure 49: gradient boosting confusion matrix (liver cancer).

True Negative (TN): 4297 (correctly predicted cases).

False Positive (FP): 24 (incorrectly predicted cases).

False Negative (FN):13 (cases missed by the model).

True Positive (TP): 3414 (correctly predicted cases).

- ✓ **Breast cancer (BRCA):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.934**	**0.935**	**0.934**	**0.934**

Table 11: Training results of the gradient boosting model for predicting breast cancer.

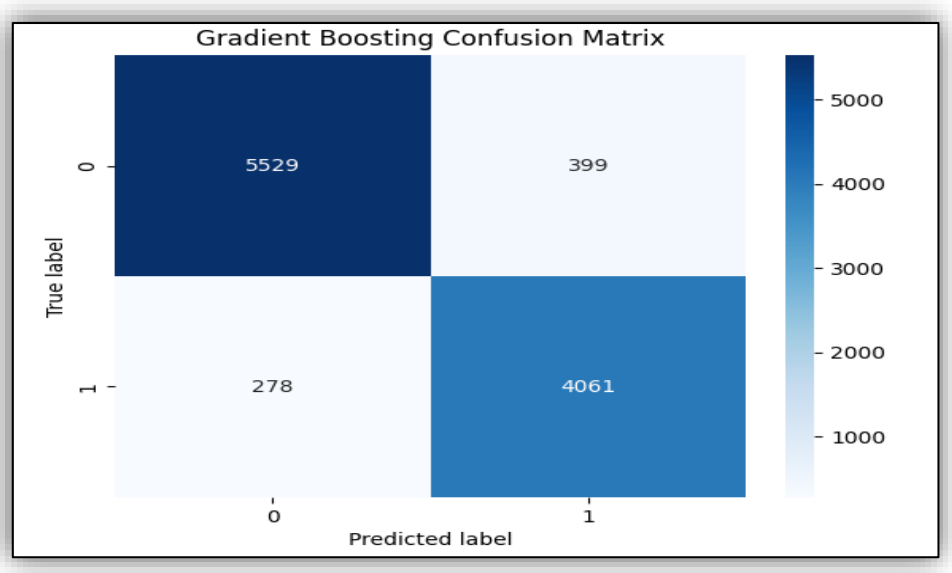

Figure 50: gradient boosting confusion matrix (breast cancer).

True Negative (TN): 5529 (correctly predicted cases).

False Positive (FP): 399 (incorrectly predicted cases).

False Negative (FN): 278 (cases missed by the model).

True Positive (TP): 4061 (correctly predicted cases).

- ✓ **Colorectal cancer (COAD):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.978**	**0.978**	**0.978**	**0.978**

Table 12: Training results of the gradient boosting model for predicting colorectal cancer.

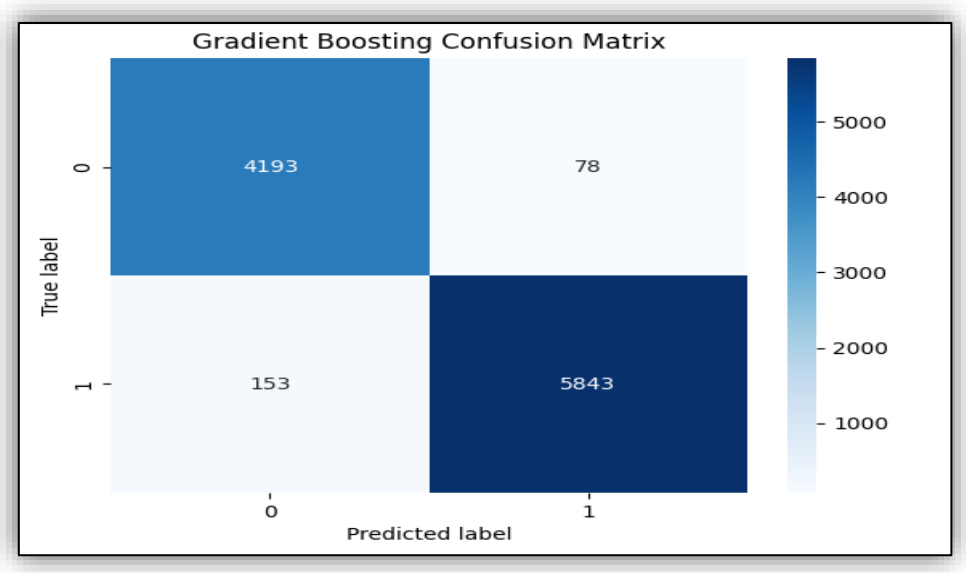

Figure 51: gradient boosting confusion matrix (colorectal cancer).

True Negative (TN): 4193 (correctly predicted cases).

False Positive (FP): 78 (incorrectly predicted cases).

False Negative (FN): 153 (cases missed by the model).

True Positive (TP): 5843 (correctly predicted cases).

- ✓ **Non-small cell lung cancer (LUSC):**

	Accuracy	**Precision**	**Recall**	**F1 score**
Model Effectiveness	**0.939**	**0.942**	**0.939**	**0.939**

Table 13: Training results of the gradient boosting model for predicting Non-small cell lung cancer.

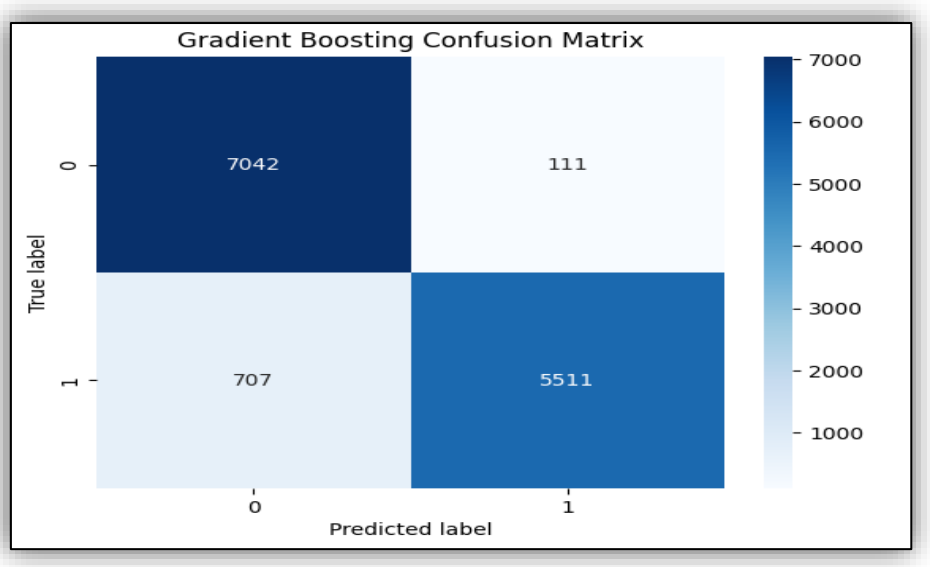

Figure 52: gradient boosting confusion matrix (Non-small cell lung cancer).

True Negatives (TN): 7042 (correctly predicted cases).

False Positives (FP): 111 (cases misclassified as cancer).

False Negatives (FN): 707 (missed cases).

True Positives (TP): 5511 (correctly predicted cases).

✓ **Pancreatic cancer (PAAD):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.943**	**0.927**	**0.943**	**0.943**

Table 14: Training results of the gradient boosting model for predicting pancreatic cancer.

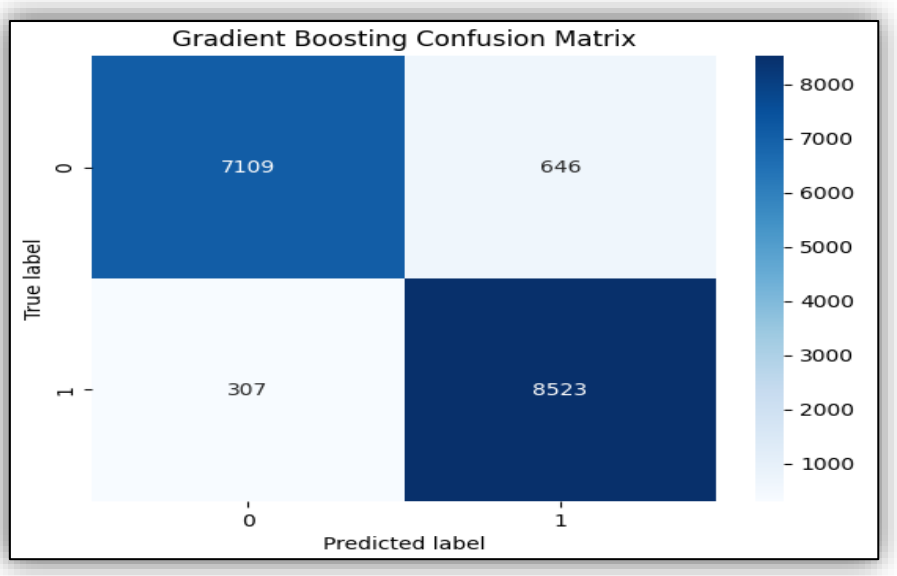

Figure 53: gradient boosting confusion matrix (pancreatic cancer).

True Negatives (TN): 7109 (correctly predicted cases).

False Positives (FP): 646 (cases misclassified as cancer).

False Negatives (FN): 307 (missed cases).

True Positives (TP): 8523 (correctly predicted cases).

✓ **Prostate cancer (PRAD):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.943**	**0.946**	**0.943**	**0.943**

Table 15: Training results of the gradient boosting model for predicting prostate cancer.

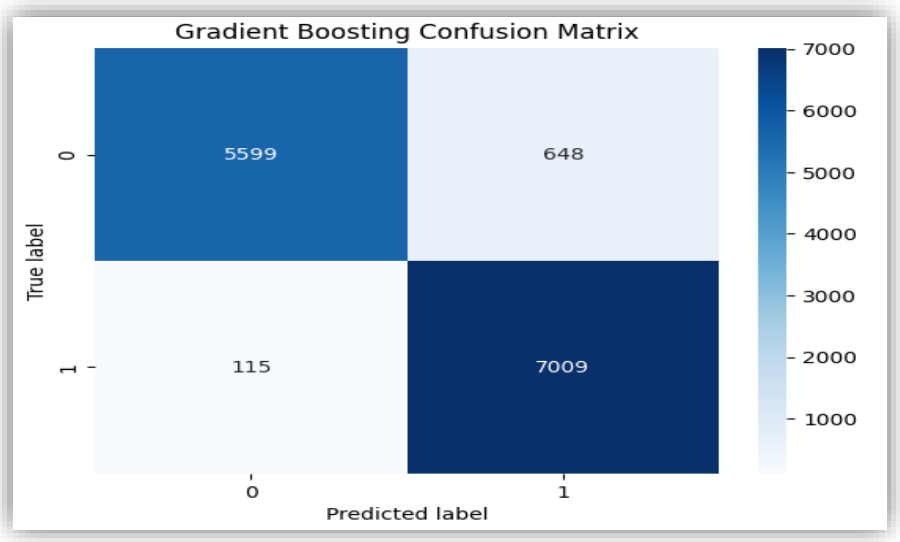

Figure 54: gradient boosting confusion matrix (prostate cancer).

True Negatives (TN): 5599 (correctly predicted cases).

False Positives (FP): 648 (misclassified cases).

False Negatives (FN): 115 (missed cases).

True Positives (TP): 7009 (correctly predicted cases).

✓ **Melanoma (SKCM):**

	Accuracy	**Precision**	**Recall**	**F1 score**
Model Effectiveness	**0.946**	**0.947**	**0.946**	**0.946**

Table 16: Training results of the gradient boosting model for predicting melanoma.

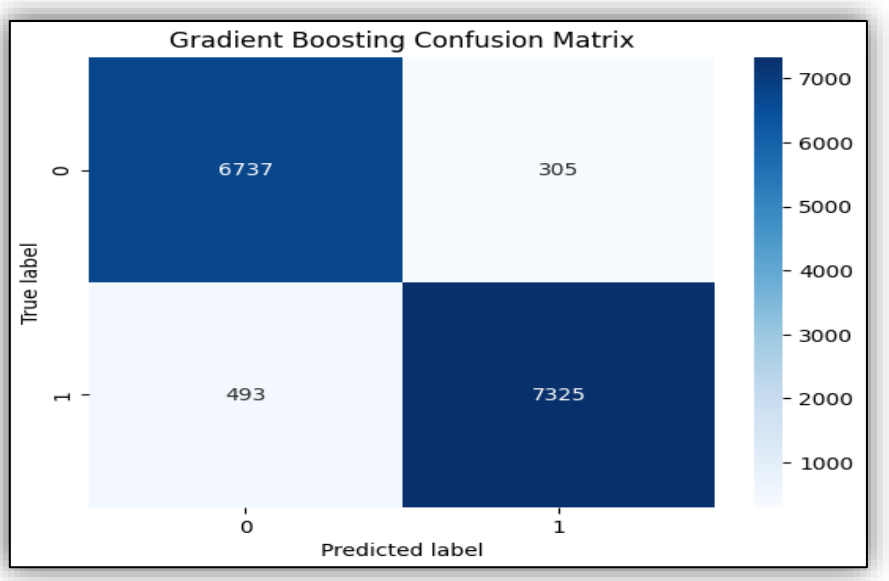

Figure 55: gradient boosting confusion matrix (melanoma).

True Negatives (TN): 6737 (correctly predicted cases).

False Positives (FP): 305(misclassified cases).

False Negatives (FN): 493 (missed cases).

True Positives (TP): 7325 (correctly predicted cases).

2-3- Decision Tree classifier:

The tables and matrices below show the results obtained after using the Decision Tree model for each cancer type:

- ✓ **Liver cancer (LIHC):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.989**	**0.989**	**0.989**	**0.989**

Table 17: Training results of the Decision Tree model for predicting liver cancer.

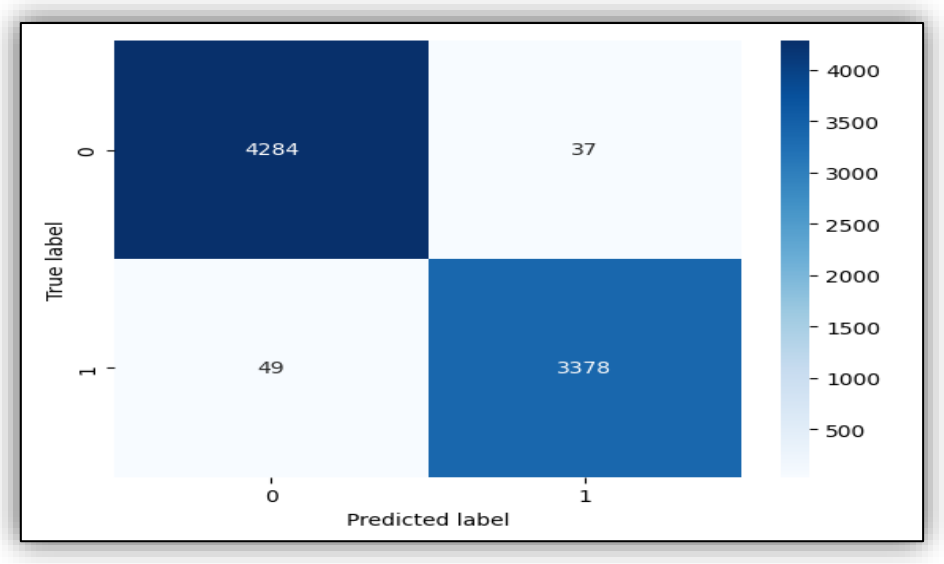

Figure 56: Decision Tree confusion matrix (liver cancer).

True Negatives (TN): 4284 (correctly predicted cases).

False Positives (FP): 37 (misclassified cases).

False Negatives (FN): 49 (missed cases).

True Positives (TP): 3378 (correctly predicted cases).

✓ **Breast cancer (BRCA):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.943**	**0.943**	**0.943**	**0.943**

Table 18: Training results of the Decision Tree model for predicting breast cancer.

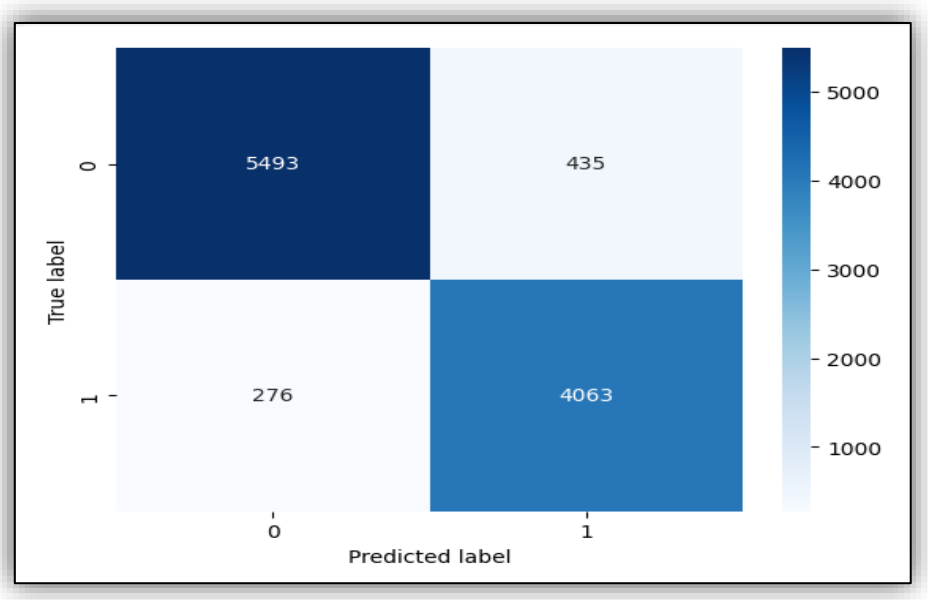

Figure 57: Decision Tree confusion matrix (breast cancer).

True Negative (TN): 5493 (correctly predicted cases).

False Positive (FP): 435 (incorrectly predicted cases).

False Negative (FN): 276 (cases missed by the model).

True Positive (TP): 4063 (correctly predicted cases).

- ✓ **Colorectal cancer (COAD):**

	Accuracy	Precission	Recall	F1 score
Model Effectiveness	**0.977**	**0.977**	**0.977**	**0.977**

Table 19: Training results of the Decision Tree model for predicting colorectal cancer.

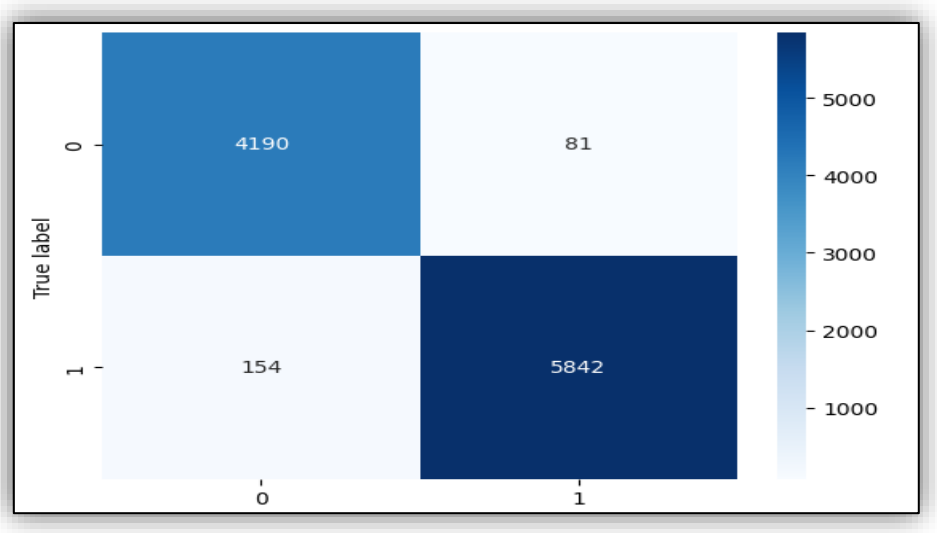

Figure 58: Decision Tree confusion matrix (colorectal cancer).

True Negative (TN): 4190 (correctly predicted non-cancer cases).

False Positive (FP): 81 (incorrectly predicted as cancer cases).

False Negative (FN): 154 (actual cancer cases missed by the model).

True Positive (TP): 5842 (correctly predicted cancer cases).

- ✓ **Non-small cell lung cancer (LUSC):**

	Accuracy	Precission	Recall	F1 score
Model Effectiveness	**0.930**	**0.932**	**0.930**	**0.929**

Table 20: Training results of the Decision Tree model for predicting Non-small cell lung cancer.

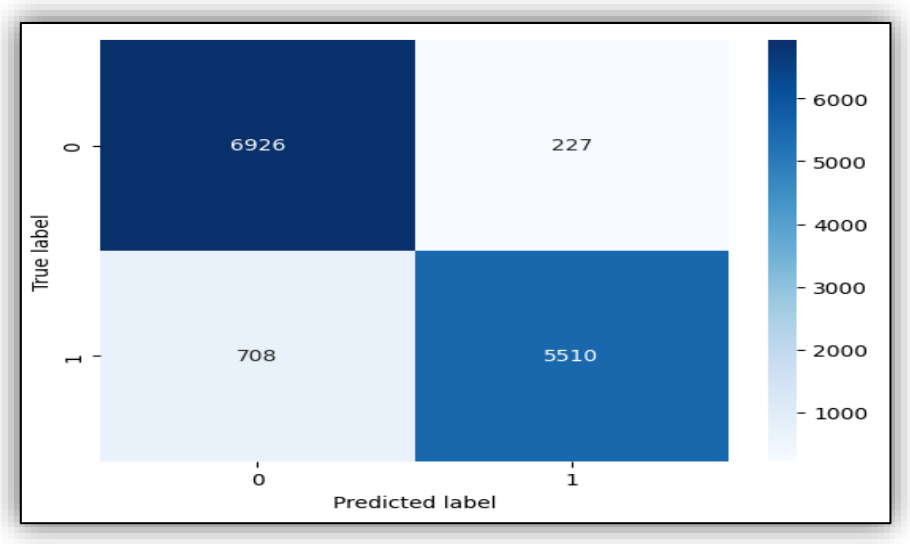

Figure 59: Decision Tree confusion matrix (Non-small cell lung cancer).

True Negatives (TN): 6936 (correctly predicted cases).

False Positives (FP): 227 (misclassified cases).

False Negatives (FN): 708 (missed cases).

True Positives (TP): 5510 (correctly predicted cases).

✓ **Pancreatic cancer (PAAD):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.885**	**0.885**	**0.885**	**0.885**

Table 21: Training results of the Decision Tree model for predicting pancreatic cancer.

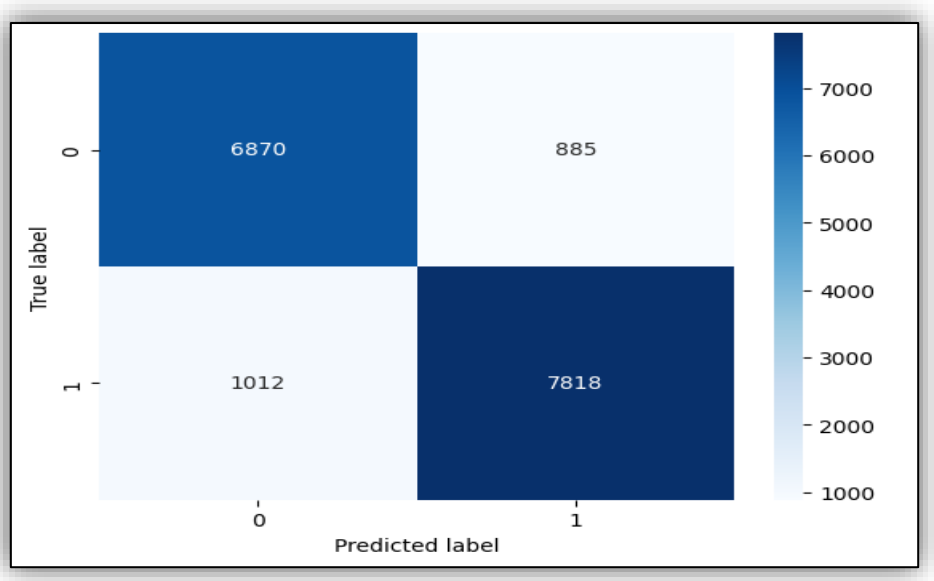

Figure 60: Decision Tree confusion matrix (pancreatic cancer).

True Negatives (TN): 6870 (correctly predicted cases).

False Positives (FP): 885 (misclassified cases).

False Negatives (FN): 1012 (missed cases).

True Positives (TP): 7818 (correctly predicted cases).

✓ **Prostate cancer (PRAD):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.932**	**0.934**	**0.932**	**0.932**

Table 22: Training results of the Decision Tree model for predicting prostate cancer.

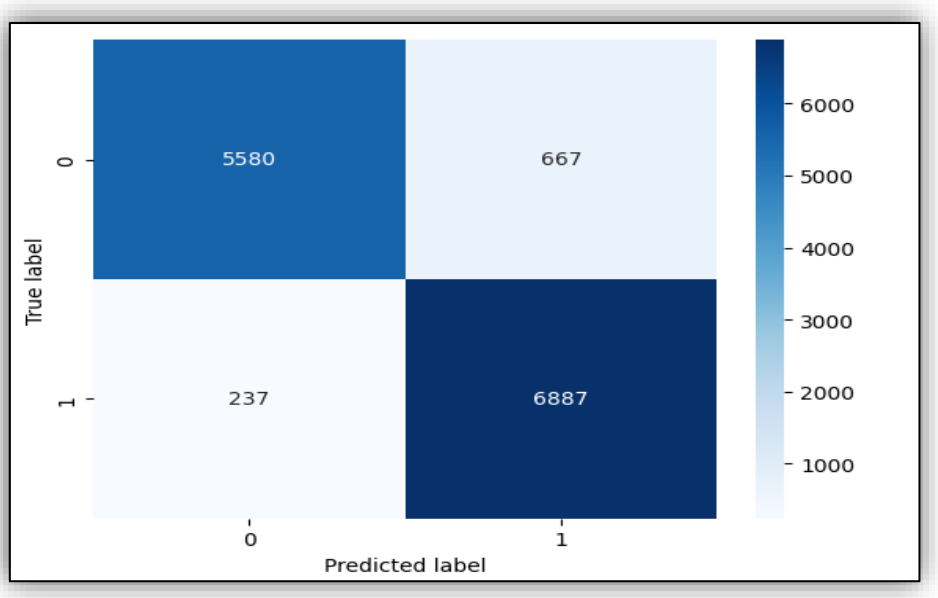

Figure 61: Decision Tree confusion matrix (prostate cancer).

True Negatives (TN): 5580 (correctly predicted cases).

False Positives (FP): 667 (cases misclassified as cancer).

False Negatives (FN): 237 (missed cases).

True Positives (TP): 6887 (correctly predicted cases).

✓ **Melanoma (SKCM):**

	Accuracy	**Precision**	**Recall**	**F1 score**
Model Effectiveness	**0.892**	**0.892**	**0.892**	**0.892**

Table 23: Training results of the Decision Tree model for predicting melanoma.

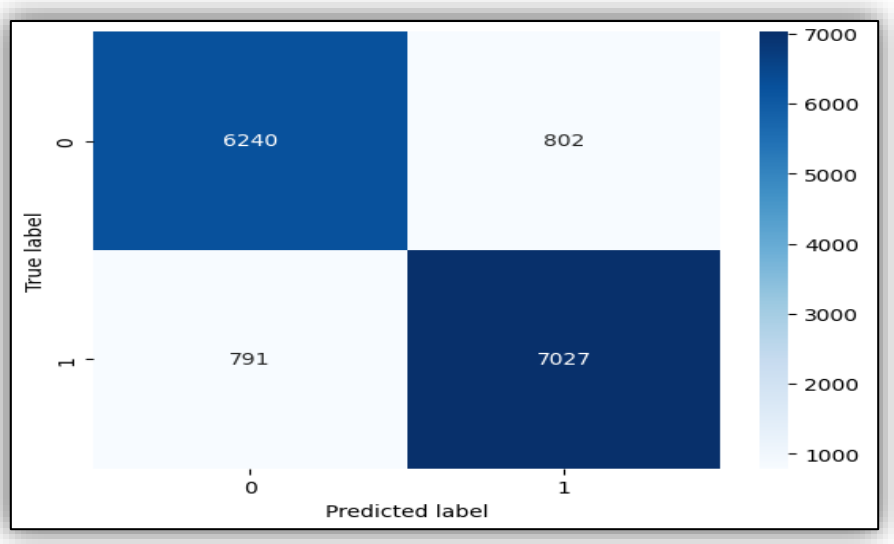

Figure 62: Decision Tree confusion matrix (melanoma).

True Negatives (TN): 6240 (correctly predicted cases).

False Positives (FP): 802 (misclassified cases).

False Negatives (FN): 791 (missed cases).

True Positives (TP): 7027 (correctly predicted cases).

2-4- Feed forward neural network:

The tables and matrices below show the results obtained after using the feedforward neural network model for each cancer type:

- ✓ **Liver cancer (LIHC):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.975**	**0.980**	**0.963**	**0.972**

Table 24: Training results of the feed forward neural network model for predicting liver cancer.

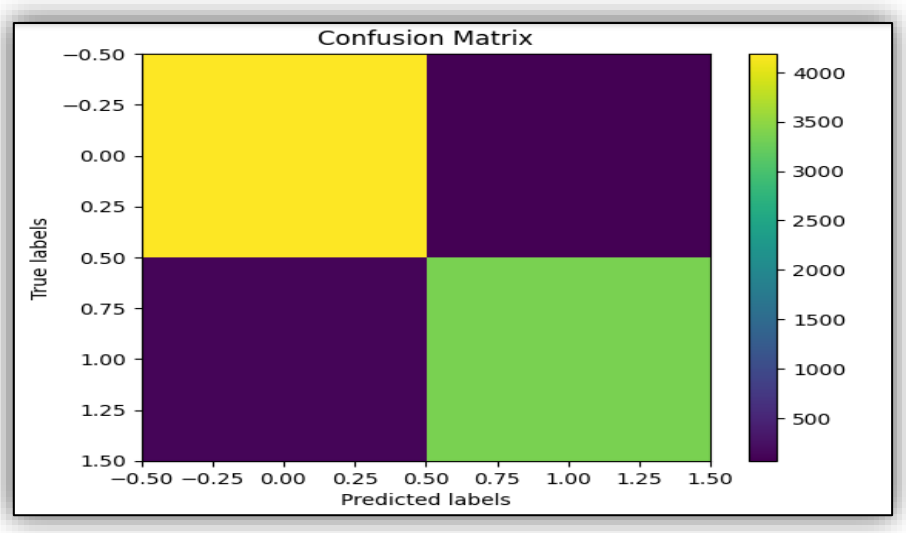

Figure 63: feed forward neural network confusion matrix (liver cancer).

True Negative (TN): 4187 (correctly predicted cases).

False Positive (FP): 67 (incorrectly predicted cases).

False Negative (FN): 128 (cases missed by the model).

True Positive (TP): 3366 (correctly predicted cases).

- ✓ **Breast cancer (BRCA):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.865**	**0.801**	**0.900**	**0.847**

Table 25: Training results of the feed forward neural network model for predicting breast cancer.

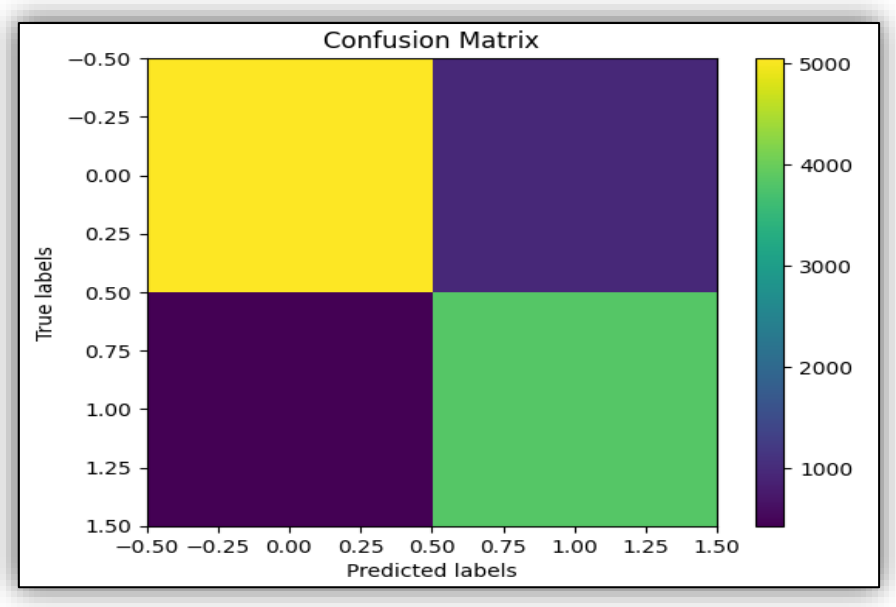

Figure 64: feed forward neural network confusion matrix (breast cancer).

True Negative (TN): 5053 (correctly predicted non-cancer cases).

False Positive (FP): 953 (incorrectly predicted as cancer cases).

False Negative (FN): 428 (actual cancer cases missed by the model).

True Positive (TP): 3833 (correctly predicted cancer cases).

- ✓ **Colorectal cancer (COAD):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.920**	**0.966**	**0.893**	**0.928**

Table 26: Training results of the feed forward neural network model for predicting colorectal cancer.

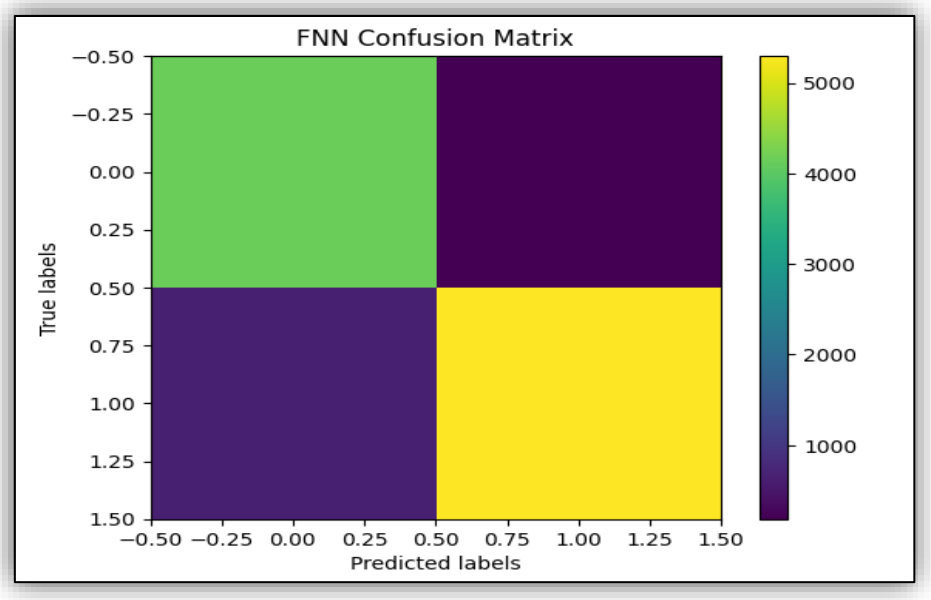

Figure 65: feed forward neural network confusion matrix (colorectal cancer).

True Negative (TN): 4149 (correctly predicted cases).

False Positive (FP): 185 (incorrectly predicted cases).

False Negative (FN): 363 (cases missed by the model).

True Positive (TP):5297 (correctly predicted cases).

- ✓ **Non-small cell lung cancer (LUSC):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.827**	**0.810**	**0.820**	**0.815**

Table 27: Training results of the feed forward neural network model for predicting Non-small cell lung cancer.

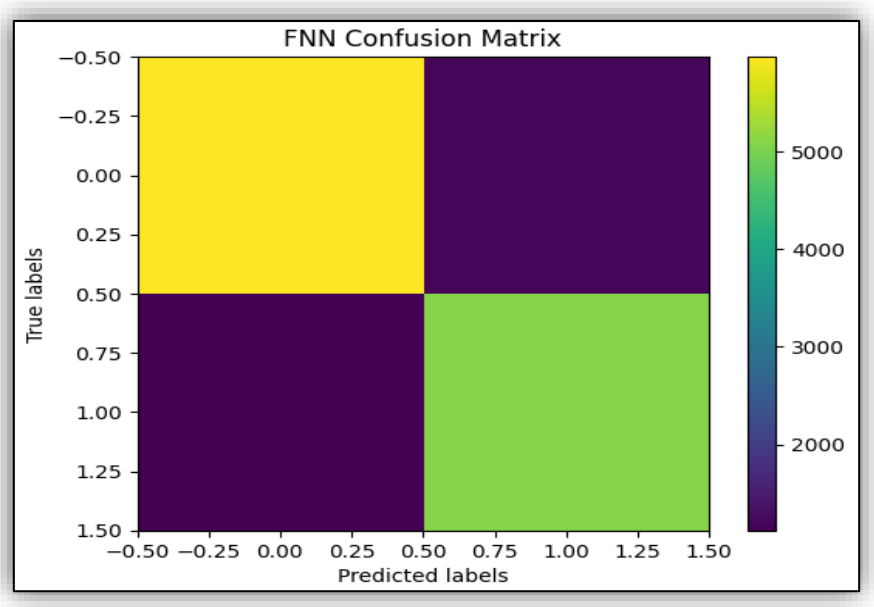

Figure 66: feed forward neural network confusion matrix (Non-small cell lung cancer).

True Negatives (TN): 5972(correctly predicted non-cancer cases).

False Positives (FP): 1193 (non-cancer cases misclassified as cancer).

False Negatives (FN): 1019 (missed cancer cases).

True Positives (TP): 5087(correctly predicted cancer cases).

✓ **Pancreatic cancer (PAAD):**

	Accuracy	**Precision**	**Recall**	**F1 score**
Model Effectiveness	**0.893**	**0.895**	**0.905**	**0.900**

Table 28: Training results of the feed forward neural network model for predicting pancreatic cancer.

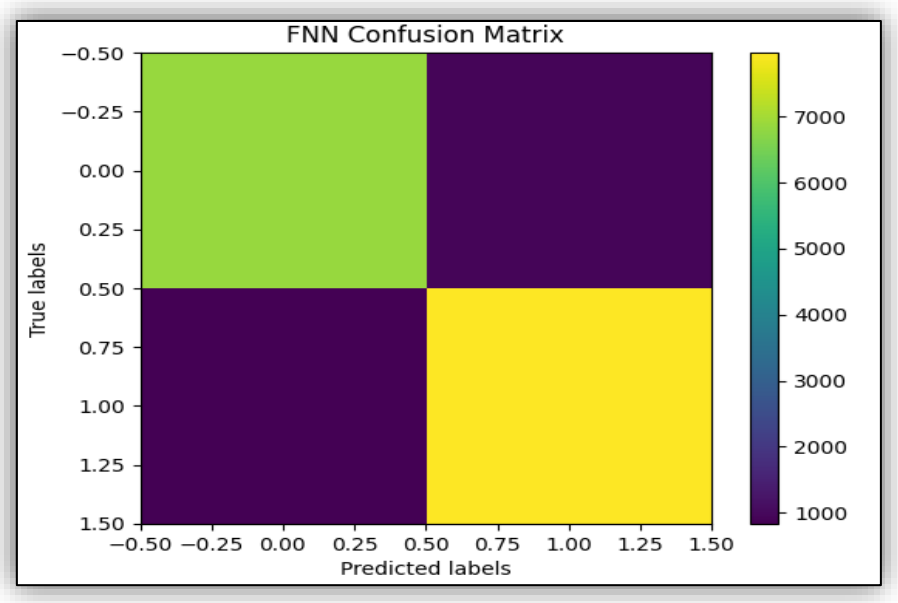

Figure 67: feed forward neural network confusion matrix (pancreatic cancer).

True Negatives (TN): 6851 (correctly predicted non-cancer cases).

False Positives (FP): 936(non-cancer cases misclassified as cancer).

False Negatives (FN): 840 (missed cancer cases).

True Positives (TP): 7958 (correctly predicted cancer cases).

✓ **Prostate cancer (PRAD):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.802**	**0.792**	**0.844**	**0.817**

Table 29: Training results of the feed forward neural network model for predicting prostate cancer.

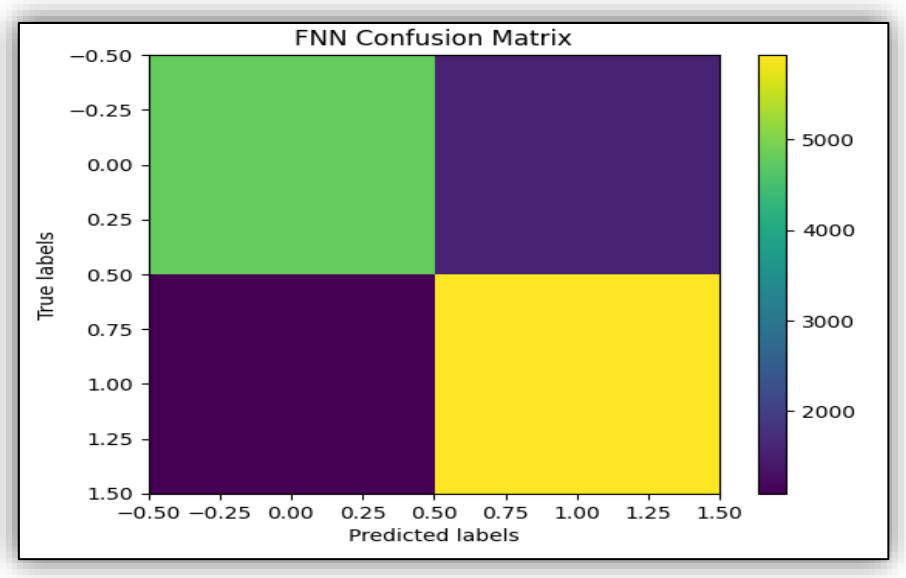

Figure 68: feed forward neural network confusion matrix (prostate cancer).

True Negatives (TN): 5931 (correctly predicted cases).

False Positives (FP): 1561 (misclassified cases).

False Negatives (FN): 1093 (missed cancer cases).

True Positives (TP): 4786 (correctly predicted cancer cases).

✓ **Melanoma (SKCM):**

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**0.874**	**0.885**	**0.869**	**0.877**

Table 30: Training results of the feed forward neural network model for predicting melanoma.

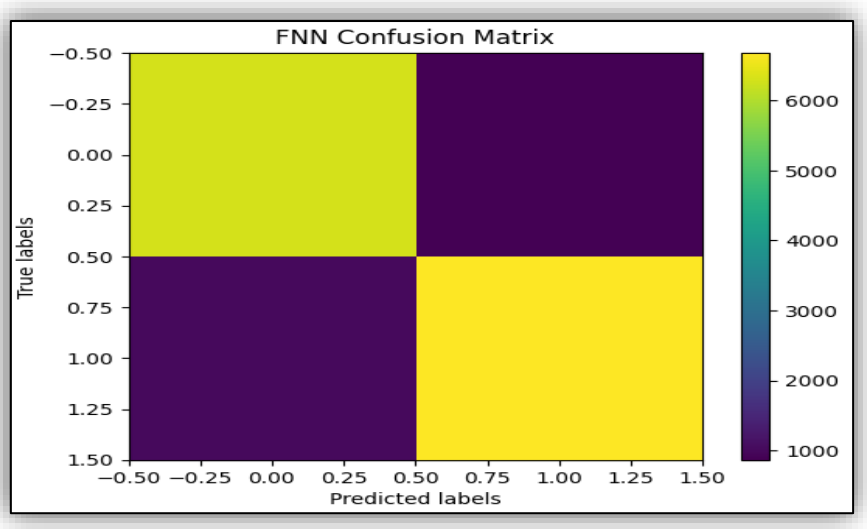

Figure 69: feed forward neural network confusion matrix (melanoma).

True Negatives (TN): 6310 (correctly predicted cases).

False Positives (FP): 864 (misclassified cases).

False Negatives (FN): 1007 (missed cases).

True Positives (TP): 6679 (correctly predicted cases).

Results interpretation:

Four binary classifiers were built to predict the presence of the seven types of cancer using gene expression data of circulating tumor cells (CTCs) and microemboli (CTMs). The classifiers used in this first experiment are Random Forest, Gradient Boosting Classifier, Decision Tree, and Feedforward Neural Network (FNN).

The Random Forest model shows great performance across all cancer types, with particularly high accuracy for liver (99.4%) cancer and colorectal cancer (97.7%). but the model's performance is slightly lower in the cases of melanoma (92.6%) and pancreatic cancer (92.5%). The confusion matrices reveal that the random forest classifier generally makes only a small number of false positives and false negatives.

The Gradient Boosting Classifier shows excellent performance across all cancer types, with a high accuracy that ranges between 93.4% and 99.5%. In addition, this model's confusion matrix reveals that it generally makes fewer false positives and false negatives, which is a good sign.

The Decision Tree classifier also performs well across all cancer types, with particularly high accuracy, precision, recall, and F1 scores for liver cancer and colorectal cancer. However, its confusion matrices reveal that the model generally makes more false positives and false negatives compared to the Random Forest and Gradient Boosting Classifiers, suggesting that this model has a lower performance than the first ones used.

The Feed forward Neural Network (FNN) also shows a very good performance for liver cancer (97.5% accuracy) and colorectal cancer (accuracy of 92%); however, this performance gets lower for all the other cancer types, with an accuracy of (89.3%) for pancreatic cancer, (87.4%) for melanoma, (86.5%) for breast cancer, (82.7%) for non-small cell lung cancer, and only (80%) for prostate cancer.

Experiment 2: Multiclass classification

Part one: building the multiclass classifiers:

In the second experiment, the selected models were used as multiclass classifiers to predict the presence of the seven cancer types.

The model	Parameters	Configuration
Random Forest	**n_estimators**	**100**
	random_state	**42**
Gradient Boosting	**n_estimators**	**100**
	max_depth	**3**
	min_samples_split	**2**
	learning_rate	**0.1**
	loss	**log_loss**
Decision Tree	**criterion**	**gini**
	max_depth	**none**
	min_samples_split	**2**
	min_samples_leaf	**1**

Feed forward Neural Network	**Dense Layers**	**1st**	**neurons**	**64**
			function	**ReLU**
		2nd	**neurons**	**32**
			function	**ReLU**
	Output Layer	**neurons**		**7**
		function		**sigmoid**
	compilation	**optimizer**		**adam**
		loss		**Categorical cross-entropy**

Table 31: The model's configurations.

1- Splitting the data:

For the machine learning classifiers, the data was split into 80% for training and 20% for testing.

```
# Assuming you have already loaded your data into X_res and y_res
X_train, X_test, y_train, y_test = train_test_split(X_res, y_res, test_size=0.2, random_state=42)
```

For the deep learning model, the data was divided into 80% for training, 10% for testing, and another 10% for validation.

```
X_train, X_temp, y_train, y_temp = train_test_split(X_res, y_res, test_size=0.2, random_state=42)
X_val, X_test, y_val, y_test = train_test_split(X_temp, y_temp, test_size=0.5, random_state=42)
```

2- The results:

2-1- Random Forest:

The results below are obtained after using the Random Forest model:

	Precision	Recall	F1 score
liver cancer	**1.00**	**1.00**	**1.00**
colorectal cancer	**1.00**	**1.00**	**1.00**
Non-small-cell-lung cancer	**1.00**	**1.00**	**1.00**
Melanoma	**1.00**	**1.00**	**1.00**
pancreatic cancer	**1.00**	**1.00**	**1.00**
prostate cancer	**1.00**	**1.00**	**1.00**
Breast cancer	**1.00**	**1.00**	**1.00**

Table 32: Training results of the Random Forest model for predicting the seven cancer types.

Model's Accuracy: 100%

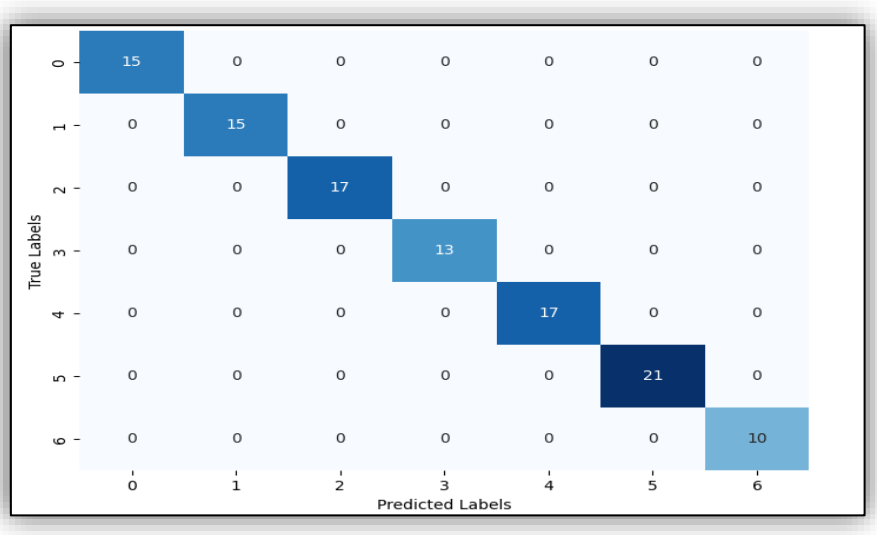

Figure 70: Random Forest confusion matrix (multiclass classification).

2-2- Gradient Boosting:

The results below are obtained after using the Gradient Boosting model:

	Precision	Recall	F1 score
liver cancer	**1.00**	**0.93**	**0.97**
colorectal cancer	**1.00**	**1.00**	**1.00**
Non-small-cell-lung cancer	**1.00**	**1.00**	**1.00**
Melanoma	**1.00**	**1.00**	**1.00**

pancreatic cancer	**1.00**	**1.00**	**1.00**
prostate cancer	**1.00**	**1.00**	**1.00**
Breast cancer	**0.91**	**1.00**	**0.95**

Table 33: Training results of the Gradient Boosting model for predicting the seven cancer types.

Model's Accuracy: 99%

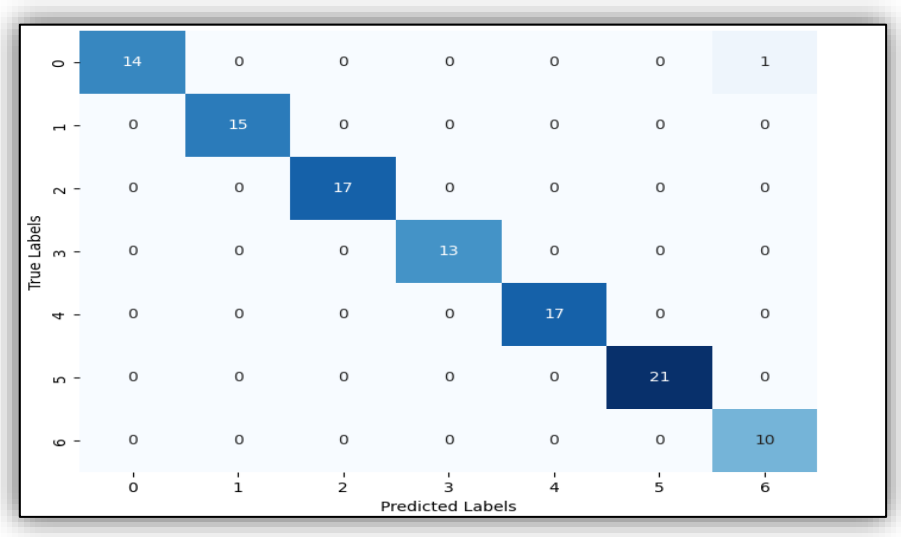

Figure 71: gradient boosting confusion matrix (multiclass classification).

2-3- Decision Tree:

The results below are obtained after using the Decision Tree model:

	Precision	Recall	F1 score
liver cancer	**1.00**	**0.93**	**0.97**
colorectal cancer	**1.00**	**0.87**	**0.93**
Non-small-cell-lung cancer	**1.00**	**1.00**	**1.00**
Melanoma	**1.00**	**1.00**	**1.00**
pancreatic cancer	**0.94**	**1.00**	**0.97**
prostate cancer	**1.00**	**1.00**	**1.00**
Breast cancer	**0.83**	**1.00**	**0.91**

Table 34: Training results of the Decision Tree model for predicting the seven cancer types.

Model's Accuracy: 97.2%

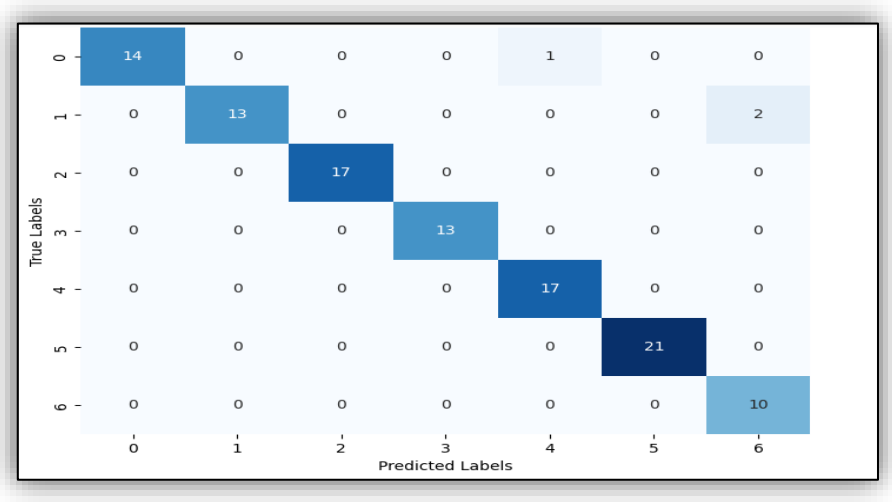

Figure 72: Decision Tree confusion matrix (multiclass classification).

2-4- Feed forward neural network:

The results below are obtained after using the Feedforward Neural Network model:

	Precision	Recall	F1 score
liver cancer	**1.00**	**0.50**	**0.67**
colorectal cancer	**1.00**	**1.00**	**1.00**
Non-small-cell-lung cancer	**1.00**	**1.00**	**1.00**
Melanoma	**1.00**	**1.00**	**1.00**
pancreatic cancer	**1.00**	**1.00**	**1.00**
prostate cancer	**0.76**	**1.00**	**0.87**
Breast cancer	**1.00**	**0.75**	**0.86**

Table 35: Training results of the Feed forward Neural Network model for predicting the seven cancer types.

Model's Accuracy: 93%

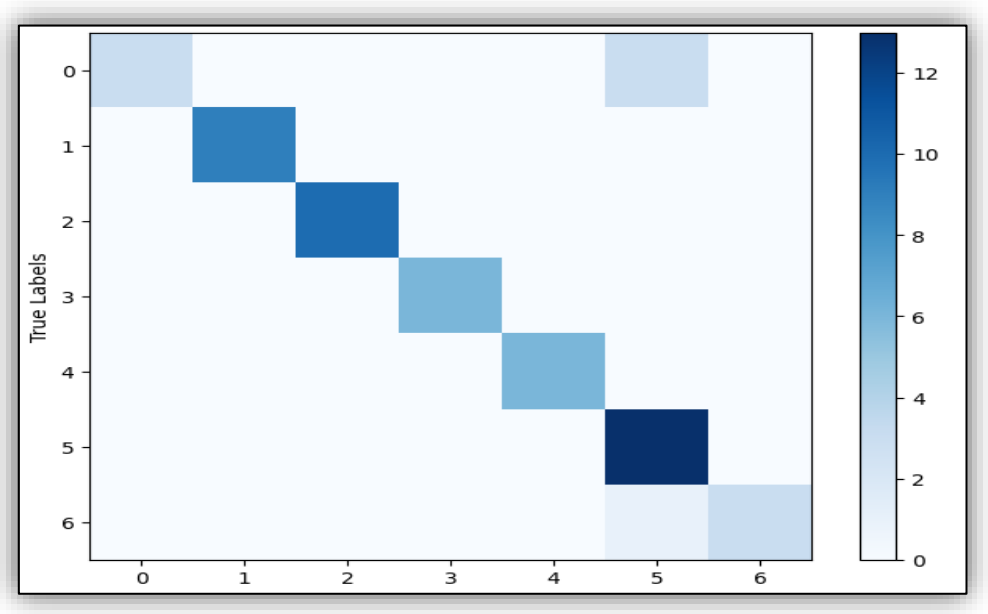

Figure 73: feed forward neural network confusion matrix (multiclass classification).

Part two: Evaluating the performance of Random Forest on unseen data:

In the first part of this experiment, Random Forest showed perfect results as a multiclass classifier. However, achieving such an accuracy of 100% on the training dataset can be both promising and concerning. It raises concerns about overfitting and the model's performance on unseen data. To address this, I evaluated its performance using two independent cancer datasets: pancreatic cancer and breast cancer. The results are shown below.

	Accuracy	Precision	Recall	F1 score
Model Effectiveness	**1.00**	**1.00**	**1.00**	**1.00**

Table 36: Evaluation Results of the Random Forest Model for Predicting pancreatic and breast Cancer (Unseen Data).

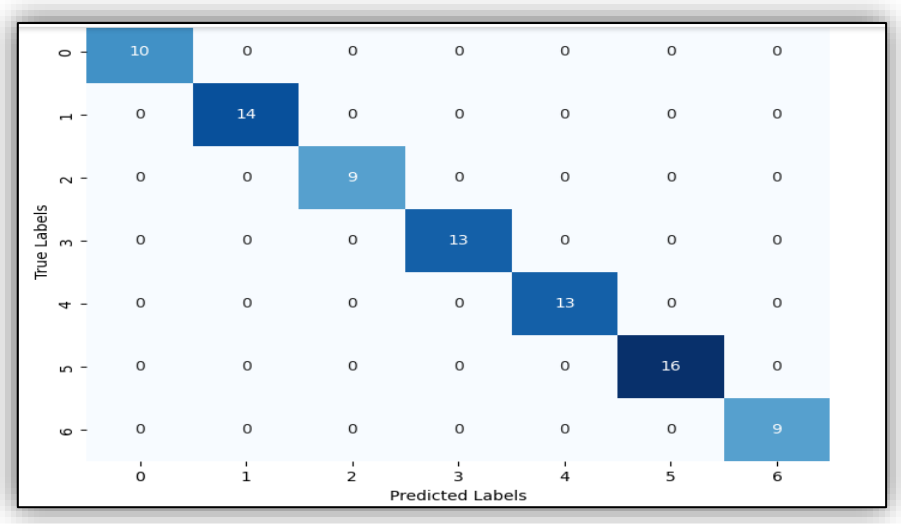

Figure 74: Random Forest Model confusion matrix for Predicting pancreatic and breast Cancer (Unseen Data).

Results interpretation:

In this experiment, Random Forest, Gradient Boosting Classifier, Decision Tree, and Feed forward Neural Network (FNN) were used as multiclass classifiers.

The Feed forward Neural Network (FNN) presents strong performance for several cancer types, achieving high precision, recall, and F1-scores for most classes. However, it exhibits some weaknesses, particularly in two classes (liver and breast cancer) where the recall drops to 0.50 and 0.75respectively. This results in a lower overall accuracy of 93%. The confusion matrix shows more misclassifications compared to the other models, indicating that while the FNN is competent, it is less consistent in its predictions and prone to more errors.

The Decision Tree classifier shows a high level of performance with an accuracy of 97%. It maintains high precision, recall, and F1-scores across most classes, though there are minor deviations. The confusion matrix reveals a few instances of misclassification, such as false positives and false negatives, suggesting that while the Decision Tree is effective, it is slightly less reliable than the Random Forest and Gradient Boosting Classifiers.

The Gradient Boosting Classifier also performs exceptionally well, with an overall accuracy of 99%. It achieves precision, recall, and F1-scores of 1.00 for most cancer types, although there is a slight drop in recall for one class, which leads to a small number of false

negatives. This model's confusion matrix indicates very few misclassifications, demonstrating its strong performance in distinguishing between different cancer types.

The Random Forest classifier demonstrates perfect performance across all cancer types, achieving an accuracy of 100%. This is reflected in its precision, recall, and F1-score, all of which are 1.00 for every class. The confusion matrix shows no misclassifications, indicating that the model has an excellent ability to correctly identify all instances of each cancer type without any false positives or false negatives.

While the perfect results on the training dataset might raise concerns about overfitting, the model's performance was further evaluated using two independent cancer datasets for pancreatic and breast cancer. Remarkably, the Random Forest model maintained its 100% accuracy, precision, recall, and F1-score on these independent datasets as well, demonstrating robust generalization capabilities. This indicates that the model effectively learned relevant patterns from the training data and performs reliably on unseen data, alleviating initial concerns about overfitting.

3- Future work:

Future research in predicting the presence of cancer types using gene expression data of circulating tumor cells (CTCs) and microemboli (CTMs) could involve developing and training the Random Forest model to predict more cancer types. Also, deep learning models can be combined with gene expression profiling for further improved accuracy in classification. Moreover, the identification of molecular markers for different types of cancer using CTCs could yield very useful insights into early detection and individualized treatment strategies. Further, expression profiling of driver cancer genes and immunotherapeutic targets in CTCs and peripheral blood mononuclear cells could provide data on a patient's response to therapies in real time, especially immunotherapy, contributing to predicting tumor progression and patient outcomes and to selecting appropriate treatment.

Conclusion:

In this chapter, I developed and evaluated machine learning models to predict seven cancer types (liver cancer, breast cancer, colorectal cancer, non-small cell lung cancer, pancreatic cancer, prostate cancer, and melanoma) using gene expression data from circulating tumor cells (CTCs) and microemboli (CTMs). I built four binary classifiers—Random Forest, Gradient Boosting Classifier, Decision Tree, and Feedforward Neural Network (FNN)—and then adapted these models to function as multiclass classifiers.

This work demonstrates the effectiveness of these machine learning models, particularly the Random Forest classifier, in predicting cancer types using gene expression data from CTCs and CTMs.

For binary classification, the Random Forest model showed excellent performance, especially for liver (99.4%) and colorectal (97.7%) cancers, with slightly lower accuracy for melanoma (92.6%) and pancreatic cancer (92.5%). The Gradient Boosting Classifier also performed well, with accuracy ranging from 93.4% to 99.5%, and fewer misclassifications. The Decision Tree classifier was effective, particularly for liver and colorectal cancers, but had more false positives and negatives. The FNN performed well for liver (97.5%) and colorectal (92%) cancers but had lower accuracy for the other types, such as pancreatic cancer (89.3%) and prostate cancer (80%).

In the multiclass classification, the Random Forest classifier achieved perfect performance with 100% accuracy, precision, recall, and F1-scores. While there were initial concerns about overfitting, the model maintained its performance on independent datasets for pancreatic and breast cancer, demonstrating robust generalization capabilities. The Gradient Boosting Classifier also had high accuracy (99%) with few misclassifications. The Decision Tree showed high precision, recall, and F1-scores with an accuracy of 97%, though it had minor misclassifications. The FNN achieved strong results for most cancer types but showed weaknesses in liver and breast cancers, resulting in a lower overall accuracy of 93%.

General conclusion

General conclusion:

This thesis aimed to develop and evaluate machine learning models for predicting the presence of seven cancer types (liver cancer, breast cancer, colorectal cancer, non-small cell lung cancer, pancreatic cancer, prostate cancer, and melanoma) using gene expression data from circulating tumor cells (CTCs) and microemboli (CTMs). The primary goal was to compare the two approaches of binary and multiclass classification to determine the best approach for cancer prediction using this data and then to identify the most effective model for accurate cancer prediction that can be applied in real-world scenarios.

Initially, binary classifiers were constructed using four different algorithms: Random Forest, Gradient Boosting Classifier, Decision Tree, and Feedforward Neural Network (FNN). These models were trained to distinguish one cancer type from all others combined. The performance evaluation revealed that the Random Forest and Gradient Boosting Classifiers provided the highest accuracy and the lowest misclassification rates. The Gradient Boosting model, in particular, demonstrated outstanding performance, with a high accuracy that ranges between 93.4% and 99.5%.

Subsequently, the same classifiers were adapted for multiclass classification to predict all seven cancer types simultaneously. The Random Forest model again excelled, achieving perfect performance with 100% accuracy, precision, recall, and F1-scores. The Gradient Boosting Classifier also performed exceptionally well, with an overall accuracy of 99%. The Decision Tree and FNN showed good performance, but with higher rates of misclassification compared to the Random Forest and Gradient Boosting Classifiers.

A critical concern of overfitting was addressed by testing the Random Forest model on independent datasets for pancreatic and breast cancer. Remarkably, the model maintained its perfect performance, indicating robust generalization capabilities and reinforcing its suitability for practical applications.

Overall, this research highlights the significant potential of machine learning models in cancer prediction using gene expression data from CTCs and CTMs. Among the models evaluated, the Random Forest multi-classifier emerged as the most reliable and effective, making it highly recommended for practical use in cancer diagnosis.

Bibliography

Bibliography

[1] Beghoul S. (2008). Les Inhibiteurs de topo isomerases I en cancerologie. Mise en evidence d'un nouveau mecanisme d'action pro-apoptotique de la camptothecine. These doctorat.

[2] Florence Coussy, Florian Bonin, Paula Azorin, Zakia Tariq and Keltouma Driouchm (2019). Biology of metastases and molecular mechanisms of their formation, published by Elsevier. Volume 106, numéro 1, janvier 2019, pages 24-36. https://www.sciencedirect.com/science/article/pii/S0007455118303667

[3] Haymour, L. (2022). Study of the expression and role of PAMR1 in colorectal cancer. (Doctoral dissertation). Université de Limoges

[4] Senga, S. S., & Grose, R. P. (2020). hallmarks of cancer - the new testament. Open Biol. 11: 200358. https://doi.org/10.1098/rsob.20.0358

[5] Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. International journal of cancer, 10.1002/ijc.33588. Advance online publication. https://doi.org/10.1002/ijc.33588

[6] Lee, S. H. Z. S. (2015). Organometallic compounds of osmium for cancer therapy (Doctoral dissertation). Nanyang Technological University; Ecole Polytechnique.

[7] Pecorino, L. (2021). Molecular biology of cancer: mechanisms, targets, and therapeutics (Fifth edition). Oxford University Press.

[8] Zhukova, L. G., Andreeva, I. I., Zavalishina, L. E., Zakiriakhodzhaev, A. D., Koroleva, I. A., Nazarenko, A. V., Paltuev, R. M., Parokonnaia, A. A., Petrovskii, A. V., Portnoi, S. M., Semiglazov, V. F., Semiglazova, T. I., Stenina, M. B., Stepanova, A. M., Trofimova, O. P., Tyulyandin, S. A., Frank, G. A., Frolova, M. A., Shatova, I. S., Nevol'skikh, A. A., Ivanov, S. A., Khailova, Z. V., & Gevorkian, T. G. (2021). Breast cancer. Journal Of Modern Oncology, 23(1), 5-40. doi: 10.26442/18151434.2021.1.200823

[09] Harris, J. R., Lippman, M. E., Veronesi, U., & Willett, W.. (1992). Breast Cancer. The New England Journal of Medicine. Publié le 6 août 1992/N Engl J Med 1992 ; 327 : 390 – 398. VOL. 327 NON. 6. DOI : 10.1056/NEJM199208063270606

[10] Veronesi, U., Boyle, P., Goldhirsch, A., Orecchia, R., & Viale, G.. (2005). Breast cancer. 365(9472). DOI : 10.1016/S0140-6736(05)66546-4

[11] Harbeck. Nadia, Penault-Liorca. Frédérique, Cardoso. Fatima (2019). Breast cancer. Nature Reviews Disease Primers 5, Article number: 67. DOI : 10.1038/s41572-019-0122-z

[12] Alshammari, F. D.. (2019). Breast cancer genetic susceptibility: With focus in Saudi Arabia. 6–12. Journal of Oncological Sciences. Volume 5, Issue 1, April 2019, Pages 6-12**.**

https://doi.org/10.1016/j.jons.2019.02.001

[13] Bennett C, Carroll C, Wright C, Awad B, Park JM, Farmer M, Brown E, Heatherly A, Woodard S (2022). Génomique du cancer du sein : métastases primaires et les plus courantes. *Cancers* . 2022 ; 14(13):3046. https://doi.org/10.3390/cancers14133046

[14] Chunfang, Hao., Chen, Wang., Ning, Lu., Weipeng, Zhao., Shufen, Li., Lei, Zhang., Wenjing, Meng., Shuling, Wang., Zhongsheng, Tong., Y., C., Zeng., Leilei, Lu. (2022). Gene Mutations Associated With Clinical Characteristics in the Tumors of Patients With Breast Cancer. Frontiers in Oncology, doi: 10.3389/fonc.2022.778511

[15] Nathan, F., Schachter., Jessica, R., Adams., Patryk, Skowron., Katelyn., J., Kozma., Christian, A., Lee., Christian, A., Lee., Nandini, Raghuram., Joanna, Yang., Amanda, J., Loch., Wei, Wang., Aaron, Kucharczuk., Katherine, L., Wright., Rita, M., Quintana., Yeji, An., Daniel, Dotzko., Jennifer, L., Gorman., Daria, Wojtal., Juhi, S., Shah., Paul, Leon-Gomez., Giovanna, Pellecchia., Adam, J., Dupuy., Charles, M., Perou., Ittai, Ben-Porath., Rotem, Karni., Eldad, Zacksenhaus., Eldad, Zacksenhaus., James, R., Woodgett., James, R., Woodgett., Susan, J., Done., Livia, Garzia., Livia, Garzia., A., Sorana, Morrissy., A., Sorana, Morrissy., Jüri, Reimand., Jüri, Reimand., Michael, D., Taylor., Sean, E., Egan. (2021). Single allele loss-of-function mutations select and sculpt conditional cooperative networks in breast cancer.. Nature Communications, doi: 10.1038/s41467-021-25467-w

[16] Ding HJ , Zhao YP, Jiang ZC, Zhou DT, Zhu R. (2022). Analysis of Mitochondrial Transfer RNA Mutations in Breast Cancer. Balkan Journal of Medical Genetics,

doi: 10.2478/bjmg-2022-0020

[17] Mariem Ben Rekaya., Farah Sassi., Linda Bel Haj Kacem., Nada Mansouri., Sinda Zarrouk., Saifeddine Azouz., Soumaya Rammeh. (2023). PIK3CA mutations in breast cancer: A Tunisian series. PLOS ONE, doi: 10.1371/journal.pone.0285413

[18] Miller, K. D., Camp, M., Steligo, K. (2021). The Breast Cancer Book: A Trusted Guide for You and Your Loved Ones. United States: Johns Hopkins University Press.

[19] Kuk, K. (1986). Colorectal cancer. The Journal of the American Osteopathic Association, 86(1), 101-110. https://doi.org/10.1515/jom-1986-860122

[20] Mattiuzzi C, Sanchis-Gomar F, Lippi G (2019). Concise update on colorectal cancer epidemiology. Ann Transl Med 2019.7(21):609. doi: 10.21037/atm.2019.07.91

[21] Richard, Barfield., Conghui, Qu., Robert, S., Steinfelder., Chenjie, Zeng., Tabitha, A., Harrison., Stefanie, Brezina., C., D., Buchanan., Peter, T., Campbell., Graham, Casey., Steven, Gallinger., Marios, Giannakis., Stephen, B., Gruber., Andrea, Gsur., Li, Hsu., Jeroen, R., Huyghe., Victor, Moreno., Polly, A., Newcomb., Shuji, Ogino., Amanda, I., Phipps., Martha, L., Slattery., Stephen, N., Thibodeau., Quang, M., Trinh., Amanda, E., Toland., Thomas, J., Hudson., Wei, Sun., Syed, H.E., Zaidi., Ulrike, Peters. (2022). Association between germline variants and somatic mutations in colorectal cancer. Dental science reports. doi: 10.1038/s41598-022-14408-2

[22] Ian Tomlinson;Alex Cornish;Andreas Gruber; Richard Houlston; Amit Sud; Philip Law; Eszter Lakatos;Richard Culliford; Jacob Househam; Trevor Graham; Henry Wood; Philip Quirke;Nuria Lopez-Bigas; Claudia Arnedo-Pac;Daniel Chubb;Maire Ni Leathlobhair;Boris Noyvert; Ben Kinnersley; William Cross;Nirupa Murugaesu;Alona Sosinsky;Jonathan Mitchell; Ludmil Alexandrov;Luis Zapata;Juan Fernandez-Tajes;Steve Thorn;Kitty Sherwood;Guler Gul;Aliah Hawari; Andrea Sottoriva;David Church; Giulio Caravagna; David Wedge;Anna Frangou (2022). Whole genome sequencing of 2,023 colorectal cancers reveals mutational landscapes, new driver genes and immune interactions. https://doi.org/10.21203/rs.3.rs-2273265/v1

[23] Julia Matas; Brendan Kohrn; Jeanne Fredrickson; Kelly Carter; Ming Yu; Ting Wang; Xianyong Gui; Thierry Soussi; Victor Moreno; William M. Grady et al. (2023). Data from Colorectal Cancer Is Associated with the Presence of Cancer Driver Mutations in Normal Colon. DOI: 10.1158/0008-5472.c.6513700.v1

[24] Alex, J., Cornish., Andreas, Gruber., Ben, Kinnersley., Daniel, Chubb., Anna, Frangou., Giulio, Caravagna., Boris, Noyvert., Eszter, Lakatos., Henry, M., Wood., Claudia, Arnedo-Pac., Richard, Culliford., Jacob, Househam., William, Cross., Amit, Sud., Philip, J., Law., Máire, Ní, Leathlobhair., Aliah, Hazmah, Hawari., Steve, Thorn.,

Bibliography

Kitty, Sherwood., Güler, Gül., Juan, Fernández-Tajes., Luis, Zapata., Ludmil, B., Alexandrov., Nirupa, Murugaesu., Alona, Sosinsky., Jonathan, S., Mitchell., Nuria, Lopez-Bigas., Philip, Quirke., David, N., Church., Ian, Tomlinson., Andrea, Sottoriva., Trevor, A., Graham., David, C., Wedge., Richard, S., Houlston. (2022). Whole genome sequencing of 2,023 colorectal cancers reveals mutational landscapes, new driver genes and immune interactions. bioRxiv, https://doi.org/10.1101/2022.11.16.515599

[25] Ionescu Adriana, Bilteanu Liviu, Ionut Geicu Ovidiu Iordache, Florin, Stanca Loredana, Pisoschi Aurelia Magdalena, Miron Adrian, Iren Serban Andreea and Calu Valentin (2022). RAS, BRAF and EGFR related genetic mutations as predictive biomarkers in colorectal cancer**.** Preprints (Posted 5 May 2022). doi : 10.20944/preprints202205.0042.v1

[26] Ili, C., Buchegger, K., Demond, H., Castillo-Fernandez, J., Kelsey, G., Zanella, L., Abanto, M., Riquelme, I., López, J., & Viscarra, T. (2020). Landscape of Genome-Wide DNA Methylation of Colorectal Cancer Metastasis. https://doi.org/10.3390/cancers12092710

[27] Poturnajova, M., Furielova, T., Balintova, S., Schmidtova, S., Kucerova, L., & Matuskova, M.. (2021). Molecular features and gene expression signature of metastatic colorectal cancer (Review). https://doi.org/10.3892/or.2021.7961

[28] Afrăsânie,V., Marinca,M., AlexaStratulat,T., Gafton,B., Păduraru,M., Adavidoaiei,A., Miron,L. & Rusu,C. (2019). KRAS, NRAS, BRAF, HER2 and microsatellite instability in metastatic colorectal cancer – practical implications for the clinician. Radiology and Oncology,53(3) 265-274. https://doi.org/10.2478/raon-2019-0033

[29] Colorectal cancer Causes, Symptoms & Treatment -Clicks Health Hub. (2024, July 10). Colorectal cancer Causes, Symptoms & Treatment -Clicks Health Hub. https://clicks.co.za/health/conditions/article-view/colorectal-cancer

[30] Ugurel, S., & Gutzmer, R.. (2023). Melanom. https://doi.org/10.1111/ddg.15053

[31] Tuong, W., Cheng, L., & Armstrong, A.. (2012). Melanoma: Epidemiology, Diagnosis, Treatment, and Outcomes. 30(1), 113–124. https://doi.org/10.1016/j.det.2011.08.006

[32] Andrey, Toropovskiy., Alexei, G., Nikitin., A., V., Solov'ev., R., M., Khuzina., O., N., Pavlova. (2022). Identifying a wide range of mutations in the BRAF gene for prescribing

targeted drugs for melanoma treatment. Vestnik medicinskogo instituta "REAVIZ": reabilitaciâ, vrač i zdorov′e, doi: 10.20340/vmi-rvz.2023.1.clin.4

[33] Alba, Loras., Marta, Gil-Barrachina., María, Ángeles, Marqués-Torrejón., Gemma, Pérez-Pastor., Conrado, Martinez-Cadenas. (2022). UV-Induced Somatic Mutations Driving Clonal Evolution in Healthy Skin, Nevus, and Cutaneous Melanoma. Reproductive and developmental Biology, doi: 10.3390/life12091339

[34] Yuchen, Guo., Yi, Chen., L., Zhang., Linglan, Ma., Keyu, Jiang., Gang, Yao., Lingjun, Zhu. (2022). TERT Promoter Mutations and Telomerase in Melanoma. Journal of Oncology, doi: 10.1155/2022/6300329

[35] NHS website (2023), March 6). Melanoma skin cancer. https://www.nhs.uk/conditions/melanoma-skin-cancer/symptoms/

[36] Chen, Z., Fillmore, C., Hammerman, P. et al (2014). Non-small-cell lung cancers: a heterogeneous set of diseases. Nat Reviews Cancer **14**, 535–546 (2014). DOI: 10.1038/nrc3775

[37] Alduais, Y. P., Zhang, H. P., Fan, F. P., Chen, J. P., & Chen, B. P.. (2023). Non-small cell lung cancer (NSCLC): A review of risk factors, diagnosis, and treatment. DOI : 10.1097/MD.0000000000032899

[38] Gridelli, C., Rossi, A., Carbone, D. & al (2015). Non-small-cell lung cancer. Nat Rev Dis Primers **1**, 15009 (2015). DOI : 10.1038/nrdp.2015.9

[39] Govindan, R., Ding, L., Griffith, M., Watson, M., Mardis, E. R., Wilson, R. K. & al (2012). Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. 150(6), P1121–1134. DOI : 10.1016/j.cell.2012.08.024

[40] Bankovic J, Stojsic J, Jovanovic D, Andjelkovic T, Milinkovic V, Ruzdijic S, Tanic N (2009). Identification of genes associated with non-small-cell lung cancer promotion and progression. Lung Cancer. 2010 Feb;67(2):151-9. DOI: 10.1016/j.lungcan.2009.04.010

[41] Ricciuti B, Elkrief A, Alessi J, Wang X, Li Y, Gupta H, Muldoon DM, Bertram AA, Pecci F, Lamberti G, Di Federico A, Barrichello A, Vaz VR, Gandhi M, Lee E, Shapiro GI, Park H, Nishino M, Lindsay J, Felt KD, Sharma B, Cherniack AD, Rodig S, Gomez DR, Shaverdian N, Rakaee M, Bandlamudi C, Ladanyi M, Janne PA, Schoenfeld AJ,

Bibliography

Sholl LM, Awad MM, Cheng ML (2023). Clinicopathologic, Genomic, and Immunophenotypic Landscape of ATM Mutations in Non-Small Cell Lung Cancer. Clin Cancer Res. 2023 Jul 5;29(13):2540-2550. doi: 10.1158/1078-0432.CCR-22-3413.

PMID: 37097610; PMCID: PMC11031845.

[42] Rosell R, Aguilar-Hernández A, González-Cao M (2023). Insights into *EGFR* Mutations and Oncogenic *KRAS* Mutations in Non-Small-Cell Lung Cancer. Cancers (Basel). 2023 Apr 28;15(9):2519. doi: 10.3390/cancers15092519

[43] Liu-sheng, Yang., Meng, Wang., Na, Li., Lubin, Yan., Wenwei, Zhou., Zhi, Yu., Xiaodong, Peng., Jun, Cai., Yonghua, Yang. (2023). TERT Mutations in Non–Small Cell Lung Cancer: Clinicopathologic Features and Prognostic Implications. Clinical Medicine: Oncology, doi: 10.1177/11795549221140781

[44] Kumar, Ashwini., Kumar, Awanish. (2022). Non-small-cell lung cancer-associated gene mutations and inhibitors. Advances in cancer biology- Metastasis. Volume 6, December 2022, 100076. https://doi.org/10.1016/j.adcanc.2022.100076

[45] Healthline Media LLC (2022, May 27). What Are the Different Types of Non-Small Cell Lung Cancer (NSCLC)?. https://www.healthline.com/health/lung-cancer/types-of-non-small-cell-lung-cancer#types

[46] Kamisawa T, Wood LD, Itoi T, Takaori K (2016). Pancreatic cancer. Lancet. 2016 Jul 2;388(10039):73-85. PMID : 26830752DOI : 10.1016/S0140-6736(16)00141-0

[47] Goral, V(2015). Pancreatic Cancer: Pathogenesis and Diagnosis. Asian Pac J Cancer Prev. 2015;16(14):5619-24. PMID : 26320426 DOI : 10.7314/apjcp.2015.16.14.5619

[48] B. Jagadeesan, P. Hari Haran, D. Praveen, P. Ranadheer Chowdary, M. Vijey Aanandhi (2021). A Comprehensive Review on Pancreatic Cancer. Research J. Pharm. and Tech. 2021; 14(1):552-554. DOI: 10.5958/0974-360X.2021.00100.1

[49] He Y, Huang W, Tang Y, Li Y, Peng X, Li J, Wu J, You N, Li L, Liu C, Zheng L, Huang X (2023). Clinical and genetic characteristics in pancreatic cancer from Chinese patients revealed by whole exome sequencing. Front Oncol. 2023 May 29;13:1167144. PMID: 37313463. doi: 10.3389/fonc.2023.1167144

[50] Liu J, Mroczek M, Mach A, Stępień M, Aplas A, Pronobis-Szczylik B, Bukowski S, Mielczarek M, Gajewska E, Topolski P, Król ZJ, Szyda J, Dobosz P (2023). Genetics, Genomics and Emerging Molecular Therapies of Pancreatic Cancer. Cancers (Basel). 2023 Jan 27;15(3):779. PMID: 36765737. DOI: 10.3390/cancers15030779

[51] Sato H, Sasaki K, Hara T, Tsuji Y, Arao Y, Otsuka C, Hamano Y, Ogita M, Kobayashi S, di Luccio E, Hirotsu T, Doki Y, Eguchi H, Satoh T, Uchida S, Ishii H (2022). Pancreatic Cancer Research beyond DNA Mutations. Biomolecules. 2022 Oct 17;12(10):1503. PMID: 36291712. DOI : 10.3390/biom12101503

[52] Nikolaos, Tsoulos., K., Potska., Konstantinos, Agiannitopoulos., Georgios, N., Tsaousis., Mustafa, Ozdogan., Bulent, Karabulut., Adina, E., Croitoru., Dimitrios, C., Ziogas., Maria, Theochari., Christos, Christodoulou., Epaminontas, Samantas., D., Janinis., Ilias, Athanasiadis., Angelos, Koutras., Eirini, Papadopoulou., G, Nasioulas. (2023). Genetic testing in patients with pancreatic cancer to reveal pathogenic variants in cancer susceptibility genes.. Journal of Clinical Oncology Volume 41, Number 16_suppl https://doi.org/10.1200/JCO.2023.41.16_suppl.e16276

[53] Hanada K, Amano H, Abe T (2017). Early diagnosis of pancreatic cancer: Current trends and concerns. Ann Gastroenterol Surg. 2017 Apr 25;1(1):44-51. PMID: 29863166. doi: 10.1002/ags3.12004

[54] Prostate Cancer: Detection & Screening VII (MP75). (2020). *Journal of Urology*, *203*(Supplement 4). https://doi.org/10.1097/JU.0000000000000961 (Original work published April 1, 2020)

[55] Davey, P., Sprigings, D., & Ajzensztejn, D.. (2018). Prostate cancer. https://doi.org/10.1093/med/9780199568741.003.0326

[56] Karolina Sienkiewicz MSc, Chunsong Yang PhD, Bryce M. Paschal PhD, Aakrosh Ratan PhD (2022). Genomic analyses of the metastasis-derived prostate cancer cell lines LNCaP, VCaP, and PC3-AR. Original article: The prostate. Volume 82 Issue 4 March 1, 2022. Pages 442- 452. https://doi.org/10.1002/pros.24290

[57] Roylance R, Spurr N, Sheer D (1997). The genetic analysis of prostate carcinoma. Semin Cancer Biol. 1997 Feb;8(1):37-44. PMID: 9299580. DOI: 10.1006/scbi.1997.0051

[58] Brothman AR (2002). Cytogenetics and molecular genetics of cancer of the prostate. Am J Med Genet. 2002 Oct 30;115(3):150-6. PMID: 12407695. DOI: 10.1002/ajmg.10694

[59] Edwards S, Meitz J, Evans C, Easton D, Hopper GG, Foulkes WD, Narod S, Simard J, Badzoich M, Maehle L, PI Eeles R (2003a): Results of a genome-wide linkage analysis in prostate cancer families ascertained through the ACTANE consortium. The international ACTANE consortium. Prostate. 57(4):270–279. https://doi.org/10.1002/pros.10301

[60] Ahmed, Elshafei., Mohammed, Al-Toubat., Allison, H., Feibus., Kashyap, Koul., Seyed, Behzad, Jazayeri., Navid, Lelani., Valencia, Henry., K.C., Balaji. (2023). Genetic mutations in smoking-associated prostate cancer.. The Prostate PMID : 37455402. DOI : 10.1002/pros.24554

[61] Zhan Y, Ruan X, Liu J, Huang D, Huang J, Huang J, Chun TTS, Ng AT, Wu Y, Wei G, Jiang H, Xu D, Na R (2023). Genetic Polymorphisms of the Telomerase Reverse Transcriptase Gene in Relation to Prostate Tumorigenesis, Aggressiveness and Mortality: A Cross-Ancestry Analysis. Cancers (Basel). 2023 May 8;15(9):2650. PMCID: PMC10177366 DOI: 10.3390/cancers15092650

[62] Gvantsa Kharaishvili, Mariam Kacheishvili, Giorgi Akhvlediani (2022). BRCA Gene Mutations and Prostate Cancer. In book: BRCA1 and BRCA2 Mutations - Diagnostic and Therapeutic Implications [Working Title]. DOI:10.5772/intechopen.108792

[63] Shridhar, C, Ghagane., Rajendra, B, Nerli. (2023). Prostate Cancer: Germline Mutations in BRCA1 and BRCA2. March 2023. Asian Pacific Journal of Cancer Biology 8(1):69-73. DOI:10.31557/apjcb.2023.8.1.69-73

[64] Vasioukhin V(2004). Hepsin paradox reveals unexpected complexity of metastatic process. Cell Cycle. 2004 Nov;3(11):1394-7. Epub 2004 Nov 28. PMID: 15539945. DOI : 10.4161/cc.3.11.1273

[65] A. I. Sherifova, A. M. Parsadanyan (2023). Predictors of Liver Cancer: a Review. Creative surgery and oncology, 2023, № 3, p. 229-237. https://doi.org/10.24060/2076-3093-2023-13-3-229-237

[66] Sergi CM (2021). Carcinoma of the Liver in Children and Adolescents. Brisbane (AU): Exon Publications; 2021 Apr 6. https://doi.org/10.36255/exonpublications.livercancer.2021

[67] Amna, Amin, Sethi., Nisar, A., Shar. (2023). Identification of Liver Cancer Driver Mutations from COSMIC Data. International Journal of Cancer Management, Volume:16 Issue: 1, Dec 2023. DOI : https://doi.org/10.5812/ijcm-131281

[68] L.S.S., Srivani, Nagam., Ramakrishna, Vadde., Rajeswari, Jinka. (2022). Polymorphisms in hepatocellular carcinoma. Theranostics and Precision Medicine for the Management of Hepatocellular Carcinoma, 2022, p. 125-133. doi: 10.1016/b978-0-323-98806-3.00013-1

[69] Magdalena, Śmiech., Pawel, Leszczynski., Christopher, P., Wardell., Piotr, Poznański., Mariusz, Pierzchała., Hiroaki, Taniguchi. (2022). Oncogenic Mutation BRAF V600E Changes Phenotypic Behavior of THLE-2 Liver Cells through Alteration of Gene Expression. International Journal of Molecular Sciences. 23(3):1548. DOI:10.3390/ijms23031548

[70] Perez S, Lavi-Itzkovitz A, Gidoni M, Domovitz T, Dabour R, Khurana I, Davidovich A, Tobar A, Livoff A, Solomonov E, Maman Y, El-Osta A, Tsai Y, Yu ML, Stemmer SM, Haviv I, Yaari G, Gal-Tanamy M (2023). High-Resolution Genomic Profiling of Liver Cancer Links Etiology With Mutation and Epigenetic Signatures. Cell Mol Gastroenterol Hepatol. 2023;16(1):63-81. Epub 2023 Mar 24. PMID: 36965814; PMCID: PMC10212990. DOI: 10.1016/j.jcmgh.2023.03.004

[71] Cui Y, Li H, Zhan H, Han T, Dong Y, Tian C, Guo Y, Yan F, Dai D, Liu P (2021). Identification of Potential Biomarkers for Liver Cancer Through Gene Mutation and Clinical Characteristics. Front Oncol. 2021 Sep 17;11:733478. PMID: 34604069; PMCID: PMC8484954. doi: 10.3389/fonc.2021.733478

[72] Ozen, C., Yildiz, G., Dagcan, A. T., Cevik, D., Ors, A., Keles, U., Topel, H., & Ozturk, M.. (2013). Genetics and epigenetics of liver cancer. 30(4), 381–384. https://doi.org/10.1016/j.nbt.2013.01.007

[73] Zemin Zhang (2012). Genomic landscape of liver cancer. Nat Genet. 2012 Oct;44(10):1075-7. PMID: 23011223DOI: 10.1038/ng.2412

[74] Dragani, T. A., Manenti, G., Gariboldi, M., De Gregorio, L., & Pierotti, M. A..(1995). Genetics of liver tumor susceptibility in mice. 82-83, 613–619.

https://doi.org/10.1016/0378-4274(95)03505-2

[75] Ding, S.-F., & Habib, N. A.. (1995). Loss of heterozygosity in liver tumours. 22(2), 230–238. https://doi.org/10.1016/0168-8278(95)80434-X

[76] CAR-T therapy & Stem Cells for Liver Cancer – Hepatocellular Carcinoma. (2024, April 8). Regeneration Center of Thailand. https://stemcellthailand.org/oncology/liver-cancer-treatment

[77] Douglas, Hanahan., Robert, A., Weinberg. (2017). Biological Hallmarks of Cancer. WILEY online library. https://doi.org/10.1002/9781119000822.hfcm002

[78] Werner H, LeRoith D (2022). Hallmarks of cancer: The insulin-like growth factors perspective. Front Oncol. 2022 Nov 21;12:1055589. PMID: 36479090; PMCID: PMC9720135. DOI: 10.3389/fonc.2022.1055589

[79] Douglas Hanahan (2022). Hallmarks of cancer: the newer proportions. Cancer Discov (2022) 12 (1): 31–46. AACR Journals. Volume 12 Issue 1 (January 2022). https://doi.org/10.1158/2159-8290.CD-21-1059

[80] Jacqueline J. Chu, Raman Mehrzad (2023). The biology of cancer. The Link Between Obesity and Cancer. Pages 35-45. https://doi.org/10.1016/B978-0-323-90965-5.00012-X

[81] Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. Cell, 144(5), 646–674. PMID: 21376230. DOI: 10.1016/j.cell.2011.02.013

[82] Castaneda, M., den Hollander, P., Kuburich, N. A., Rosen, J. M., & Mani, S. A. (2022). Mechanisms of cancer metastasis. Seminars in cancer biology. Volume 87, p17–31. https://doi.org/10.1016/j.semcancer.2022.10.006

[83] Majidpoor, J., & Mortezaee, K. (2021). Steps in metastasis: an updated review. Medical oncology (Northwood, London, England), 38(1), 3.

https://doi.org/10.1007/s12032-020-01447-w

[84] Zhen Sun (2017). Mechanisms of Tenascin-C dependent tumor migration and metastasis. Cancer. Université de Strasbourg, 2017. English. ⟨NNT : 2017STRAJ038⟩.

[85] Tianyue, Fan., Guicheng, Kuang., Runmin, Long., Junwei, Han., Jing, Wang. (2022). The overall process of metastasis: From initiation to a new tumor.. Biochimica Et Biophysica Acta - Reviews On Cancer, PMID: 35728735. DOI: 10.1016/j.bbcan.2022.188750

[86] Jamal, Majidpoor., Keywan, Mortezaee. (2021). Steps in metastasis: an updated review.. Medical Oncology, PMID : 33394200. DOI : 10.1007/s12032-020-01447-w

[87] M., Shibata., Kohei, Taniguchi. (2023). Metastasis Inhibition. International Journal of Molecular Sciences. 2023, 24(8), 7123. **https://doi.org/10.3390/ijms24087123**

[88] van Zijl F, Krupitza G, Mikulits W (2011). Initial steps of metastasis: cell invasion and endothelial transmigration. Mutat Res. 2011 Jul-Oct;728(1-2):23-34. Epub 2011 May 12. PMID: 21605699; PMCID: PMC4028085. DOI: 10.1016/j.mrrev.2011.05.002

[89] Helmut, Popper. (2020). Primary tumor and metastasis-sectioning the different steps of the metastatic cascade.. Translational lung cancer research, 2020 Oct;9(5):2277-2300. PMID: 33209649. PMCID: PMC7653118. DOI: 10.21037/tlcr-20-175

[90] Elisa, C., Woodhouse., Rodrigo, F., Chuaqui., Lance, A., Liotta. (1997). General mechanisms of metastasis. Cancer, An International Interdisciplinary Journal of The American Cancer Society. Volume 80, issue S8. P1529- 1537. https://doi.org/10.1002/(SICI)1097-0142(19971015)80:8+<1529::AID-CNCR2>3.0.CO;2-F

[91] Fares J, Fares MY, Khachfe HH, Salhab HA, Fares Y (2020). Molecular principles of metastasis: a hallmark of cancer revisited. Signal Transduct Target Ther. 2020 Mar 12;5(1):28. PMID: 32296047. PMCID: PMC7067809. DOI: 10.1038/s41392-020-0134-x

[92] Nikolaos, Sfakianakis., Mark, A., J., Chaplain. (2020). Mathematical Modelling of Cancer Invasion: A Review. https://hdl.handle.net/10023/25867

[93] Tatiana, S., Gerashchenko., Nikita, M., Novikov., Nikita, M., Novikov., N., V., Krakhmal., Sofia, Y., Zolotaryova., M.V., Zavyalova., M.V., Zavyalova., Nadezhda, V, Cherdyntseva., Nadezhda, V, Cherdyntseva., Evgeny, V., Denisov., Evgeny, V., Denisov.,

Bibliography

Vladimir, M., Perelmuter. (2019). Markers of Cancer Cell Invasion: Are They Good Enough?. Journal of Clinical Medicine, July 2019. 8(8):1092. DOI:10.3390/jcm8081092

[94] Jianzhou, Zheng., Liuyang, He., Lei, Xia., Yong, Wang. (2017). Research advances in molecular mechanisms of the invasion and metastasis of circulating tumor cells. doi: 10.3760/CMA.J.ISSN.1673-4386.2017.06.010

[95] Abdullah, Norul Hidayah (2014). Mathematical model for cancer cell invasion of tissue. Masters thesis, Universiti Teknologi Malaysia, Faculty of Science.

[96] Krakhmal NV, Zavyalova MV, Denisov EV, Vtorushin SV, Perelmuter VM (2015). Cancer Invasion: Patterns and Mechanisms. Acta Naturae. 2015 Apr-Jun;7(2):17-28. PMID: 26085941; PMCID: PMC4463409

[97] Verbitsky,, Igor, E.. (2023). Data from Mitosis-Mediated Intravasation in a Tissue-Engineered Tumor–Microvessel Platform. https://doi.org/10.1158/0008-5472.c.6509396.v1

[98] Wong, Andrew D.; Searson, Peter C. (2017). Data from Mitosis-Mediated Intravasation in a Tissue-Engineered Tumor–Microvessel Platform. American Association for Cancer Research. Collection. https://doi.org/10.1158/0008-5472.c.6509396.v1

[99] Mayo LN, Kutys ML (2022). Conversation before crossing: dissecting metastatic tumor-vascular interactions in microphysiological systems. Am J Physiol Cell Physiol. 2022 Nov 1;323(5):C1333-C1344. Epub 2022 Sep 19. PMID: 36121131; PMCID: PMC9602802. DOI: 10.1152/ajpcell.00173.2022

[100] Adrien, Vu, Van., William, B., English. (2023). Data from ErbB3-Dependent Motility and Intravasation in Breast Cancer Metastasis. doi: 10.1158/0008-5472.c.6495188

[101] Magdalena K. Sznurkowska, Nicola Aceto (2022). The gate to metastasis: key players in cancer cell intravasation. The FEBS Journal. Volume 289, Issue 15. P4336- 4354. https://doi.org/10.1111/febs.16046

[102] Kurma K, Alix-Panabières C (2023). Mechanobiology and survival strategies of circulating tumor cells: a process towards the invasive and metastatic phenotype. Front Cell Dev Biol. 2023 May 5;11:1188499. PMID: 37215087; PMCID: PMC10196185. DOI : 10.3389/fcell.2023.1188499

[103] Chen Qian; Asurayya Worrede-Mahdi; Fei Shen; Anthony DiNatale; Ramanpreet Kaur; Qiang Zhang; Massimo Cristofanilli; Olimpia Meucci; Alessandro Fatatis (2023). Data from Impeding Circulating Tumor Cell Reseeding Decelerates Metastatic Progression and Potentiates Chemotherapy. DOI: 10.1158/1541-7786.c.6540148

[104] Diane S. Kang, Aidan Moriarty, Jeong Min-Oh & Hydari Masuma Begum (2023). Biophysical Properties and Isolation of Circulating Tumor Cells. In book: Engineering and Physical Approaches to Cancer (pp.255-283). DOI:10.1007/978-3-031-22802-5_9

[105] Gaetan, Aime, Noubissi, Nzeteu., Claudia, Geismann., Alexander, Arlt., Frederik, J.H., Hoogwater., Maarten, W., Nijkamp., N., Helge, Meyer., Maximilian, Bockhorn. (2022). Role of Epithelial-to-Mesenchymal Transition for the Generation of Circulating Tumors Cells and Cancer Cell Dissemination. Cancers, DOI : 10.3390/cancers14225483

[106] Tao J, Zhu L, Yakoub M, Reißfelder C, Loges S, Schölch S (2022). Cell-Cell Interactions Drive Metastasis of Circulating Tumor Microemboli. Cancer Res. 2022 Aug 3;82(15):2661-2671. PMID: 35856896. DOI: 10.1158/0008-5472.CAN-22-0906

[107] Zavyalova MV, Denisov EV, Tashireva LA, Savelieva OE, Kaigorodova EV, Krakhmal NV, Perelmuter VM (2019). Intravasation as a Key Step in Cancer Metastasis. Biochemistry (Mosc). 2019 Jul;84(7):762-772. PMID: 31509727. DOI: 10.1134/S0006297919070071

[108] Giusti I, Poppa G, Di Fazio G, D'Ascenzo S, Dolo V(2023). Metastatic Dissemination: Role of Tumor-Derived Extracellular Vesicles and Their Use as Clinical Biomarkers. Int J Mol Sci. 2023 May 31;24(11):9590. PMID: 37298540; PMCID: PMC10253525. DOI: 10.3390/ijms24119590

[109] Cheng X, Cheng K. Visualizing cancer extravasation (2020): from mechanistic studies to drug development. Cancer Metastasis Rev. 2021 Mar;40(1):71-88. Epub 2020 Nov 6. PMID: 33156478; PMCID: PMC7897269. DOI: 10.1007/s10555-020-09942-2

[110] Kim S, Wan Z, Jeon JS, Kamm RD (2022). Microfluidic vascular models of tumor cell extravasation. Front Oncol. 2022 Nov 11;12:1052192. PMID: 36439519; PMCID: PMC9698448. DOI: 10.3389/fonc.2022.1052192

[111] Wang Z, Wu X, Chen HN, Wang K. Amino acid metabolic reprogramming in tumor metastatic colonization (2023). Front Oncol. 2023 Mar 14;13:1123192. PMID: 36998464; PMCID: PMC10043324. DOI: 10.3389/fonc.2023.1123192

[112] Jin B, Zhang YY, Pan JX (2023). [The Role and Significance of Hepatic Environmental Cells in Tumor Metastatic Colonization to Liver]. Sichuan Da Xue Xue Bao Yi Xue Ban. 2023 May;54(3):469-474. Chinese. PMID: 37248570; PMCID: PMC10475444. DOI: 10.12182/20230560301

[113] San Juan BP, Garcia-Leon MJ, Rangel L, Goetz JG, Chaffer CL (2019). The Complexities of Metastasis. Cancers (Basel). 2019 Oct 16;11(10):1575. PMID: 31623163; PMCID: PMC6826702. DOI: 10.3390/cancers11101575

[114] Liu J, Lian J, Chen Y, Zhao X, Du C, Xu Y, Hu H, Rao H, Hong X (2021). Circulating Tumor Cells (CTCs): A Unique Model of Cancer Metastases and Non-invasive Biomarkers of Therapeutic Response. Front Genet. 2021 Aug 25;12:734595. PMID: 34512735; PMCID: PMC8424190. DOI : 10.3389/fgene.2021.734595

[115] Castro-Giner F, Aceto N (2020). Tracking cancer progression: from circulating tumor cells to metastasis. Genome Med. 2020 Mar 19;12(1):31. PMID: 32192534; PMCID: PMC7082968. DOI: 10.1186/s13073-020-00728-3

[116] Tao J, Zhu L, Yakoub M, Reißfelder C, Loges S, Schölch S (2022). Cell-Cell Interactions Drive Metastasis of Circulating Tumor Microemboli. Cancer Res. 2022 Aug 3;82(15):2661-2671. PMID: 35856896. DOI : 10.1158/0008-5472.CAN-22-0906

[117] Guan Y, Xu F, Tian J, Chen H, Yang C, Huang S, Gao K, Wan Z, Li M, He M, Chong T (2020). Pathology of circulating tumor cells and the available capture tools (Review). Oncol Rep. 2020 May;43(5):1355-1364. Epub 2020 Mar 6. PMID: 32323847. DOI : 10.3892/ou.2020.7533

[118] Karabacak, N. M., Spuhler, P. S., Fachin, F., Lim, E. J., Pai, V., Ozkumur, E., Martel, J. M., Kojic, N., Smith, K., Chen, P.-. i ., Yang, J., Hwang, H., Morgan, B., Trautwein, J., Barber, T. A., Stott, S. L., Maheswaran, S., Kapur, R., Haber, D. A., & Toner, M.. (2014). Microfluidic, marker-free isolation of circulating tumor cells from blood samples. 9(3), 694–710. https://doi.org/10.1038/nprot.2014.044

[119] Abdulla, A., Zhang, T., Li, S., Guo, W., Warden, A. R., Xin, Y., Maboyi, N., Lou, J., Xie, H., & Ding, X.. (2021). Integrated microfluidic single-cell immunoblotting chip enables high-throughput isolation, enrichment and direct protein analysis of circulating tumor cells. 8(1), 1–12. https://doi.org/10.1038/s41378-021-00342-2

[120] Fachin, F., Spuhler, P., Martel-Foley, J. M., Edd, J. F., Barber, T. A., Walsh, J., Karabacak, M., Pai, V., Yu, M., Smith, K., Hwang, H., Yang, J., Shah, S., Yarmush, R., Sequist, L. V., Stott, S. L., Maheswaran, S., Haber, D. A., Kapur, R., & Toner, M.. (2017). Monolithic Chip for High-throughput Blood Cell Depletion to Sort Rare Circulating Tumor Cells. 7(1), 10936–10936. https://doi.org/10.1038/s41598-017-11119-x

[121] Errico, A.. (2014). Breast cancer: CTCs - a predictive approach for targeted cancer therapy. 11(9), 501–501. https://doi.org/10.1038/nrclinonc.2014.129

[122] Ting, D. T., Wittner, B. S., Ligorio, M., Vincent Jordan, N., Shah, A. M., Miyamoto, D. T., Aceto, N., Bersani, F., Brannigan, B. W., Xega, K., Ciciliano, J. C., Zhu, H., MacKenzie, O. C., Trautwein, J., Arora, K. S., Shahid, M., Ellis, H. L., Qu, N., Bardeesy, N., Rivera, M. N., … Haber, D. A. (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. Cell reports, 8(6), 1905–1918. https://doi.org/10.1016/j.celrep.2014.08.029

[123] Amir, Seyfoori., S.A., Seyyed, Ebrahimi., Mohamadmahdi, Samandari., Ehsan, Samiei., E., Štefanek., Cathie, Garnis., Mohsen, Akbari. (2023). Microfluidic-Assisted CTC Isolation and In Situ Monitoring Using Smart Magnetic Microgels..Small, doi: 10.1002/smll.202205320

[124] Md., Sadiqul, Islam., Xiaolin, Chen. (2023). Continuous CTC separation through a DEP-based contraction-expansion inertial microfluidic channel..Biotechnology Progress, doi: 10.1002/btpr.3341

[125] Lucie, Descamps., D., J., Le, Roy., Anne-Laure, Deman. (2022). Microfluidic-Based Technologies for CTC Isolation: A Review of 10 Years of Intense Efforts towards Liquid Biopsy. International Journal of Molecular Sciences, doi: 10.3390/ijms23041981

[126] Ajanth, P., Sudeepthi, A. & Sen, A.K (2020). Microfluidics Technology for Label-Free Isolation of Circulating Tumor Cells. J. Inst. Eng. India Ser. C 101, 1051–1071 (2020). https://doi.org/10.1007/s40032-020-00617-z

[127] Jiang, X., Wong, K. H. K., Khankhel, A. H., Zeinali, M., Reategui, E., Phillips, M. J., Luo, X., Aceto, N., Fachin, F., Hoang, A. N., Kim, W., Jensen, A. E., Sequist, L. V., Maheswaran, S., Haber, D. A., Stott, S. L., & Toner, M.. (2017). Microfluidic isolation of platelet-covered circulating tumor cells. 17(20), 3498–3503. https://doi.org/10.1039/C7LC00654C

[128] Adams, D. L., Zhu, P., Makarova, O. V., Martin, S. S., Charpentier, M., Chumsri, S., Li, S., Amstutzd, P., & Tang, C.-M.. (2014). The systematic study of circulating tumor cell isolation using lithographic microfilters. 4(9), 4334–4342. https://doi.org/10.1039/C3RA46839A

[129] Balasubramanian, S., Kagan, D., Hu, C.-M. J., Campuzano, S., Lobo-Castañon, M. J., Lim, N., Kang, D. Y., Zimmerman, M., Zhang, L., & Wang, J.. (2011). Micromachine-enabled capture and isolation of cancer cells in complex media.. 50(18), 4161–4164. https://doi.org/10.1002/anie.201100115

[130] Marilena, L., Currey., Viswajit, Kandula., Ronald, Biggs., John, F., Marko., Andrew, D., Stephens. (2022). A Versatile Micromanipulation Apparatus for Biophysical Assays of the Cell Nucleus. Cellular and Molecular Bioengineering, doi: 10.1007/s12195-022-00734-y

[131] Khoo, B.L., Chaudhuri, P.K., Lim, C.T., Warkiani, M.E. (2017). Advancing Techniques and Insights in Circulating Tumor Cell (CTC) Research. In: Aref, A., Barbie, D. (eds) Ex Vivo Engineering of the Tumor Microenvironment. Cancer Drug Discovery and Development. Humana Press, Cham. https://doi.org/10.1007/978-3-319-45397-2_5

[132] Magbanua, M.J., & Park, J. (2013). Isolation of circulating tumor cells by immunomagnetic enrichment and fluorescence-activated cell sorting (IE/FACS) for molecular profiling. Methods, 64 2, 114-8 . https://doi.org/10.1016/j.ymeth.2013.07.029

[133] Chen, H., Li, Y., Zhang, Z., & Wang, S. (2020). Immunomagnetic separation of circulating tumor cells with microfluidic chips and their clinical applications. Biomicrofluidics, 14(4), 041502. https://doi.org/10.1063/5.0005373

[134] Che J., Yu V., Dhar M., Renier C., Matsumoto M., Heirich K., Garon E. B., Goldman J., Rao J., Sledge G. W., Pegram M. D., Sheth S., Jeffrey S. S., et al (2016). Classification of large circulating tumor cells isolated with ultra-high throughput

microfluidic Vortex technology. Oncotarget. 2016; 7: 12748-12760. Retrieved from https://www.oncotarget.com/article/7220/text/

[135] Kuske, A., Gorges, T. M., Tennstedt, P., Tiebel, A. K., Pompe, R., Preißer, F., Prues, S., Mazel, M., Markou, A., Lianidou, E., Peine, S., Alix-Panabières, C., Riethdorf, S., Beyer, B., Schlomm, T., & Pantel, K. (2016). Improved detection of circulating tumor cells in non-metastatic high-risk prostate cancer patients. Scientific reports, 6, 39736. https://doi.org/10.1038/srep39736

[136] Man, Y., Wang, Q., & Kemmner, W. (2011). Currently Used Markers for CTC Isolation - Advantages, Limitations and Impact on Cancer Prognosis. Experimental pathology, 2011. https://doi.org/10.4172/2161-0681.1000102

[137] van der Toom, E. E., Verdone, J. E., Gorin, M. A., & Pienta, K. J. (2016). Technical challenges in the isolation and analysis of circulating tumor cells. Oncotarget, 7(38), 62754–62766. https://doi.org/10.18632/oncotarget.11191

[138] Descamps, L., Le Roy, D., & Deman, A. L. (2022). Microfluidic-Based Technologies for CTC Isolation: A Review of 10 Years of Intense Efforts towards Liquid Biopsy. International journal of molecular sciences, 23(4), 1981. https://doi.org/10.3390/ijms23041981

[139] Macaraniag, C., Luan, Q., Zhou, J., & Papautsky, I. (2022). Microfluidic techniques for isolation, formation, and characterization of circulating tumor cells and clusters. APL bioengineering, 6(3), 031501. https://doi.org/10.1063/5.0093806

[140] Chen, J., Liu, C. Y., Wang, X., Sweet, E., Liu, N., Gong, X., & Lin, L. (2020). 3D printed microfluidic devices for circulating tumor cells (CTCs) isolation. Biosensors & bioelectronics, 150, 111900. https://doi.org/10.1016/j.bios.2019.111900

[141] Warkiani, M. E., Khoo, B. L., Wu, L., Tay, A. K., Bhagat, A. A., Han, J., & Lim, C. T. (2016). Ultra-fast, label-free isolation of circulating tumor cells from blood using spiral microfluidics. Nature protocols, 11(1), 134–148. https://doi.org/10.1038/nprot.2016.003

[142] Du, J., Liu, X., & Xu, X. (2014). Advances in isolation and enrichment of circulating tumor cells in microfluidic chips. Chinese journal of chromatography, 32(1), 7–12. https://doi.org/10.3724/sp.j.1123.2013.08009

[143] Myung, J. H., & Hong, S. (2015). Microfluidic devices to enrich and isolate circulating tumor cells. Lab on a chip, 15(24), 4500–4511. https://doi.org/10.1039/c5lc00947b

[144] Sun, C., Hsieh, YP., Ma, S. et al (2017). Immunomagnetic separation of tumor initiating cells by screening two surface markers. Sci Rep **7**, 40632 (2017). https://doi.org/10.1038/srep40632

[145] Tang, M., Wen, C. Y., Wu, L. L., Hong, S. L., Hu, J., Xu, C. M., Pang, D. W., & Zhang, Z. L. (2016). A chip assisted immunomagnetic separation system for the efficient capture and in situ identification of circulating tumor cells. Lab on a chip, 16(7), 1214–1223. https://doi.org/10.1039/c5lc01555c

[146] Wang, Z., Wu, W., Wang, Z., Tang, Y., Deng, Y., Xu, L., Tian, J., & Shi, Q. (2016). Ex vivo expansion of circulating lung tumor cells based on one-step microfluidics-based immunomagnetic isolation. The Analyst, 141(12), 3621–3625. https://doi.org/10.1039/c5an02554k

[147] Earhart, C. M., Hughes, C. E., Gaster, R. S., Ooi, C. C., Wilson, R. J., Zhou, L. Y., Humke, E. W., Xu, L., Wong, D. J., Willingham, S. B., Schwartz, E. J., Weissman, I. L., Jeffrey, S. S., Neal, J. W., Rohatgi, R., Wakelee, H. A., & Wang, S. X. (2014). Isolation and mutational analysis of circulating tumor cells from lung cancer patients with magnetic sifters and biochips. Lab on a chip, 14(1), 78–88. https://doi.org/10.1039/c3lc50580d

[148] Meng, Q. F., Cheng, Y. X., Huang, Q., Zan, M., Xie, W., Sun, Y., Li, R., Wei, X., Guo, S. S., Zhao, X. Z., Rao, L., & Liu, W. (2019). Biomimetic Immunomagnetic Nanoparticles with Minimal Nonspecific Biomolecule Adsorption for Enhanced Isolation of Circulating Tumor Cells. ACS applied materials & interfaces, 11(32), 28732–28739. https://doi.org/10.1021/acsami.9b10318

[149] Horgan, K., Shaw, S., & Boirivant, M. (2009). Immunomagnetic purification of T cell subpopulations. Current protocols in immunology, Chapter 7, 7.4.1–7.4.9. https://doi.org/10.1002/0471142735.im0704s85

[150] Kim, S., Han, S. I., Park, M. J., Jeon, C. W., Joo, Y. D., Choi, I. H., & Han, K. H. (2013). Circulating tumor cell microseparator based on lateral magnetophoresis and immunomagnetic nanobeads. Analytical chemistry, 85(5), 2779–2786. https://doi.org/10.1021/ac303284u

[151] Bhat, M. P., Thendral, V., Uthappa, U. T., Lee, K. H., Kigga, M., Altalhi, T., Kurkuri, M. D., & Kant, K. (2022). Recent Advances in Microfluidic Platform for Physical and Immunological Detection and Capture of Circulating Tumor Cells. Biosensors, 12(4), 220. https://doi.org/10.3390/bios12040220

[152] Naeem, A., James, N., Tanvir, M., Marriam, M., & Nathaniel, S. (2017). Fluorescence Activated Cell Sorting (FACS): An Advanced Cell Sorting Technique.

[153] Basu, S., Campbell, H. M., Dittel, B. N., & Ray, A. (2010). Purification of specific cell population by fluorescence activated cell sorting (FACS). Journal of visualized experiments : JoVE, (41), 1546. https://doi.org/10.3791/1546

[154] Sino Biological (2024). Fluorescence-activated Cell Sorting (FACS). Retrieved April 30, 2024, from https://www.sinobiological.com/category/fcm-facs-facs

[155] Xu, M., Liu, W., Zou, K., Wei, S., Zhang, X., Li, E., & Wang, Q. (2021). Design and Clinical Application of an Integrated Microfluidic Device for Circulating Tumor Cells Isolation and Single-Cell Analysis. Micromachines, 12(1), 49. https://doi.org/10.3390/mi12010049

[156] LINH, Nguyen Hue & al (2022). RNA Drugs. VNU Journal of Science: Medical and Pharmaceutical Sciences**,** [S.l.], v. 38, n. 1, mar. 2022. ISSN 2588-1132. Available at: <https://js.vnu.edu.vn/MPS/article/view/4388>. Date accessed: 27 june 2024. doi: https://doi.org/10.25073/2588-1132/vnumps.4388.

[157] Schuster P, Stadler PF, Renner A (1997). RNA structures and folding: from conventional to new issues in structure predictions. Curr Opin Struct Biol. 1997 Apr;7(2):229-35. doi: 10.1016/s0959-440x(97)80030-9. PMID: 9094330.

[158] Frédérique Théry Théry (2013). L'importance biologique des ARN non codants : perspectives historique et philosophique.. Université Panthéon-Sorbonne - Paris I, 2013. Français. ffNNT : 2013PA010620ff. fftel-01080000f

[159] Anant, Manish, Singh., Wasif, Bilal, Haju. (2022). Artificial IntelligenceArtificial Intelligence. International Journal For Science Technology And Engineering, doi: 10.22214/ijraset.2022.44306

[160] MATTIA, G. M.. (2022). Artificial Intelligence for Neurological Diseases: aiding Diagnose and Improve Understanding of Convolutive Neural Networks Behavior.

[161] Panesar, A. (2019). Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes. Germany: Apress.

[162] Pierre, Marquis., Odile, Papini., Henri, Prade. (2020). Elements for a History of Artificial Intelligence. doi: 10.1007/978-3-030-06164-7_1

[163] Klaus, Henning. (2021). How Did Artificial Intelligence Come into Being and Where Do We Stand Today.doi: 10.1007/978-3-030-52897-3_3

[164] Fatima, Hameed, Khan., Muhammad, Adeel, Pasha., Shahid, Masud. (2021). Advancements in Microprocessor Architecture for Ubiquitous AI-An Overview on History, Evolution, and Upcoming Challenges in AI Implementation..Micromachines, doi: 10.3390/MI12060665

[165] Ravit, Kumar. (2022). Machine Learning. International Journal For Science Technology And Engineering, doi: 10.22214/ijraset.2022.44376

[166] Panesar, A. (2019). Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes. Germany: Apress.

[167] Christophe Genevey-Metat (2023). Apprentissage automatique pour les attaques par canaux auxiliaires. Machine Learning [cs.LG]. Université de Rennes, 2023. English. ffNNT : 2023URENS030ff. fftel-04241537f

[168] Matt Crabtree (2023). What is Machine Learning? Definition, Types, Tools & More. Datacamp (2023, July 22).. https://www.datacamp.com/blog/what-is-machine-learning

[169] Zulkarnain, S., Mehjabin, M., Sultana, R., Hasan, M. A., Arefin, S., &Farid, D. M.. (2023). A New Method for Learning Decision Tree Classifier. doi: 10.1109/ecce57851.2023.10101557

[170] Dewan, Md., Farid. (2023). A New Method for Learning Decision Tree Classifier. doi: 10.1109/ECCE57851.2023.10101557

[171] Kenneth Dwyer Robert Holte (2007). Decision Tree Instability and Active Learning. Conference: Machine Learning: ECML 2007, 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007, Proceedings. 4701:128-139 DOI:10.1007/978-3-540-74958-5_15

[172] Leqi, Tian., Wenbin, Wu., Tianwei, Yu. (2023). Graph Random Forest: A Graph Embedded Algorithm for Identifying Highly Connected Important Features. Biomolecules, doi: 10.3390/biom13071153

[173] Grzegorz, Dudek. (2022). A Comprehensive Study of Random Forest for Short-Term Load Forecasting. Energies, doi: 10.3390/en15207547

[174] Ion-Margineanu, A. (2017). Machine learning for classifying abnormal brain tissue progression based on multi-parametric Magnetic Resonance data. Bioengineering. Université de Lyon; KU Leuven (1970-…). (ffNNT: 2017LYSE1224ff). Retrieved from fftel-01769443.

[175] Haihao, Lu., Rahul, Mazumder. (2020). Randomized Gradient Boosting Machine. Siam Journal on Optimization, doi: 10.1137/18M1223277

[176] David, N., Bresch., Samuel, Lüthi., Rajendra, Ghadwal. (2023). Gradient boosting for socio-economic wildfire risk assessment. doi: 10.5194/egusphere-egu23-11387

[177] Dominic, Lagrois., Tyler, R., Bonnell., Ankita, Shukla., Clément, Chion. (2022). The Gradient-Boosting Method for Tackling High Computing Demand in Underwater Acoustic Propagation Modeling. Journal of Marine Science and Engineering,doi: 10.3390/jmse10070899

[178] student, H., & Raju, M. M.. (2022). A Study on Deep Learning. 10(XI). https://doi.org/10.22214/ijraset.2022.47486

[179] Interaction Design Foundation - IxDF. (2023, October 26). *What are Neural Networks (NN)?*. Interaction Design Foundation - IxDF. https://www.interaction-design.org/literature/topics/neural-networks

[180] Pajankar, A., Joshi, A. (2022). Feedforward Neural Networks. In: Hands-on Machine Learning with Python. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-7921-2_13

[181] Fernando, P., Guiomar. (2023). The Concept of Forward-Forward Learning Applied to a Multi Output Perceptron. doi: 10.48550/arxiv.2304.03189

[182] Ahmet, Cevahir, Cinar. (2020). Training Feed-Forward Multi-Layer Perceptron Artificial Neural Networks with a Tree-Seed Algorithm. Arabian Journal for Science and Engineering, doi: 10.1007/S13369-020-04872-1

[183] NASSER, Y.. (2023). ENHANCED DEEP NEURAL NETWORKS FOR EARLY DIAGNOSIS OF KNEE OSTEOARTHRITIS.

[184] Hugo, Siqueira., Ivette, Luna. (2019). Performance comparison of feedforward neural networks applied to stream flow series forecasting. Journal | MESA,

[185] Tugba, Ozdemir. (2022). Comparison of Feedforward Perceptron Network with LSTM for Solar Cell Radiation Prediction. Applied Sciences, doi: 10.3390/app12094463

[186] Georgios, N., Kouziokas. (2019). Unemployment Prediction in UK by Using a Feedforward Multilayer Perceptron. doi: 10.1007/978-3-319-95666-4_5

[187] Soulaimane Guedria (2020). A scalable and component-based deep learning parallelism platform : an application to convolutional neural networks for medical imaging segmentation.. Logic in Computer Science [cs.LO]. Université Grenoble Alpes [2020-..], 2020. English. ⟨NNT : 2020GRALM023⟩.

[188] Dr., Deepa, A. (2023). Back Propagation. International Journal For Science Technology And Engineering, doi: 10.22214/ijraset.2023.50077

[189] András, Attila, Csontos. (2023). Backpropagation and F-adjoint. doi: 10.36227/techrxiv.22650193

[**190] LAURA, BEATRIZ, DE, FREITAS, NOVAIS. (2023).** Backpropagation and F-adjoint. doi: 10.36227/techrxiv.22650193.v1

[191] Dubravka, Bozic., Biserka, Runje., Dragutin, Lisjak., Davor, Kolar. (2023). Metrics Related to Confusion Matrix as Tools for Conformity Assessment Decisions. Applied Sciences, doi: 10.3390/app13148187

[192] Görtler, J., Hohman, F., Moritz, D., Wongsuphasawat, K., Ren, D., Nair, R., Kirchner, M., & Patel, K. (2022). Neo: Generalizing Confusion Matrix Visualization to Hierarchical and Multi-Output Labels. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI).doi: 10.1145/3491102.3501823

[193] Kevin Riehl, Michael Neunteufel and Martin Hemberg (2023). Hierarchical confusion matrix for classification performance evaluation. Journal of the Royal Statistical

Society Series C: Applied Statistics, 2023, 72, 1394–1412**.** https://doi.org/10.1093/jrsssc/qlad057

[194] Jacob Murel Ph.D., Eda Kavlakoglu (2024). What is a confusion matrix?. (2024, June 19). https://www.ibm.com/topics/confusion-matrix

[195] Rajvir, Kaur., Jeewani, Anupama, Ginige. (2018). Comparative Evaluation of Accuracy of Selected Machine Learning Classification Techniques for Diagnosis of Cancer: A Data Mining Approach. World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering,

[196] Slamet, Riyanto., Imas, Sukaesih, Sitanggang., Taufik, Djatna., Tika, Dewi, Atikah. (2023). Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification. International Journal of Advanced Computer Science and Applications, doi: 10.14569/ijacsa.2023.01406116

[197] Kuldeep Singh Kaswan, Jagjit Singh Dhatterwal, B Balamurugan (2022). Python for Beginners. Chapman and Hall/CRC. New York. DOIhttps://doi.org/10.1201/9781003202035

[198] Welcome to Python.org. (2024, June 1). https://www.python.org/

[**199] ARTHUR, SANTOS, DA, SILVA. (2023).** Analysis of Python Libraries for Artificial Intelligence. doi: 10.1007/978-981-99-0071-8_13

[200] Mingu, Kang., Suntae, Kim., Duksan, Ryu., JaeHyuk, Cho. (2022). Which Exceptions Do We Have to Catch in the Python Code for AI Projects?. International Journal of Software Engineering and Knowledge Engineering, doi: 10.1142/s0218194022500814

[201] Kanda, Rongsawad. (2023). Python for Deep Learning: A General Introduction. Synthesis lectures on engineering, science, and technology, doi: 10.1007/978-3-031-35737-4_6

[202] byteXD (2022, October 1). What is Google Colab: A Beginner's Guide. https://bytexd.com/what-is-google-colab-a-beginner-guide/

[203] Zhao, L., Wu, X., Li, T., Luo, J., & Dong, D. (2020). ctcRbase: The gene expression database of circulating tumor cells and microemboli. Database, 2020, baaa020. DOI: 10.1093/database/baaa020