HELLENIC REPUBLIC
UNIVERSITY OF IOANNINA
SCHOOL OF HEALTH SCIENCES
DEPARTMENT OF BIOLOGICAL APPLICATIONS AND TECHNOLOGY

Multidimensional Computational Methods for Modeling Cancer Diagnosis, Prognosis and Treatment

Konstantina D. Kourou

PhD Thesis

IOANNINA 2020

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΩΝ

Πολυδιάστατες υπολογιστικές μέθοδοι για την μοντελοποίηση της διάγνωσης, πρόγνωσης και θεραπείας του καρκίνου

Κωνσταντίνα Δ. Κούρου

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΙΩΑΝΝΙΝΑ 2020

«Η έγκριση της διδακτορικής διατριβής από το Τμήμα Βιολογικών Εφαρμογών και Τεχνολογιών της Σχολής Επιστημών Υγείας του Πανεπιστημίου Ιωαννίνων δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα Ν. 5343/32, άρθρο 202, παράγραφος 2».

Date of Application of Ms. Konstantina Kourou: 14 November 2014
Date of Appointment of PhD Advisory Committee: 16 January 2015

Members of the 3 Member Advisory Committee:

Thesis Advisor

Costas Papaloukas, Associate Professor, Department of Biological Applications and Technology, School of Health Sciences, University of Ioannina

Members

Date of Thesis Subject Definition: 16 January 2015

PhD Thesis Title: Multidimensional Computational Methods for Modeling Cancer Diagnosis, Prognosis and Treatment

Date of Appointment of the 7-member Examination Committee: 11-05-2020

Costas Papaloukas	Associate Professor, Department of Biological Applications and Technology, School of Health Sciences, University of Ioannina
Dimitrios I. Fotiadis	Professor, Department of Materials Science and Engineering, School of Engineering, University of Ioannina
Anastasios Troganis	Professor, Department of Biological Applications and Technology, School of Health Sciences, University of Ioannina
Michalis Mitsis	Professor, Faculty of Medicine, School of Health Sciences, University of Ioannina
Manolis Tsiknakis	Professor, Department of Electric and Computer Engineering, School of Engineering, Hellenic Mediterranean University
Konstantinos Marias	Associate Professor, Department of Electric and Computer Engineering, School of Engineering, Hellenic Mediterranean University
Themis Exarchos	Assistant Professor, Department of Informatics, School of Information Science and Informatics, Ionian University

The PhD thesis is **approved**, with «Excellent» on **06-06-2020**

Ημερομηνία Αίτησης της κας Κωνσταντίνας Κούρου: 14 Νοεμβρίου 2014

Ημερομηνία Ορισμού Τριμελούς Συμβουλευτικής Επιτροπής: 16 Ιανουαρίου 2015

Μέλη Τριμελούς Συμβουλευτικής Επιτροπής:

Επιβλέπων

Κωνσταντίνος Παπαλουκάς, Αναπληρωτής Καθηγητής, Τμήμα Βιολογικών Εφαρμογών και Τεχνολογιών, Σχολή Επιστημών Υγείας, Πανεπιστήμιο Ιωαννίνων

Μέλη

Ημερομηνία Ορισμού Θέματος: 16 Ιανουαρίου 2015

Θέμα Διατριβής: Πολυδιάστατες υπολογιστικές μέθοδοι για την μοντελοποίηση της διάγνωσης, πρόγνωσης και θεραπείας του καρκίνου

Διορισμός Επταμελούς Εξεταστικής Επιτροπής: 11-05-2020

Κωνσταντίνος Παπαλουκάς	Αναπληρωτής Καθηγητής, Τμήμα Βιολογικών Εφαρμογών και Τεχνολογιών, Σχολή Επιστημών Υγείας, Πανεπιστήμιο Ιωαννίνων
Δημήτριος Ι. Φωτιάδης	Καθηγητής, Τμήμα Μηχανικών Επιστήμης Υλικών, Πολυτεχνική Σχολή, Πανεπιστήμιο Ιωαννίνων
Αναστάσιος Τρογκάνης	Καθηγητής, Τμήμα Βιολογικών Εφαρμογών και Τεχνολογιών, Σχολή Επιστημών Υγείας, Πανεπιστήμιο Ιωαννίνων
Μιχαήλ Μήτσης	Καθηγητής, Τμήμα Ιατρικής, Σχολή Επιστημών Υγείας, Πανεπιστήμιο Ιωαννίνων
Μανώλης Τσικνάκης	Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Σχολή Μηχανικών, Ελληνικό Μεσογειακό Πανεπιστήμιο
Κωνσταντίνος Μαριάς	Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Σχολή Μηχανικών, Ελληνικό Μεσογειακό Πανεπιστήμιο
Θεμιστοκλής Έξαρχος	Επίκουρος Καθηγητής, Τμήμα Πληροφορικής, Σχολή Επιστήμης της Πληροφορίας & Πληροφορικής, Ιόνιο Πανεπιστήμιο

Έγκριση Διδακτορικής Διατριβής με βαθμό «ΑΡΙΣΤΑ» στις **06-06-2020**

DEDICATION

To my parents

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to my supervisor Associate Professor Costas Papaloukas, Department of Biological Applications and Technology, University of Ioannina for his support and great assistance in order to complete this thesis. He guided and encouraged me to be professional and do the right thing even when the road got tough. I would also like to emphasize my sincere appreciation to Professor Dimitrios I. Fotiadis, the second member of my advisory committee, for his inspiration, guidance and continuous support through the completion of this thesis. I would also like to sincerely thank Professor Fotiadis for the financial support I had all these years through the Unit of Medical Technology and Intelligent Information Systems research programs. I would also like to thank the third member of my advisory committee, Professor Anastasios Troganis, for his guidance and continuous support during my Ph.D. studies.

I would also like to express my sincere appreciation to Professor Michalis Mitsis, Professor Manolis Tsiknakis, Associate Professor Konstantinos Marias and Assistant Professor Themis Exarchos for their collaboration and support all these years and their participation in my examination committee.

I would like to warmly thank all my colleagues and friends from the Unit of Medical Technology and Intelligent Information Systems for the excellent cooperation, friendship and support all these years.

In addition, I would like to refer to the support of The Hellenic Foundation for Research and Innovation (H.R.F.I) for the scholarship I received under the program of Research and Innovation and the opportunity to complete my thesis through their financial support. Through my participation in this program, I broadened the perspectives and knowledge on my thesis, and I had also the chance to participate in conferences and thereby present my research work.

I would like to thank my family, who all these years support my dreams and my efforts. My parents deserve special thanks for their unlimited support and encouraging. They kept me going on and this work would not have been possible without their input. Without their persistent help, the goal of this thesis would not have been realized.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

X	a set of random variables
B	a Bayesian network
G	an annotated directed acyclic graph
Θ	the parameters that quantify the Bayesian network
$P_B(x_1,..x_n)$	a joint probability distribution over X in B
$pa(x_i)$	the parents of $x_i$in G
$P(X[t+1]\|X[t])$	transition probability
t	time
B_0	prior network
$X[0]$	initial state
$B_{\rightarrow}$	transition network
$P_B(x[0],..,x[T])$	a joint probability distribution over $X[0],...,X[T]$ in a DBN
$V_{i,j}$	the i_{th} variable in the j_{th} time-slice
$E[y_g]$	a linear model
g	a gene
y_g	gene expression values
β_g	coefficients
$S(k)$	the score value
k	the radius of pathway steps
M_k	the number of input transcription factors
N_k	total number of all steps
$M_{max,k}$	the highest values among all master regulator nodes
L	a learning problem
F	a non-linear parameterized function

x	a set of features
Z	a set of training samples
x	the random “input” or “explanatory” features
y	the “output” or the “response” variable
x_i	input vector
P_i	a patient
$L(y, F(x))$	loss function
$F^*(x)$	a function mapping x to y
P	a set of parameters
$h(x;\ a)$	a generic parameterized function
a	a set of parameters
m	the m^{th} adaptive (parameterized) simple function
s_t	the best split for each node t
$\mathit{\Delta}_i(s, t)$	the decrease of some impurity measure i_t
t_L	the left child nodes
t_L	the right child nodes
N_T	trees in the forest
t	node
K	variables
$Imp(X_m)$	the weighted impurity decrease
p_t	the proportion $\frac{N_t}{N}$ of training samples reaching node t
$v(s_t)$	the feature used in split s_t
R	a ranked list of variables
D	the original training data
T	the test set
D_i	a training set
C_i	a base learner
x	test record
$C_i(x)$	a classifier
$C^*(x)$	the class
$\{a_m\}_0^M$	a set of parameters

$\{\beta_m\}_0^M$	expansion coefficients
$F_0(x)$	an initial guess
$L(y, F(x))$	differentiable loss function
$h(x;\ a)$	a function that is fitted to the current “pseudo”-residuals by least squares
y_{im}	the current “pseudo” –residuals
P	a set of minority class samples
N	a set of majority class samples
s_i	the number of iterations to train the gradient boosting ensemble
Θ_κ	a random vector
$\bar{\rho}$	the average correlation between the trees
k_{th}	a tree
s	a quantity that expresses the “strength” of the tree classifiers
$margin, M(X, Y)$	a classifier’s margin
$\widehat{Y_\theta}$	the predicted class of X based on the random vector θ

LIST OF ABBREVIATIONS

Acc	Accuracy
AI	Artificial Intelligence
AML	Acute Myeloid Leukemia
ANNs	Artificial Neural Networks
ATRA	All-Trans Retinoic Acid
AUC	Area Under The Curve
BC	Breast Cancer
BD4BC	Big Data for Breast Cancer
BNs	Bayesian Networks
CC	Colon Cancer
CCA	Canonical Correlation Analysis
CCLE	Cancer Cell Line Encyclopedia
cDNA	complementary DNA
CNNs	Convolution Neural Networks
CPD	Conditional Probability Distribution
DBmcmc	Dynamic Bayesian Markov Chain Monte Carlo
DBNs	Dynamic Bayesian Networks
DEGs	Differentially Expressed Genes
DL	Deep Learning
DNA	Deoxyribonucleic acid
DNNs	Deep Neural Networks
DTs	Decision Trees
f1	F1 score
FDR	False Discovery Rate
GB	Gradient Boosting
GCO	Global Cancer Observatory

GEO	Gene Expression Omnibus
GRC	Genome Reference Consortium
GSE	Gene Expression Series
GT	Generic Transcription
GWAS	Genome Wide Association Studies
HDI	Human Development Index
HGP	Human Genome Project
HPRD	Human Protein Reference Database
IARC	International Agency for Research on Cancer
IHGSC	International Human Genome Sequencing Consortium
KIRC	Kidney Renal Clear Cell Carcinoma
LR	Logistic Regression
MALT	Mucosa-Associated Lymphoid Tissue
MDS	Myelodysplastic Syndrome
miRNA	micro RNA
MKL	Multiple Kernel Learning
ML	Machine Learning
mRNA	messenger RNA
MRs	Master Regulators
MS	Mass Spectrometry
NB	Naive Bayes
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NHL	Non-Hodgkin Lymphoma
OSCC	Oral Squamous Cell Carcinoma
PCR	Polymerase Chain Reaction
PDAC	Pancreatic ductal adenocarcinoma
PLS	Partial Least Squares
PNEP	Pre-NOTCH Expression & Processing pathway
RFs	Random Forests
RNA	Ribonucleic acid
RNNs	Recursive Neural Networks

ROC	Receiver Operating Characteristic
SAM	Significance Analysis of Microarrays
SD	Standard Deviation
Sens	Sensitivity
SNP	Single Nucleotide Polymorphisms
SNVs	Single Nuclotide Variations
Spec	Specificity
SS	Sjögren's Syndrome
SV	Structural Variations
SVMs	Support Vector Machines
TCGA	The Cancer Genome Atlas
TFBSs	Transcrition Fatcor Binding Sites
TFs	Transcription Factors
tRNA	Tranfer RNA
US	United States

ABSTRACT

The present thesis deals with the modeling of cancer diagnosis, prognosis and treatment by utilizing and implementing well-established computational approaches that can efficiently and effectively contribute to cancer care research and precision oncology. The main objective of this thesis is to study and further understand the molecular basis underlying cancer progression and risk prediction by combining high-throughput data with patient information. Towards this direction, we seek to investigate how the integration of heterogeneous datasets related to cancer development, such as genomic changes and single nucleotide polymorphisms, could provide subsequently a better understanding on cancer classification and progression based on Dynamic Bayesian Networks (DBNs) and ensemble Machine Learning (ML) methodologies, respectively.

The first part of the thesis concerns the interactions of the molecules and especially of differentially expressed genes (DEGs) that contribute to cancer progression. Based on this knowledge the identification of DEGS and their related molecular pathways is therefore of great importance. We exploited DEGs in order to further perform pathway enrichment analysis. According to our results we found significant pathways in which the disease associated genes have been identified as strongly enriched. Based on the performed pathway analysis we further proposed a methodology for predicting oral cancer recurrence using DBNs. The methodology takes into consideration time series gene expression data in order to predict a disease recurrence. Subsequently, we can conjecture about the causal interactions between genes in consecutive time intervals. A considerable overall performance of the predictive models was achieved with reference to the knowledge obtained from the pathway level. Cancer classification through DBN-based approaches that could reveal the importance of exploiting knowledge from statistically significant genes and key regulatory molecules was also explored. We identified the genes that act as regulators and mediate the activity of transcription factors that have been found in all promoters of our list with DEGs. These features serve as potential priors for distinguishing tumour from normal samples using a

DBN-based classification approach. We employed three microarray datasets from the Gene Expression Omnibus (GEO) public functional repository and performed differential expression analysis. Promoter and pathway analysis of the identified genes revealed the key regulators which influence the transcription mechanisms of these genes. We applied the DBN algorithm on selected genes and identified the features that can accurately classify the samples into wild type and controls. Both accuracy and area under the receiver operating characteristic (ROC) curve (AUC) were high for the gene sets comprising of the DEGS along with their master regulators.

In the second part of the thesis we explored the contribution of the genetic susceptibility patients’ profiles and by combining them with known clinical, histological and serological risk factors we enhanced the accuracy of predicting lymphoma development in this patient population. The potential predictive role of both genetic variants and laboratory risk factors were investigated through a ML-based framework which encapsulated ensemble classifiers, such as Gradient Boosting (GB) and Random Forests (RFs) with Gini and entropy measures. Ensemble methods enhance the classification accuracy with approaches that are sensitive to minor perturbations in the training phase. The evaluation of the proposed methodology based on a 10-fold stratified cross validation procedure yielded considerable results in terms of balanced accuracy. The initial clinical, histological and serological findings at an early diagnosis were exploited to establish ML-based predictive tools in clinical practice and further enhance our understanding towards cancer development.

In the present thesis, we studied the potential of integrating transcriptomic data with knowledge from the pathway level to model cancer progression and patient risk stratification. The development and application of novel DBN-based analysis methods allowed to infer models that could classify different phenotypes into groups with high classification accuracy. We also demonstrated that robust ensemble ML-based models could contribute to the prediction of cancer development based on the integration of genotype data along with clinical information; thus, contributing to improved disease prognosis and treatment.

ΕΚΤΕΤΑΜΕΝΗ ΠΕΡΙΛΗΨΗ

Η παρούσα διατριβή πραγματεύεται τη μοντελοποίηση της διάγνωσης, της πρόγνωσης και της θεραπείας του καρκίνου, αναλύοντας και εφαρμόζοντας ευρέως χρησιμοποιούμενες υπολογιστικές προσεγγίσεις οι οποίες μπορούν αποτελεσματικά να συμβάλουν στην έρευνα για τη διαχείριση του καρκίνου, την ογκολογία και την ιατρική ακριβείας. Κύριος στόχος αυτής της διατριβής είναι να μελετήσει και να κατανοήσει περαιτέρω τη μοριακή βάση της εξέλιξης του καρκίνου και την πρόβλεψη κινδύνου συνδυάζοντας τα ιατρικά δεδομένα του ασθενούς με δεδομένα υψηλής απόδοσης. Προς αυτή την κατεύθυνση, επιδιώξαμε να διερευνήσουμε τον τρόπο με τον οποίο η ενσωμάτωση ετερογενών συνόλων δεδομένων που σχετίζονται με την ανάπτυξη του καρκίνου, όπως οι γονιδιωματικές αλλαγές και οι πολυμορφισμοί ενός νουκλεοτιδίου, θα μπορούσε στη συνέχεια να επιτρέψει την καλύτερη και πιο έγκυρη ταξινόμηση διαφορετικών φαινοτύπων σχετικών με την εξέλιξη του καρκίνου.

Ο τομέας της Συστημικής Βιολογίας έχει αναπτυχθεί σημαντικά τα τελευταία χρόνια και αφορά την ερμηνεία συγκεκριμένων βιολογικών συμβάντων χρησιμοποιώντας τη θεωρία των συστημάτων και των δικτύων. Τα βιολογικά δίκτυα ή δίκτυα γονιδιακής αλληλεπίδρασης, αποτελούν κοινή έννοια στη Συστημική Βιολογία ενώ ταυτόχρονα παρέχουν κρίσιμη πληροφορία σχετικά με τους βιολογικούς μηχανισμούς των υγιών και μη υγιών φαινοτύπων.

Στην παρούσα διδακτορική διατριβή, δεδομένα γονιδιακής έκφρασης τα οποία έχουν εξαχθεί από διατάξεις μικροσυστοιχιών, αναλύονται περαιτέρω με στόχο την μοντελοποίηση δικτύων μέσω της τεχνικής των Μπαγιεσιανών (Bayesian) και Δυναμικών Μπαγιεσιανών (Dynamic Bayesian) δικτύων. Απώτερος στόχος είναι η ανίχνευση αλληλεπιδράσεων και σχέσεων μεταξύ των γονιδίων, καθώς και η διεξοδική ανάλυση των παραγόμενων δικτύων αλληλεπίδρασης που συμβάλλουν στη διάγνωση και πρόγνωση της ασθένειας του καρκίνου καθώς και στην ταξινόμηση των δειγμάτων σε διαφορετικές κλάσεις. Επιπλέον, μελετήθηκαν και υλοποιήθηκαν σειρά αλγορίθμων Μηχανικής

Μάθησης (Machine Learning) με σκοπό την αναπαράσταση της γνώσης και την εξαγωγή συμπερασμάτων αναφορικά με τα κλινικά, ιστολογικά και γενετικά ευρήματα ασθενών σε πρώιμη διάγνωση τα οποία αξιοποιήθηκαν περαιτέρω σε μια προσπάθεια δημιουργίας μοντέλων πρόβλεψης στην κλινική πρακτική και την ενίσχυσης της κατανόησής μας για την ανάπτυξη λεμφώματος.

Το πρώτο μέρος της διατριβής αναφέρεται στις αλληλεπιδράσεις των μορίων και ιδιαίτερα των διαφορικά εκφρασμένων γονιδίων (differentially expressed genes) που συμβάλλουν στην διάγνωση και εξέλιξη της νόσου του καρκίνου. Με βάση αυτή την γνώση, ο προσδιορισμός και η αναγνώριση των διαφορικά εκφρασμένων γονιδίων και των σχετικών μοριακών μονοπατιών στα οποία συμμετέχουν είναι μεγάλης σημασίας. Εκμεταλλευτήκαμε τα σημαντικά ως προς την έκφρασή τους γονίδια για να πραγματοποιήσουμε περαιτέρω ανάλυση των βιολογικών μονοπατιών. Σύμφωνα με τα αποτελέσματα, προσδιορίσαμε σημαντικές βιολογικές οδούς στις οποίες τα γονίδια που σχετίζονται με την ανάπτυξη καρκίνου έχουν αναγνωριστεί ως έντονα εμπλουτισμένες και συμμετέχουν σε αυτές. Με βάση την ανάλυση που πραγματοποιήθηκε, προτείναμε μεθοδολογία για την πρόβλεψη της υποτροπής του καρκίνου του στόματος χρησιμοποιώντας Δυναμικά Μπαγιεσιανά δίκτυα. Η προτεινόμενη μεθοδολογία δέχεται ως είσοδο δεδομένα έκφρασης γονιδίων από διάφορες χρονικές στιγμές προκειμένου να προβλέψει την υποτροπή της νόσου. Στη συνέχεια και βάσει της μεθόδου των δυναμικών δικτύων, μπορούμε να εξάγουμε υποθέσεις για τις αιτιώδεις αλληλεπιδράσεις μεταξύ των γονιδίων σε διαδοχικά χρονικά διαστήματα. Επιτεύχθηκε έτσι η ανάπτυξη έγκυρων και ακριβών μοντέλων πρόβλεψης με αναφορά στα δεδομένα που αποκτήθηκαν από το επίπεδο των βιολογικών μονοπατιών στα οποία συμμετέχουν τα γονίδια προς μελέτη. Τα δεδομένα που αναφέρθηκαν παραπάνω χρησιμοποιήθηκαν ώστε να καθοριστούν η δομή και οι παράμετροι δύο μοντέλων Δυναμικών Μπαγιεσιανών δικτύων που σχετίζονται με την κατάσταση συγκεκριμένων ασθενών, δηλαδή εκείνων που επανεμφάνισαν ή όχι καρκίνο. Οι παράμετροι προσδιορίστηκαν μεταξύ των μεταβλητών του πρώτου χρονικού διαστήματος και κατά τη διάρκεια του πρώτου και δεύτερου διαστήματος. Έτσι, μπορέσαμε να υποθέσουμε σχετικά με τις σχέσεις - αλληλεπιδράσεις μεταξύ των γονιδίων. Επιπλέον, η χαρτογράφηση αυτών των αλληλεπιδράσεων με γνωστές και επαληθευμένες αλληλεπιδράσεις στην βιβλιογραφία είναι σε θέση να προσφέρει καλύτερη εικόνα στις υποκείμενες μοριακές διεργασίες της νόσου. Η συνολική απόδοση των μοντέλων πρόβλεψης ήταν ίση με 81,8% ακρίβεια και περιοχή κάτω από την ROC καμπύλη ίση με 0.892, αναφορικά με τις γνώσεις που αποκτήθηκαν από την ανάλυση εμπλουτισμού των σηματοδοτικών μονοπατιών.

Στην συνέχεια, διερευνήθηκε η ταξινόμηση των ασθενών με καρκίνο σε προκαθορισμένες κλάσεις μέσω προσεγγίσεων που βασίζονται και πάλι στα Δυναμικά Μπαγιεσιανά δίκτυα τα οποία επιτρέπουν τη συνεκμετάλλευση της γνώσης από στατιστικά σημαντικά γονίδια και τα βασικά ρυθμιστικά τους μόρια. Προσδιορίσαμε τα γονίδια που λειτουργούν ως ρυθμιστές και μεσολαβούν στη δραστηριότητα παραγόντων μεταγραφής τα οποία έχουν βρεθεί σε όλους τους υποκινητές της λίστας με τα διαφορικά εκφρασμένα γονίδια. Τα χαρακτηριστικά αυτά χρησιμοποιήθηκαν ως προγενέστερη γνώση στα Δυναμικά Μπαγιεσιανά δίκτυα για τη διάκριση του όγκου από τα υγιή δείγματα. Χρησιμοποιήσαμε τρία σύνολα δεδομένων μικροσυστοιχιών από το αποθετήριο Gene Expression Omnibus (GEO) και πραγματοποιήσαμε αρχικά ανάλυση διαφορικής έκφρασης. Η μετέπειτα ανάλυση των υποκινητών και των σηματοδοτικών οδών των αναγνωρισμένων γονιδίων αποκάλυψε τους βασικούς ρυθμιστές που επηρεάζουν τους μηχανισμούς μεταγραφής των εν λόγω γονιδίων. Εφαρμόσαμε τον προτεινόμενο αλγόριθμο σε επιλεγμένα γονίδια και προσδιορίσαμε τα χαρακτηριστικά που μπορούν να ταξινομήσουν με ακρίβεια τα δείγματα στις ομάδες ελέγχου (controls) και άγριου τύπου (wild type). Τόσο η ακρίβεια όσο και η περιοχή κάτω από την καμπύλη ROC ήταν υψηλές, στηριζόμενοι στα τελικά σύνολα γονιδίων (δηλαδή στα διαφορικά εκφρασμένα γονίδια και τους υποκινητές τους). Συγκεκριμένα, η ακρίβεια κυμάνθηκε μεταξύ 70,8% - 98,5%, ενώ η καμπύλη ROC μεταξύ 0,562 - 0,985.

Στο δεύτερο μέρος της διατριβής μελετήσαμε τη συμβολή των προφίλ γενετικής ευαισθησίας σε ασθενείς με σύνδρομο Sjögren. Συνδυάζοντας τα γενετικά δεδομένα με γνωστούς κλινικούς, ιστολογικούς και ορολογικούς παράγοντες κινδύνου, ενισχύσαμε την ακρίβεια της πρόβλεψης ανάπτυξης λεμφώματος σε αυτόν τον πληθυσμό ασθενών. Ο δυνητικός προγνωστικός ρόλος τόσο των γενετικών παραλλαγών όσο και των εργαστηριακών παραγόντων κινδύνου διερευνήθηκε μέσω μεθοδολογίας Μηχανικής Μάθησης, η οποία ενσωματώνει ταξινομητές, όπως ο ταξινομητής Ενίσχυσης Σύστασης (Gradient Boosting - GB) και τα Τυχαία Δέντρα (Random Forests - RFs) με συγκεκριμένα μέτρα εντροπίας. Οι μέθοδοι συνόλου (ensemble) που αναπτύχθηκαν βελτίωσαν την ακρίβεια της ταξινόμησης των ασθενών βάσει προσεγγίσεων ευαίσθητων σε μικρές διακυμάνσεις στη φάση της εκπαίδευσης. Η αξιολόγηση της προτεινόμενης μεθοδολογίας έγινε με διαδικασία διασταυρούμενης επικύρωσης και έδωσε σημαντικά αποτελέσματα ως προς την ακρίβεια, την ευαισθησία και την ειδικότητα (GB: ακρίβεια = 0.7780, RF με ευρετήριο Gini: ακρίβεια = 0.7626, RF με εντροπία: ακρίβεια = 0.7590). Επομένως, τα κλινικά, ιστολογικά και ορολογικά ευρήματα κατά την πρώιμη διάγνωση χρησιμοποιήθηκαν στον σχεδιασμό προγνωστικού μοντέλου που βασίζεται σε τεχνικές Μηχανικής Μάθησης και έχει ως στόχο την εφαρμογή του στην κλινική πράξη ενισχύοντας περαιτέρω την κατανόησή μας για την ανάπτυξη του καρκίνου.

Συνοψίζοντας, στην παρούσα διατριβή, μελετήσαμε τις δυνατότητες συνδυασμού δεδομένων μεταγραφής με γνώσεις από σηματοδοτικά μονοπάτια στα οποία συμμετέχουν γονίδια σημαντικά στην εξέλιξη του καρκίνου, με στόχο την διαστρωμάτωση του κινδύνου των ασθενών. Η εφαρμογή νέων μεθόδων ανάλυσης που βασίζονται σε Δυναμικά Μπαγιεσιανά δίκτυα επέτρεψε την ανάπτυξη μοντέλων ικανών να ταξινομήσουν στις επιμέρους ομάδες διαφορετικούς φαινοτύπους με υψηλή ακρίβεια. Δείξαμε επίσης, ότι τα προβλεπτικά μοντέλα που βασίζονται σε τεχνικές Μηχανικής Μάθησης μπορούν να συμβάλουν στην πρόβλεψη της ανάπτυξης του καρκίνου μέσω της ενσωμάτωσης δεδομένων γενωμικής στις υπάρχουσες κλινικές πληροφορίες, συμβάλλοντας έτσι στη βελτίωση της πρόγνωσης και της θεραπείας της νόσου.

CHAPTER 1 INTRODUCTION

1.1 Background and Thesis Motivation

Cancer is a genetic disease with diverse subtypes according to the tumor type, and it can start in almost any organ or tissue of the body [1-3]. It is characterized by abnormal cell growth that invades uncontrollably healthy cells in the body and interfere with the function of normal tissues and organs, initiating thereby metastases that leads eventually to death. The early diagnosis and prognosis of a cancer type as well as cancer prevention has become a necessity in cancer research, as it can facilitate the subsequent optimization of cancer treatment and the improvement of patients' management in clinical practice [4].

According to the status report on the global cancer mortality and incidence [5], cancer burden is expected to grow drastically by 2040 to 27.5 million new cancer cases and 16.3 million cancer deaths based on the growth and aging of the population [6]. Due to the increasing prevalence of risk factors, such as smoking, obesity, UV radiation and physical inactivity, cancer burden will be even larger in the future; thus, resulting in the tumor development. The estimated number of incidences across all cancer sites at the age of 0-69 years, is expected to increase by 5.4% for males and by 4.6% for females in 2020 comparing to 2018 estimates, according to the Global Cancer Observatory (GCO) [7, 8].

The European Society for Medical Oncology (ESMO) [9] and the American Society of Clinical Oncology (ASCO) [10] develop and publish clinical practice guidelines and opinions, for providing evidence-based recommendations that will serve as a guide for both doctors and researchers to select appropriate methods for cancer treatment and care. The vision of these initiatives is to offer the best of care to cancer patients through tailoring treatment and fostering new precise cancer care that will support oncologists in their professional development while maintaining the sustainability of care worldwide for people suffering from cancer.

Rapid advances in cancer research during the last decades revealed that cancer is a disease which includes dynamic genomic changes. Several molecular, cellular and biochemical characteristics have been suggested as the acquired capabilities that are shared by almost all types of human cancers. According to [2], a set of rules have been studied to provide the clues that control the transformation of normal cells into malignant tumors. In addition, a variety of published studies specify that tumor growth in humans is a multistage procedure which reveals the genetic mutations that govern the alteration of normal human cells [11-15]. Cancer heterogeneity refers not only to the complex network of interacting signaling pathways, but also to the interactions among cancer cells and their microenvironment. Hence, the increased cancer complexity is based on the large number of interacting molecules, the information exchange between pathways and the (non-)linear connections between the molecules. This multiparametric functioning of a system defines cancer as a systems biology disease. Consequently, there is a recent trend within the cancer research community to study cancer as a complex biological system and thus predict its behavior.

With the advent of modern technologies, the scientific community has embraced the promise of high-throughput sequencing, microarray technology, and other large-scale approaches for exploring many questions related to cancer diagnosis, prognosis and treatment. One of the main objectives of microarray experiments is the class prediction based on gene expression data, which concerns the creation of gene-based predictive models that can be applied to new samples to assign the class labels and further clinical decisions (i.e. who will and who will not suffer from a disease relapse). Moreover, the identification of different types of cancer based on gene expression profiles of the tumors has been studied extensively in the literature. Microarray data has been also used to identify DEGs that may influence tumor progression, metastases, and survival outcomes.

Towards this direction, in the era of precision oncology, data-driven approaches have been proposed in terms of computational and especially Artificial Intelligence (AI)-based models for improving the decision making in healthcare systems. Predictive modeling frameworks have been established and facilitated the combination of heterogenous data sources (i.e. clinical variables, imaging and omics data) for further utilization regarding the clinical unmet needs. In cancer research, the introduction of well-known machine learning approaches and newly introduced methods such as deep learning and multi-modal integrative schemes paved the way for predictive capabilities in precision oncology aiming at patient stratification and risk prediction. To date, computational models with advanced predictive capabilities reveal the potential usefulness of ML algorithms which play a prominent role in accelerating the application of robust methodologies for the clustering, correlation, and classification of various data views towards patient risk stratification as well as a more precise diagnosis, prognosis, and treatment of cancers.

In order to provide a better understanding on how cancer progresses across time and during the follow-up period of diagnosed patients and further comprehend the medical and genomic features for the clinical diagnosis and treatment of cancer, we herein propose certain computational approaches that have been adopted and developed for modeling cancer data. Based on the large amount of available biomedical data we investigated the modeling of gene expression measurements and the identification of DEGs among samples of different phenotypes for classification purposes. Special emphasis was given to the integration of knowledge from the pathway level based on the transcription factors of the significant genes in order to extract models that predict accurately the class labels of new samples. DBNs was employed aiming at identifying the changes of gene interactions in terms of gene expression data. Hence, we could model cancer progression and therefore conjecture about the underlying relationships among genes for classification purposes. Going one step further and motivated by the potential usefulness of integrating genetic information along with clinical findings we proposed a ML-based framework aiming at investigating the contribution of the genetic susceptibility profiles of patients at the time of disease diagnosis for predicting the risk for lymphoma development.

On this basis, individualized treatment for the clinical management of cancer or diagnosis could be enhanced. The data-driven analysis of all biological information facilitates the detection of differences in gene expression due to the phenotype of each cancer sample. Therefore, predictive models could be a valuable and increasingly necessary tool for elucidating the behavior of cancer in clinical practice and modern healthcare systems.

1.2 Overview of the Thesis

The present thesis is structured as follows:

In the following chapter, an introduction to cancer biology and its causes is given with details about the hallmarks that have been assessed toward elucidating cancer development. We briefly discuss cancer genetics and the high throughput technologies that are used for producing the relevant biological data. Genes that have been found to initiate cancer progression along with the pathways of cancer pathogenesis are also presented. A summary report related to cancer descriptive epidemiology is provided as well as a description of the cancer burden today and in the future. Finally, a thorough explanation can be found in this chapter concerning the biomedical data that are stored and analyzed towards cancer management.

The third chapter concerns the use of data science and AI in cancer research. The big data era is described with reference to initiatives that highlight the data for data science and big data in order to improve patient care and optimize cancer therapies. Furthermore, special emphasis is given on the use of AI and ML in precision oncology. We give details about well-established ML applications in cancer, prognosis and survival. The modern approach of Deep Learning (DL) is also described which is used currently for enhancing cancer diagnosis and classification. A few paradigms are also presented with reference to the expandability and reproducibility of ML models and how the new findings and results can be validated and reproduced by separate research groups.

In the fourth chapter a literature overview on modeling gene expression data by means of BNs and DBNs is given for cancer prognosis prediction. Several BNs approaches are presented for the analysis of gene expression microarray data. Other ML methodologies that have been proposed in the literature for modeling cancer diagnosis, prognosis and treatment are also given with emphasis on cancer prediction and survival assessment. In addition, ensemble ML-based models for cancer prognosis and prediction are introduced aiming at pinpointing the multi-modal fusion strategies that can be applied at both the feature and the decision levels based on AI and ML techniques as a promising framework for cancer management and better decision making in clinical practice.

The fifth chapter presents our first proposed methodology for predicting cancer recurrence through DBNs. Transcriptomic data are utilized for identifying the DEGs among different samples. We perform a pathway enrichment analysis and further predict the disease recurrence through the utilization of DBN models while an overrepresentation analysis is conducted in order to detect the pathways that are enriched in the defined gene set. Promising results were derived in terms of cancer prediction and we showed that the combination of the specific gene set with the highly connected nodes from the selected pathway can provide accurate disease recurrence prediction.

The sixth chapter deals with the next proposed methodology for classifying cancer samples and tissues based on time series gene expression data as well as on regulatory molecules. In this study we investigate the transcription factors and master regulators that are involved in the provided list of DEGs. DBNs are also employed with reference to the prediction of the class label of each provided sample. We performed an upstream analysis and further model the relative pathway data for developing gene regulatory networks from microarray time series gene expression data for cancer classification through DBNs.

In the seventh chapter we introduce a ML-based methodology with ensemble classifiers aiming at exploring the contribution of combined initial clinical, serological and histopathological features with genetic variants in predicting lymphoma development. A robust pipeline with a list of estimators was developed for the accurate prediction of cancer risk. The sequential application of certain preprocessing steps, class imbalance handling, and model's performance evaluation constitute the main procedure followed for predicting caner development.

Finally, in the last chapter we discuss the main findings of the current thesis in accordance to the literature and state-of-the-art studies. The plan for our future steps for advancing the proposed computational methodologies is given and several key ideas are presented related to the integration of different data sources along with the application of multi-modal deep learning frameworks that facilitate the design and deployment of predictive models towards improved cancer management.

CHAPTER 2 BACKGROUND ON CANCER BIOLOGY AND DATA USED FOR DIAGNOSIS, PROGNOSIS AND TREATMENT

2.1 The biology of cancer

2.1.1 The development of cancer

Cancer, a broad term for a class of diseases, is characterized by the uncontrolled proliferation and spread of abnormal cells. It results from a breakdown of the regulatory mechanisms that control cell division and behavior. Cancer cells can infiltrate adjacent healthy cells and interfere with the function of normal tissues and organs, initiating the metastases and leading to death eventually. Understanding cancer at the molecular and cellular levels has been an objective for many years in the field of experimental biology. The thorough study of cancer cells has improved our understanding on the (i) regulation of the cell cycle, (ii) control of cell death and (iii) cell signaling. It is true that key molecules involved in cell regulation, have been identified by their abnormalities which contribute to the uncontrolled growth of cancer cells [1]. The study of cancer's biology empowers our understanding of the fundamentals of human cell regulation. Based on the continual uncontrolled growth and the accumulated abnormalities in cell regulatory aspects, cancer cells are well discriminated from their normal counterparts. Because cancer cells can result from any kind of cells in a healthy body, more than 100 distinct cancer types have been assessed. These cancer types differ substantially in their initiation, their progression and their response to treatment.

In cancer pathology, the characterization of a tumor (i.e. a cell mass created because of the abnormal cell proliferation) as benign or malignant is one of the most important and critical parts for the subsequent disease management. While benign tumors do not invade surrounding healthy tissues and remains in their initial locations, malignant tumors can spread to other parts of the body through the circulatory or lymphatic systems (Figure 2.1). Malignant tumors are dangerous enough and are mentioned as cancers because of their ability to metastasize. In addition, the spread of malignant tumors to other body sites makes them resistant to localized treatments, such as surgery, and cannot thus be removed.

Based on the type of cells that both benign and malignant tumors are initiated, they can be categorized into three main groups, namely: (i) carcinomas, (ii) sarcomas and (iii) leukemias or lymphomas. Sarcomas are rare in humans and refer to solid tumors of mesenchymal origin. The connective tissue that sarcomas can arise from includes bone, cartilage, fat, vascular, or hematopoietic tissues. Leukemias and lymphomas are malignancies that can arise from cells of the immune system and/or cells that form the blood. This type of tumors account for 8% of the human cancers. On the contrary, carcinomas include 90% of human malignancies and they are referring to malignancies in epithelial cells. Tumors can be further characterized based on the tissue and the cell type they are involved in.

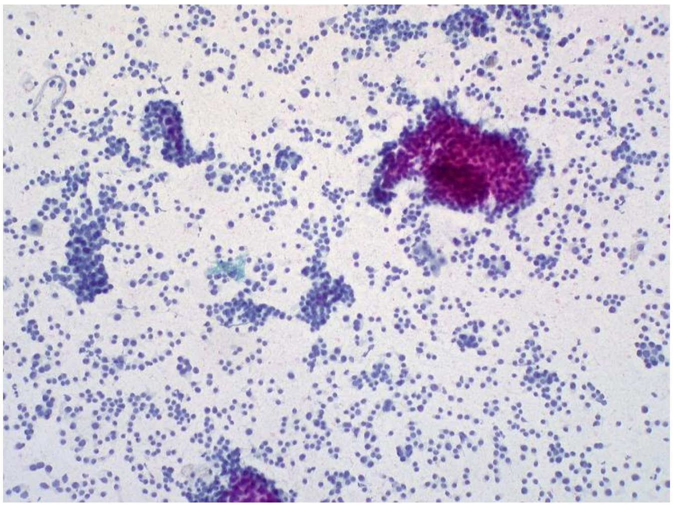


Figure 2.1 A micrograph showing carcinoid tumor, metastatic to liver, where the cancer cells have dark purple nuclei and are invading the normal tissue (pink) [1].

One of the fundamental characteristics of cancer is that tumor cells come from an original cell which begins to proliferate abnormally. In other words, the cells of a tumor constitute a cell clone, and this has been proven in many situations with reference to the chromosome X inactivation [1].

Because of the multistep process during cancer development, where the cells are transformed to malignant through progressive mutations, the progenitor cell does not adopt all the features of a cancer cell from the beginning. At the cellular level, we can consider cancer development as a multistep procedure where mutated cells that progressively proliferate, survive, invade and finally metastasize are selected. Tumor initiation is the first step of this process.

Genetic alterations enhance the uncontrolled growth of a single cell and then this proliferation allow the formation of a population of clonally derived tumor cells. When additional mutations occur within the cancer cells, the tumor progression continues. Some of these mutations enable the cell to adopt selective advantages, such as more rapid growth.

Therefore, the descendants of the cell with such mutations will become dominant within the cell population of the tumor. This is called clone selection, and properties such as increased growth rate, survival, invasion or metastasis characterize the new clones of tumor cells giving them a selective advantage. During the development of the tumor clonal selection persist; thus, tumors evolve more rapidly while they become increasingly malignant.

Figure 2.2 presents a clear example of tumor progression and clonal selection of colon carcinoma. The increased proliferation of colon epithelial cells constitutes the very first step during tumor development. A benign neoplasm (an adenoma or polyp) of increasing size is then created by one of the cells resulting in tumor population. During the next clonal selection steps the adenomas is growing and it also increases in size. Later, the benign adenomas give rise to the malignant carcinomas which is denoted by the invasion of the tumor cells into the underlying connective tissue. Apparently, the cancer cells continue to proliferate and spread through the connective tissues [2, 3]. Finally, other abdominal organs (such as the small intestine) are invaded by the cancer cells after penetrating the blood and lymphatic vessels. As mentioned above, the rapid uncontrolled tumor growth results from accumulated mutations that affect regulatory mechanisms of the cell. These

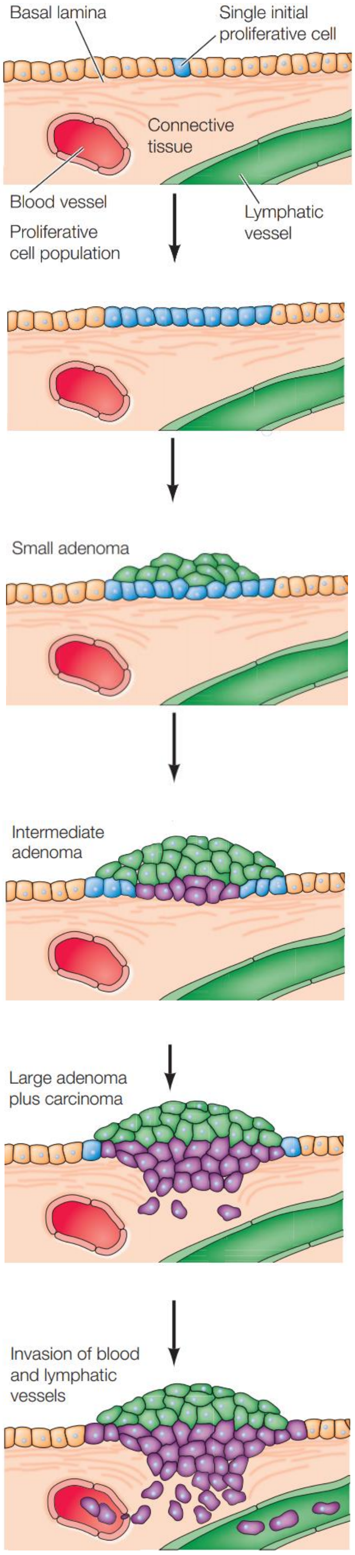

Basal lamina
Single initial proliferative cell
Connective tissue
Blood vessel
Lymphatic vessel
Proliferative cell population
Small adenoma
Intermediate adenoma
Large adenoma plus carcinoma
Invasion of blood and lymphatic vessels

Figure 2.2 Colon cancer development. A single mutated cell starts to proliferate abnormally. A benign adenoma of increasing size is firstly created and then results in malignant carcinoma. Invasion of the cancer cells into the underlying connective tissue occurs, which then penetrate blood and lymphatic vessels [2, 3].

mechanisms concern the normal cell proliferation, differentiation, and survival. Cancer cells are therefore characterized by abnormalities in certain regulatory mechanisms and display attributes which distinguish them from their normal counterparts. The cancer cell characteristics constitute the main study objective of malignancy, at the cellular level. The primary differentiations among cancer and normal cells are (i) the density-dependent inhibition and contact inhibition of their growth and movement, (ii) the production of growth factors that stimulate their own proliferation and (iii) the less adhesive behavior of most cancer cells than normal [1].

2.1.2 The causes of cancer

Carcinogens are the primary initiators of cancer. They have been identified either by experimental studies and/or by epidemiological studies based on the analysis of human cancer frequencies (e.g. high prevalence of lung cancer in the cigarette smokers). Due to the complex nature of cancer which reflects a multistep process, it is very simplistic to talk about unique reasons that cause cancer. Many factors can affect the possibility of cancer incidence, with radiation, chemicals, and viruses being the most prominent. Most chemical carcinogens as well as radiation act by inducing mutations in the genome. Solar ultraviolent, the major cause of skin cancer, and the carcinogenic chemicals in tobacco smoke are among the main carcinogens that contribute to the development of human cancers. Aflatoxin, another liver carcinogen that is produced by some molds that affect stored grain supplies, is also considered to contribute to cell mutations [1]. Carcinogens included within the tobacco smoke are the main cause of nearly 90% of lung cancers while they are also involved in cancers of the oral cavity, pharynx, larynx, esophagus, and other sites [1]. It is estimated that tobacco smoking is responsible for one-third of cancer deaths.

Some carcinogens, known as tumor promoters, facilitate the increased cell division; thus, invoking the growth of a cell population that proliferates at the early stages of tumor development. This abnormal cell proliferation occurs by mutations during Deoxyribonucleic acid (DNA) replication. Moreover, several kinds of viruses can also cause cancers (Table 2.1), such as liver cancer and cervical carcinoma. The frequency rate of this type of cancers stands for 10-20% worldwide [1]. Experimental studies on tumor viruses has contributed to the elucidation of the molecular events responsible for human cancers development by both viral and nonviral carcinogens. While chemical carcinogens act by inducing mutations in cellular genes, tumor viruses introduce new genetic material into contaminated cells. Tumor viruses are characterized by small genome size revealing our ability to detect through molecular analysis the viral genes responsible for cancer development. This knowledge paves the way to better understand mutations at the molecular level.

Table 2.1 Tumor viruses [1]. The virus family of each human tumor is presented alongside the genome size in kilobase (kb).

Virus family	Human tumors	Genome size (kb)
DNA Genomes		
Hepatitis B virus	Liver cancer	3
Polyomaviruses and SV40	Merkel cell carcinoma	5
Papillomaviruses	Cervical carcinoma	8
Adenoviruses	-	35
Herpesviruses	Burkitt's lymphoma, nasopharyngeal carcinoma, Kaposi's sarcoma	100-200
Ribonucleic acid (RNA) Genomes		
Hepatitis C virus	Liver cancer	10
Retroviruses	Adult T-cell leukemia	9-10

2.1.3 The hallmarks of cancer

During the multistep process of human tumor development, several traits need to be acquired by the cancer cells to become ultimately tumorigenic and malignant [3]. In 2000, the six hallmarks of cancer were established enabling the better understanding of the diversity of neoplastic diseases [2, 3]. For the normal cells to progress to a neoplastic state (i.e. the state of excessive and abnormal growth of cells known as tumor), they should successfully acquire one or more of the hallmark capabilities.

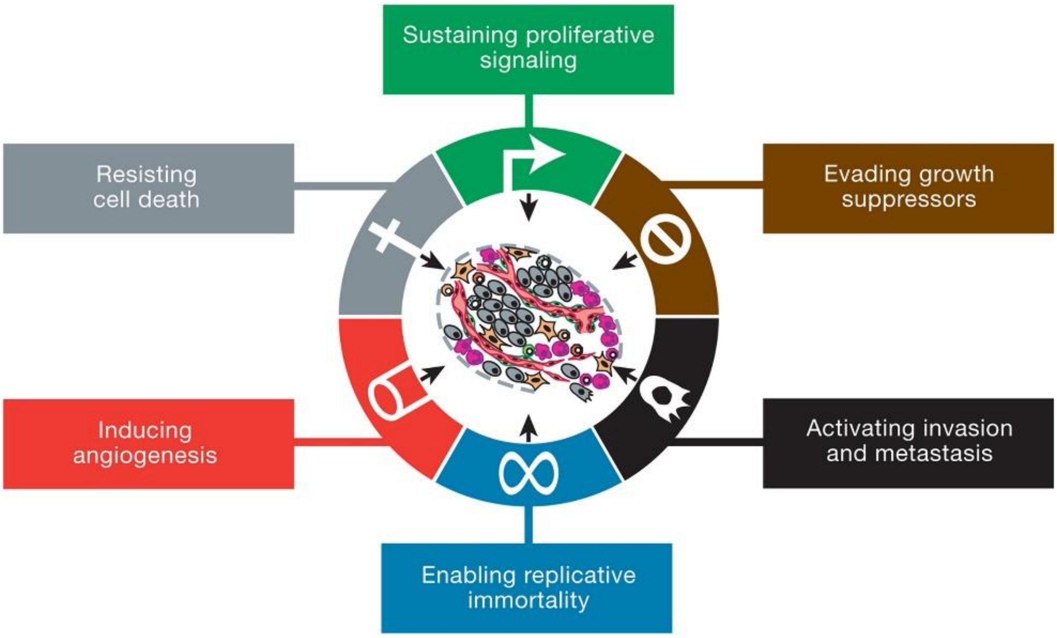

Sustaining proliferative signaling
Resisting cell death
Evading growth suppressors
Inducing angiogenesis
Activating invasion and metastasis
Enabling replicative immortality

Figure 2.3 The hallmarks of cancer as originally proposed in 2000 [2].

Tumors can be characterized as complex tissues which are composed of distinct cell types that interact with one another. The last decade, “tumor microenvironment” has been also identified to contribute to tumorigenesis, demonstrating that tumor biology can no longer be understood simply by the six primary traits of cancer cells which enable tumor growth and metastases. The distinctive and complementary hallmarks of cancer provide therefore a general framework for understanding the biology of this complex disease and the tumor development (Figure 2.3).

2.1.3.1 Sustaining Proliferative Signaling

Chronic proliferation is the most fundamental characteristic of cancer cells. The normal growth and conservation of living organisms is based on a complex and extremely accurate system which controls the cellular growth and differentiation. Therefore, the normal tissues ensure a homeostasis of cell growth and maintenance. On the contrary, cancer cells cause the aberration from normal cell growth leading to creation of abnormal somatic cells which cause then tumor development. The signals that are dysregulated by the malignant cells are propagated typically by growth factors that bind receptors in the cell-surface. In general, the sources that convey the proliferative signals to normal cells remain unknown. These mechanisms are poorly understood because the growth factor signaling that controls the cells number and division is transmitted in a temporal and spatial manner from one cell to its neighbors [2, 3].

2.1.3.2 Evading growth suppressors

Cancer cells can induce and sustaining growth signals towards their preservation. However, they must also inhibit powerful mechanisms that depend on the action of tumor suppressor genes. These genes represent the opposite side of cell growth control and cell proliferation and they act normally to circumvent tumor development [2, 3]. During cancer development, tumor suppressor genes are lost or inactivated, therefore contributing to the abnormal proliferation of tumor cells. It is worth mentioning that studies of retinoblastoma, a rare childhood eye tumor, paved the way for the identification of the first tumor suppressor gene (*Rb* retinoblastoma-associated gene). The *p53* tumor suppressor gene is the second suppressor that has been identified to be inactivated in a wide variety of human cancers. *p53* is the most common target of mutations in human cancers, including leukemias, lymphomas and sarcomas among others [16].

2.1.3.3 Resisting cell death

Programmed cell death, serves as a barrier to cancer development since it is responsible for balancing cell proliferation and maintaining the numbers of cells in tissues constant. Moreover, the programmed cell death by apoptosis (i.e. a series of cellular changes), provides a defense framework by which damaged and unwanted cells are eliminated for the normal function of the organism. The most common strategy that tumor cells develop gradually in order to limit apoptosis is the loss of function of *p53* tumor suppressor [17]. In a similar manner, tumors may influence the normal mechanisms of apoptosis by increasing the expression of antiapoptotic regulators or of survival signals [18].

2.1.3.4 Enabling replicative immortality

Cancer cells require the potential of replication towards the development of macroscopic tumors. This contrasts with the behavior of normal cells in the body which can undergo a limited number of consecutive cellular growth and division cycles. This limitation of the normal cells is associated with two barriers of proliferation, namely the (i) senescence (i.e.

a non-proliferative but viable state of the cells) and (ii) crisis/apoptosis which corresponds to programmed cell death. Furthermore, cancer cells may revert to a pre-differentiated phenotype allowing the uninhibited cellular division and other metabolic adaptations. These characteristics enable thereby the survival in adverse conditions, known as immortalization [2, 3]. Two key signaling pathways are involved, among others, in these changes enabling the cancer cells replicative immortality: (i) the Hippo signaling and (ii) the Wnt signaling pathways. Hippo signaling controls organ size by adjusting cell proliferation, apoptosis, and stem cell self-renewal. Dysregulation of this pathway contributes to cancer development [19]. The Wnt/β-Catenin pathway controls stem cell pluripotency and cell fate decisions during cells' development [20]. Both pathways are evolutionary conserved, and their alterations contribute to cancer's ability to replicate abnormally and indefinitely.

2.1.3.5 Inducing angiogenesis and activating invasion and metastasis

Additional properties of malignant cells affect their interactions with other tissues; hence, playing significant role in invasion and metastasis. First, secretion of proteases by malignant cells permits the digestion of extracellular matrix components. This enables cancer cells to invade underlying connective normal tissues. For example, as depicted in Figure 2.2, the secretion of proteases that digest collagen implies an important determinant of carcinomas to penetrate through basal laminae into adjacent tissues. Second, through the process of angiogenesis, cancer cells excrete growth factors that are known to promote the formation of new blood vessels. Angiogenesis supports the growth of a tumor beyond the size of about a million cells [2, 3]. Obviously, at this point oxygen and nutrients supplied only by new blood vessels are required towards the growth and proliferation of tumor cells. The new blood vessels are then formed in response to growth factors enhancing therefore the proliferation of endothelial cells and the growth of new capillaries into the tumor. Angiogenesis is an important hallmark capability in terms of tumor growth and metastasis. The new capillaries that are formed based on the angiogenic stimulation can be penetrated by the cancer cells allowing the aberrant cells enter the circulatory system and start the metastatic process. The available new research tools and the refined experimental models have accelerated the research for angiogenesis, invasion and metastasis. In addition, the identification of critical regulatory genes permits the elucidation of significant features of these complex hallmark traits.

During the multistep tumorigenesis, the functional hallmark capabilities that are obtained from the tumor cells in order to proliferate, survive and metastasize can be adopted at different times and in different tumor types. Two enabling characteristics and two newly proposed emerging hallmarks could facilitate then the acquisition of core capabilities and the development and progression of cancer, respectively (Figure 2.4). The most discrete characteristic is the genomic instability developed in cancer cells including rearrangements of chromosomes. These genetic changes can influence and guide the presence of hallmark traits. The second enabling characteristic corresponds to the inflammatory state of malignant and premalignant lesions which is driven by cells of the immune system and promotes tumor progression. Concerning the distinct attributes of malignant cells that enhance the possibility of cancer development, two new hallmarks have evoked our interest the last years. The first includes the deregulation of cellular energetics which support the continuous cell growth and proliferation while it influences the metabolic program that characterizes normal cells. The second attribute refers to the ability of cancer cells to avoid immune destruction. Cancer cells can evade actively the attack and elimination by immune cells. These two additional hallmarks are involved in the pathogenesis of cancer and can therefore be considered as emerging hallmarks of cancer.

2.2 Human cancer genetics

Genomic medicine has evolved from the classical clinical genetics towards exploiting the knowledge coming from the human genome analysis. Although genome analysis has been conducted for several decades, genomic medicine essentially began in 2001, when the first version of the human genome was completed.

2.2.1 The Human Genome

With the advent of large-scale sequencing in the 1980s [21], the term of genomics started to be used. For the first time, researchers were able to design new strategies for patient management based on the complete genetic knowledge of an organism. This genomics-based research has been characterized by its massive scale, including the study of features involved in hundreds of millions of nucleotides or the analysis of the expression levels of thousands of genes simultaneously. The genomics era was launched by the first sequenced bacterial genome with the sequencing of the human genome being the major event in genomics.

The human genome is found in the nucleus and mitochondria of the cells. The genome located within the nucleus comprises 23 pairs of chromosomes. Each chromosome is made up of a large linear DNA molecule. The chromosome pairs are divided into 22 autosomes, which are common in both sexes, and into a pair that differentiates the sex (i.e. XY for men and XX for women). Each cell contains hundreds of copies of mitochondrial DNA. Nuclear chromosomes are inherited by 50% of the father and 50% of the mother, whereas mitochondrial DNA originates exclusively from one's mother side.

The Human Genome Project (HGP) was presented as an idea in the middle of 1985s (eight years after the invention of Sanger sequencing described in 1977) and started officially at 1990 [22, 23]. It consisted of the sequencing and studying of the genomes of other organisms. However, the overarching goal was to determine the sequence of the 3 billion base pairs that constitute the 24 human chromosomes. In 2001, a landmark goal was met by two groups, namely: (i) a large international consortium of funded researchers and scientists (the International Human Genome Sequencing Consortium or IHGSC) and (ii) a private company called Celera Genomics. These groups independently published draft sequences of the human genome, each about 90% complete. In April 2003, a defining moment occurred

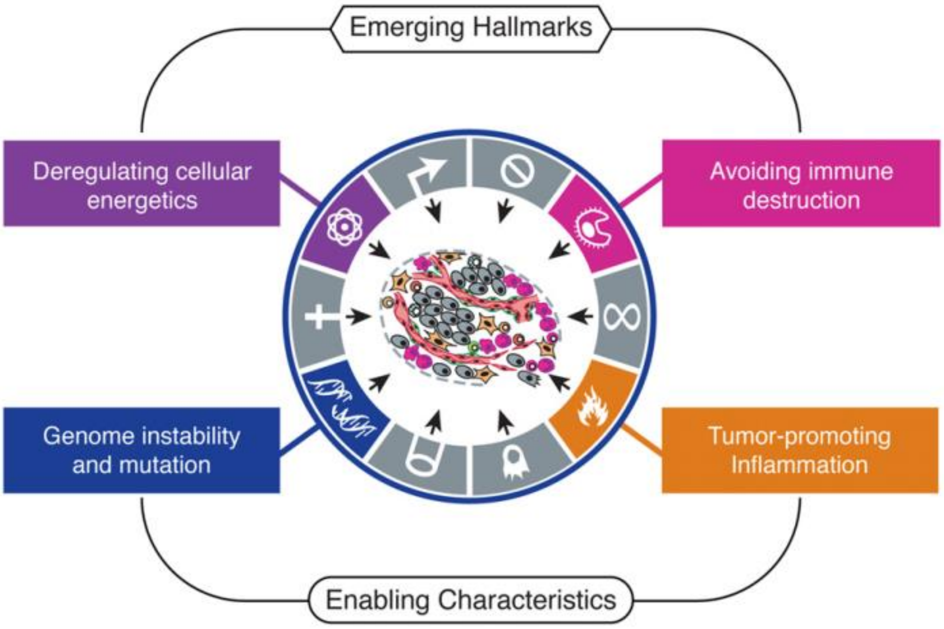

Emerging Hallmarks
Deregulating cellular energetics
Avoiding immune destruction
Genome instability and mutation
Tumor-promoting Inflammation
Enabling Characteristics

Figure 2.4 Emerging hallmarks and enabling characteristics of malignant cells which promote the development and progression of cancer [2].

with reference to the presentation of the finished reference sequence (99% completeness) of the human genome by the IHGSC. The reference sequence contained 100% of the entire sequence while it was characterized by higher quality and accuracy (>99,9%) than the draft sequences. The total cost of the first sequence reading was more than 3 billion while it took 13 years in order to be completed.

The haploid form of the genome consists of about 3 billion bases, which means that the first version had tens or even hundreds of thousands of errors. Based on this knowledge, the effort has been and is still being pursued by the Genome Reference Consortium (GRC) [24]. Hence, the latest version of the human genome, the 38th, was published in December 2013 while there are still small areas that have not been read.

The human genome, which is used as a reference genome and is the basis of all genetic analysis that is carried out, is not the result of thousands of human reads. Moreover, it does not represent the average in terms of the bases and their frequency. Instead, it is made up of parts of the genome of nine people who volunteered and donated their DNA for this purpose. Specifically, in the public effort by the IHGSC, the DNA of four people was used for reading the genome, but for technical reasons 71% of the total came from one of them. In the private effort by the Celera Genomics company, the genome of five volunteers was read. It should be noted that the five volunteers who participated in the second effort are members of the Caucasian race and therefore the reference genome largely represents the reference genome of the European population.

Since 2000 many efforts have been made to best integrate the human genome revealing that there are differences between large geographic populations. In addition, there are entire regions that can cover hundreds of millions of bases. Within these regions even the arrangement of genes may vary. Concerning the typical features of the genome, the version 38.5 of 2015 [24] reveals that the total haploid human genome has 3,099,734,149 bases, which are distributed among the 24 human chromosomes of nuclear DNA. The largest of them, chromosome 1, has about 250 million bases (250Mbp), while the smallest, chromosome 21, has about 46 million (46Mbp). We should also consider the mitochondrial genome, a 16,571-base (16.5Kbp) circular DNA molecule, which is reported to be in duplicate within each cell. However, there are still regions of the human genome that have been read but have not been identified in chromosomes or in the genome.

2.2.2 Functional elements of the Human Genome

The human genome consists of hundreds of thousands of functional elements, such as: (i) genes coding for proteins or functional Ribonucleic acid (RNA) molecules, (ii) promoters, (iii) enhancers, and (iv) insulators among others. These elements act jointly in order to ensure the proper functioning of the genome. The last years, it has been understood that DNA should not be treated as a simple linear molecule. DNA folds and acquires a complex three-dimensional structure, resulting in adjacent genetic elements that may be even 1 million bases apart. About 50% of the human genome includes repeating elements [25], namely: (i) the small sequences that make up to 8% of the genome and (ii) the retrotransposons. It is worth mentioning the fact that the human genome (DNA) is capable of a very large number of epigenetic modifications which determine its structure and therefore reveal which areas can be functional and which are not [26].

2.2.2.1 Genes

Genes encode proteins and are the most well-studied part of the human genome. They constitute 1.5% of the human genome. Two main databases record the human genes, namely (i) the RefSeq [27, 28] which is the main reference system in clinical practice and (ii) the GENCODE [29] which contains more exons and more isoforms than RefSeq. Difference between these databases can be found in [30] which compares the gene annotation and the impact of reference genes on the prediction of variants' effects.

According to GENCODE and the statistics about the current release (01.2020 GRCh38), human genes are estimated to be around 60,662 (2020 version 33, Table 2.2). 19,957 of these encode proteins, about 17,952 large non-coding RNAs and approximately 7,576 small non-coding RNAs. There are also more than 14,768 pseudogenes in the genome. These numbers are indicative, as there are many genes that have been characterized as "predicted" by the genetic features of the genome but have not yet been confirmed experimentally. Regarding the names of genes encoding proteins, an important database is the HGNC (HUGO gene nomenclature committee [31], which is responsible for giving names to new genes as well as to make updates on the older gene names.

In addition to the genes encoding for proteins, there are approximately 26,000 genes encoding RNA molecules present in the human genome. They are discriminated into long noncoding RNA genes greater than 200 bp and small noncoding RNA genes smaller than 200 bp. However, these two categories are completely arbitrary concerning the inadequate knowledge we have about their function and how they affect cellular function.

2.2.2.2 Mutations

With respect to the reference genome there is on average one change per thousand bases in each human genome. It has been proved that these changes are not distributed constantly throughout the genome. On the contrary, they depend to a great extent on the function of the main genetic element they affect. It should be noted that among genes there are differences on the number of changes they carry on. These differences depend largely on the function of

Table 2.2 Statistics about the current GENCODE Release (version 33). The statistics contains only the annotation of the main chromosomes, as derived from the respective gtf file.

General Statistics			
Total No of Genes	60,662	Total No of Transcripts	22,7912
Protein-coding genes	19,957	Protein-coding transcripts	84,107
Long non-coding RNA genes	17,952	○ *full length protein-coding*	58,048
Small non-coding RNA genes	7,576	○ *partial length protein-coding*	26,059
Pseudogenes	14,768	Nonsense mediated decay transcripts	15,937
○ *processed pseudogenes*	10,672	Long non-coding RNA loci transcripts	48,438
○ *unprocessed pseudogenes*	3,554		
○ *unitary pseudogenes*	232		
○ *polymorphic pseudogenes*	55	Total No of distinct translations	62,357
○ *pseudogenes*	18	Genes that have more than one distinct translation	13,739
Immunoglobulin/T-cell receptor gene segments			
○ *protein coding segments* ○ *pseudogenes*	408 237		

protein(s) that each specific gene encodes [1]. In the recent years, the term "mutation" has been used to distinguish pathological from polymorphic changes in the genome. The discrimination of mutations can be difficult considering that some of them may be necessary but not enough for the occurrence of a disease or they may increase the risk for a complex disease in cases of common genetic changes in a population. In genetics, genome wide association studies (GWAS) are thereby conducted to detect any polymorphism that is associated with a trait in different individuals. However, any diagnosis and risk prediction must be carefully considered since most cases correspond to polymorphisms that are in linkage disequilibrium with the associated genetic mutation.

Before the advent of modern cell biology, researchers proposed that a small number of events (i.e. chromosomal abnormalities) is needed for carcinogenesis. Concerning the origin of malignant tumours [32], they can arise as a consequence of certain abnormal chromosomal mutations. Figure 2.5 depicts a timeline of these mutations related to cancer development along with the significant reports on somatic mutations during tumorigenesis [33].

2.2.3 Cancer genetics

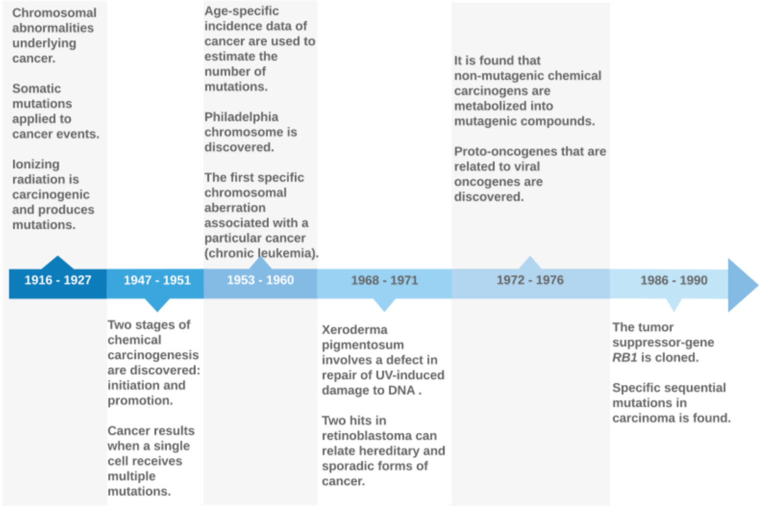
The completion of the HGP and the recent advancements in molecular biology and genetics, verified the main idea that neoplasia is caused by acquired genetic mutations [34]. Based on this knowledge, cancer genetics have been reached its maturity the last three decades. Through different approaches, such as the microscopic (chromosomes) and sub-microscopic (genes) we can thereby conclude that cancer is a genetic disease [35].

Cancer is a genetic disease in a sense that the primary "mutated" material is the genetic material of the cells. It has been well reported that any damage caused to the genetic material is responsible for the onset and progression of carcinogenesis. Nowadays, a huge amount of genetic information related to neoplasia has been collected while it continues to increase with high rate. This knowledge and information have contributed substantially to two major aims in cancer treatment: (i) the better understanding of the phenomena and mechanisms in oncogenesis and (ii) their direct application in clinical practice as indicators of diagnosis, prognosis, prevention, and cancer treatment.

The analysis level of the genetic material corresponds mainly to its observed defects. In the cellular level we can detect chromosomal abnormalities while in the molecular level gene anomalies. The term "genetic change" (i.e. rearrangement, damage or abnormality) refers to any minor or major change that cause the alteration of the normal function of a portion or portions of the genetic material. In addition, it has been shown that the chromosomal abnormalities result in gene rearrangements. Therefore, the basic concepts and definitions of genes involved in tumor development and progression are an essential part in the study of cancer genetics.

2.2.3.1 Genes related to cancer

From the huge number of genes that have been found in the cancer cells, more than 350 genes have been recorded whose genetic abnormalities are involved in carcinogenesis [36, 37]. Genes that are present in malignant cells have been classified into three categories, according to the genetic studies in oncology: (i) the oncogenes, (ii) the tumor suppressor genes and (iii) the stability genes.



Chromosomal abnormalities underlying cancer.
Somatic mutations applied to cancer events.
Ionizing radiation is carcinogenic and produces mutations.
Age-specific incidence data of cancer are used to estimate the number of mutations.
Philadelphia chromosome is discovered.
The first specific chromosomal aberration associated with a particular cancer (chronic leukemia).
It is found that non-mutagenic chemical carcinogens are metabolized into mutagenic compounds.
Proto-oncogenes that are related to viral oncogenes are discovered.
1916 - 1927
1947 - 1951
1953 - 1960
1968 - 1971
1972 - 1976
1986 - 1990
Two stages of chemical carcinogenesis are discovered: initiation and promotion.
Cancer results when a single cell receives multiple mutations.
Xeroderma pigmentosum involves a defect in repair of UV-induced damage to DNA .
Two hits in retinoblastoma can relate hereditary and sporadic forms of cancer.
The tumor suppressor-gene *RB1* is cloned.
Specific sequential mutations in carcinoma is found.

Figure 2.5 Mutations related to carcinogenesis across time [33].

2.2.3.2 Oncogenes

The alterations in key regulatory genes which control cell proliferation, differentiation and survival is the main cause of cancer. Towards the study of tumor viruses, it has been found that certain genes, namely the oncogenes, are capable of inducing cell mutation; thus, the first insights into the molecular basis of cancer were revealed [38]. However, more than 80% of human cancers arise due to other reasons (i.e. errors during DNA replication, radiation and chemical carcinogens) than viruses. The thorough study of viral oncogenes, for the overall understanding of cancer, revealed the cellular oncogenes which are involved in tumor development and are virus-related. This knowledge enhanced our better understanding of the molecular mechanisms of tumorigenesis. The key association between viral and cellular oncogenes was clarified during the study of retroviruses [39]. From the dozens of oncogenes that have been discovered until today, more information is given for those of the *Ras* family (*Harvey-Ras, Kirsten-Ras and N-Ras* [40]), for both their structure and function. Nowadays, the oncogenes are further studied as biomarkers for cancer prevention and as target genes for cancer therapy.

2.2.3.3 Tumor suppressor genes

The two types of genetic alterations that contribute to tumor development include (i) the activation of cellular oncogenes and (ii) the inactivation of tumor suppressor genes. The oncogenes stimulate the continuous abnormal cell proliferation in terms of either increase in gene expression or in the uncontrolled activity of respective oncogene-encoded proteins [41]. Tumor suppressor genes correspond to the opposite side of the mechanism controlling cell growth and under normal conditions they inhibit proliferation and tumor development. Tumor suppressor genes are lost or inactivated in cancer cells; hence, the negative regulators of cell proliferation are removed. This phenomenon results in the unusual malignant cell proliferation. Genetic changes in both oncogenes and tumor suppressor genes affect the normal function of cells similarly and comparably.

2.2.3.4 Stability genes

The third class of genes related to cancer refers to the stability genes that are responsible to retain the accumulation of genetic mutations caused by extracellular or intracellular factors in low rate. They are also known as guard genes and their alteration or suspension (i.e. loss of function) can increase the acquisition of changes in other genes, including the oncogenes or tumor suppressor genes among others.

2.2.3.5 From genes to pathways of cancer pathogenesis

The accumulation of data related to the type of genes involved in the process of carcinogenesis, led to their classification both functionally and clinically. This effort is still ongoing, and the data extracted have their origin on the extensive analyses of tumor genomes and the study of the interaction between the genes and their products involved in the process. To this end, a complex network of intracellular and extracellular signal transduction pathways is created. Concerning the vast amount of “diseased” genes that have been found in malignant cells, only a small proportion (~200-300 genes), called driver genes, are supposed to contribute definitely to the development and progression of tumors [36, 37]. These genes are involved in a certain number of signal transduction pathways within the cells. A basic classification of the pathways that are involved in carcinogenesis can be achieved based on their primary function. They can be further discriminated into three groups, namely the pathways that: (i) have major effect on cell differentiation, (ii) affect importantly cell proliferation and survival and (iii) control the integrity of the genetic material within the cells [35, 42].

The deregulation of the normal cell differentiation is one of the basic characteristics of cancer cells. Genetic mutations to the genes in the respective pathways result in changes to the direction of cell “de-differentiation” and to the ability for continuing cell division. The related pathways that belong to this category are: (i) the *APC* pathways (*APC, CDH1, CTNNB1* etc.) [43], (ii) the *WNT1* pathway [43], (iii) the *NOTCH* (*Notch 1, 2*) pathway [44] and pathways related to chromatin alterations [45, 46].

The genes that influence cell proliferation and survival are involved in pathways that have direct effect on cell cycle control (*TGF-β, MAPK*) and apoptosis (*p53, STAT*). These genes and their products are also involved in pathways that permit the survival ability under adverse circumstances (*RAS, PI3K*) [47, 48].

The pathways responsible for preserving the integrity of genetic material within the cells ensure the correct DNA replication and the regulated cell death when they have accumulated genetic abnormalities [49]. The deregulation or damage of genes that participate in these pathways result in the genetic predisposition of organisms to accumulate genetic mutations in their cells [50, 51]

Towards discovering and investigating the effect of gene pathways on carcinogenesis, we can conclude that important theoretical and practical implications exist. For instance, the way a targeted therapy may affect or restore the consequences of genetic damages could be predicted for genes that participate in the initial steps of the respective pathway. The fact that a small number of pathways are involved in different tumor types enhances the ability of revealing similar results in these tumors based on a targeted therapy. As mentioned above, these pathways cover each other forming thereby a larger network which participates in more than one cellular process. Obviously, within the cells several networks operate in terms of interactions among the various pathways.

In modern cancer genetics, the main conclusion is that the genetic changes of tumor cells are distributed unequally in the genome [42]. In different neoplasms, several genes, chromosomes, regions and other chromosomal zones are selectively involved in the genetic rearrangements. In addition, an increasing number of certain abnormalities have been found to be linked with specific diseases. These abnormalities and mutations in the genes are currently used as biomarkers for cancer diagnosis, prognosis, prevention and treatment in clinical practice.

Among the different diagnostic biomarkers, the chimeric genes (such as the BCR/ABL genes in chronic myeloid leukemia) that result from chromosomal translocations, inversions, and intrusions is the most indicative group. The major advantage of studying the chimeric genes among other biomarkers is the fact that their presence characterizes only the cancer cells where they can be identified. Additionally, molecular biomarkers have been associated with cancer prognosis in several cases of hematologic malignancies. These biomarkers could help to accurately assess the actual status of disease progression and the possibility of a metastases or disease recurrence. In the last few years, the identification of inherited genetic abnormalities or polymorphisms associated with cancer development has been the main issue with extensive implications in prevention healthcare programs.

The diverse role of genetics in cancer treatment has been elucidated through the clearer diagnosis and prognosis. Based on the rapid evolution of cancer genetics, a recent advancement in cancer therapy, i.e. the targeted therapy, has been unveiled. Targeted therapy refers to the treatment with specific molecules that prevent tumor development and progression. Its goal is to target precisely certain molecular functions to be more efficient and with less side effects than traditional chemotherapy. In general, targeted therapy makes use of monoclonal antibodies or small molecules in terms of adjustment.

2.2.3.6 Molecular approaches to cancer treatment

Current and future research on cancer will focus on better understanding of this complex disease by improving its prevention and treatment. The last decade, the elucidation of the molecular cancer biology led to the development of new targeted strategies and therapies. Dealing with cancer implies the prevention of its development through the improved genetic understanding of its origin and metastasis [52].

The most efficient way to deal with this malignant disease is to block its development. A second effective way is the early diagnosis of premalignant stages of tumor development that could be treated. Localized therapies, such as surgery or radiation, could be applied successfully for curing cancer if it has been diagnosed early before metastasis. The success rate of treating early carcinomas that remain localized to their initial positions is 90%. On the contrary, the survival rate decreases to 70% in patients that have been diagnosed with cancer that invaded adjacent tissues and lymph nodes. To this end, the early disease diagnosis is a major determinant of the disease progression and outcome [1].

Regarding the applications of molecular biology to cancer prevention and early diagnosis, they correspond mainly to the identification of individuals with inherited predisposition to tumor development. The inherited susceptibility may refer to mutations on tumor suppressor genes, on at least two oncogenes and to the inactivation of the stability genes. These abnormalities can be discovered with molecular techniques, allowing the detection of individuals at high risk for cancer evolvement.

Among the total cancer incidence, 5% correspond to the inheritance of mutations on well-established genes. The most common case of inherited cancer susceptibility is hereditary nonpolyposis colon cancer. This incidence corresponds to 15% of colon cancers. In breast cancer, mutations on the BRCA1 and BRCA2 tumor suppressor genes are very common and account for 5% among this type of cancer. Additional genes contribute to the increase susceptibility of cancer development and have been shown to be involved in common adult tumors. The detection of these susceptibility genes can thereby enhance the practical implications in the new approaches for molecular cancer treatment. The identification of individuals at high risk along with the early diagnosis and prevention could have a great impact on cancer mortality rates.

Most of the drugs used towards cancer treatment cause DNA damage or suspend DNA replication. Therefore, these drugs are toxic both for the cancer cells and the normal. An alternative strategy for curing cancer is the use of drugs that do not act immediately to malignant cells but inhibit tumor development in terms of angiogenesis [53]. A more promising personalized treatment corresponds to the creation of targeted drugs that act against the oncogenes that drive malignant cells and tumor development [54]. However, we should consider the fact that oncogenes do not function only to cancer cells; thus, they may affect negatively both benign and malignant cells.

Several cancer types can be treated if they are diagnosed early in premalignant stages. The genetic approaches for identifying high-risk individuals with inherited mutations allow the immediate diagnosis and efficient treatment of these cancer patients. The scientific effort for designing targeted drugs has already been started aiming at developing new therapeutic agents that act selectively against cancer cells.

2.3 Cancer descriptive epidemiology

Cancer epidemiology is dedicated to cancer prevention, prognosis and control and aims at increasing understanding about cancer in terms of studying the distribution and determinants of cancer incidence. Epidemiology of cancer can be used to detect events that increase or decrease the possibility of cancer development in certain populations and/or regions [55].

A variety of methodological approaches have been evaluated within cancer descriptive epidemiology. These approaches correspond to (i) the identification of novel risk factors, (ii) the evaluation of tumor heterogeneity among different cancers and (iii) the description of current and new trends of common and rare malignant tumors [56].

Research on cancer descriptive epidemiology concerns three main areas of interest regarding the distribution of the disease occurrence in populations related to time, tissue origin and individuals. The first research area encompasses the description and interpretation of disease patterns regarding the incidence and mortality worldwide. Additional exploration of differences due to age, sex, socio-economic status, area of residence and time is also considered. In the second study of cancer descriptive epidemiological approaches, the same philosophy is followed related to the patterns of survival rate of cancer patients along with the associated factors. The last research area contains the trends that are detected based on the comparison of morbidity, mortality and survival rates in Europe with those in other countries.

2.3.1 Cancer Statistics

According to the status report on the global cancer mortality and incidence provided by the International Agency for Research on Cancer (IARC) [5], it is expected that cancer burden will grow remarkably by 2040 to 27.5 million new cancer cases and 16.3 million cancer deaths based on the growth and aging of the population [6].

Figure 2.6 presents a global map with the estimated number of new cancer cases by world area, in 2018 [57]. Similarly, Figure 2.7 illustrates the estimated new cancer incidences and deaths worldwide for leading cancer sites, except non-melanoma skin cancers as patients have not been tracked by cancer registries. We can observe that the corresponding new cancer cases and deaths in 2018 were estimated to be 17.0 million and 9.5 million, respectively [57].

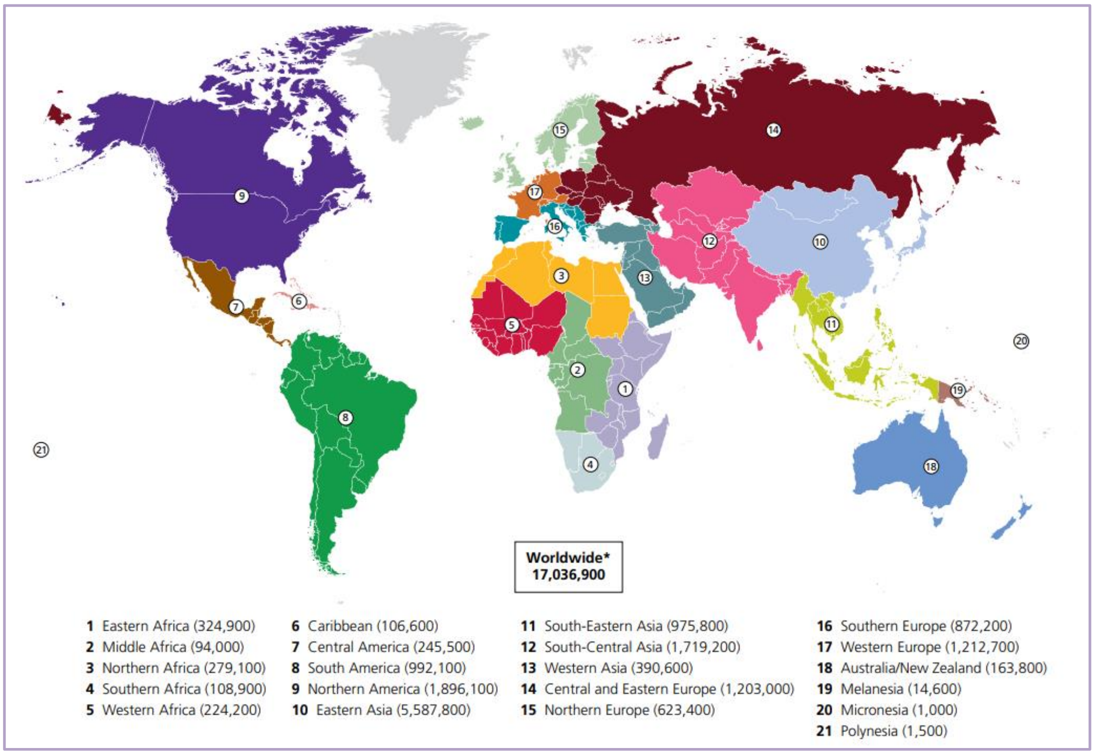
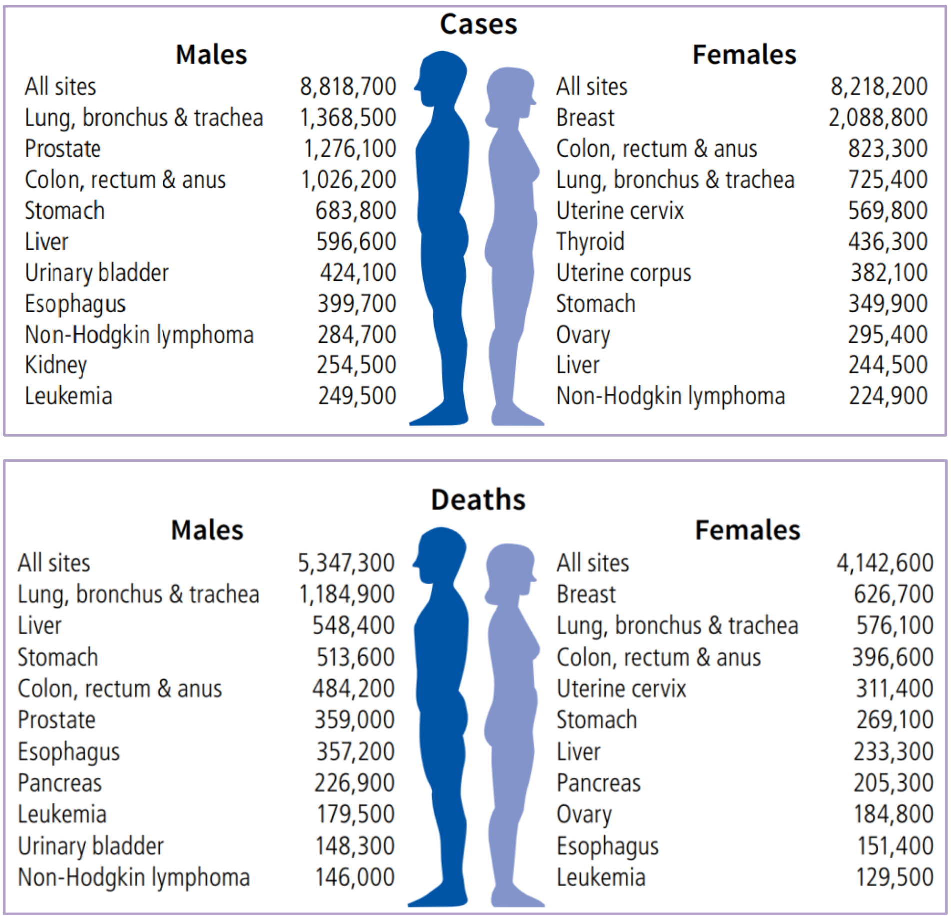

Worldwide*
17,036,900
1 Eastern Africa (324,900)
2 Middle Africa (94,000)
3 Northern Africa (279,100)
4 Southern Africa (108,900)
5 Western Africa (224,200)
6 Caribbean (106,600)
7 Central America (245,500)
8 South America (992,100)
9 Northern America (1,896,100)
10 Eastern Asia (5,587,800)
11 South-Eastern Asia (975,800)
12 South-Central Asia (1,719,200)
13 Western Asia (390,600)
14 Central and Eastern Europe (1,203,000)
15 Northern Europe (623,400)
16 Southern Europe (872,200)
17 Western Europe (1,212,700)
18 Australia/New Zealand (163,800)
19 Melanesia (14,600)
20 Micronesia (1,000)
21 Polynesia (1,500)

Figure 2.6 Estimated numbers of new cancer incidence worldwide. Region estimates do not sum to the worldwide estimate due to calculation method as noted by [57].

The cancer burden will be probably even larger in the future due to the increasing prevalence of risk factors contributing to tumor development. Smoking, obesity, UV radiation, physical inactivity, and fewer pregnancies in economically developed countries, are attributable factors to cancer burden. According to the Global Cancer Observatory (GCO) [8], the estimated number of incidences across all cancer sites at the age of 0-69 years, will increase 5.4% for males and 4.6% for females by 2020 comparing to 2018 estimates [7]. Respectively, for the age group of 70+ years, the cancer incidence will increase by 2020 with equal rates (i.e. 5.4% for males and 4.6% for females).

According to the GLOBOCAN 2018 estimates [6, 8] of incidence and mortality worldwide for 36 cancers in 185 countries, the most commonly diagnosed cancer is lung cancer (11.6% of the total cases) which is also the leading cause of mortality (18.4% of the total cancer deaths) for males. Among males, lung cancer is followed by prostate and colorectal cancer for incidence and by liver and stomach cancers for mortality. Breast cancer is the most common diagnosed cancer type and the leading cause of cancer death among women. For new cancer cases, breast cancer is followed by colorectal and lung cancer while for mortality it is followed by lung and colorectal cancers. Depending on the economic development of each country and the associated risk factors (i.e. lifestyle, social life and physical activity) the incident and cancer death cases vary across regions. The estimated cumulative risk of mortality in 2018, across all cancer sites for both sexes in the age group 0-69 is illustrated in the global map Figure 2.7 (top), as extracted from the International Agency for Research on Cancer 2020. In a similar manner, Figure 2.7 (bottom) depicts the estimated cumulative risk of incidence in 2018, across all cancer sites for both sexes. We can also observe in Figure 2.8 (top) that the greatest number of deaths will be in Eastern Asia, followed by South-Central Asia, Northern American and Western Europe. The greatest number of new cancer cases in 2018 will be in Eastern Asia, followed by Northern America and South-Central Asia. These numbers concern the size of the population, as well as cancer incidence and survival across countries. In addition, in Table 2.3 the estimated numbers of new cancer incidents and deaths by world areas and based on the GLOBOCAN 2018 estimates are listed [6, 8]. As mentioned previously, Eastern Asia has the highest number of cases and deaths for both males and females. In Eastern Europe, the cancer cases are higher in females than in males, while concerning the cancer mortality rate, males are most commonly die from cancer than females in countries of this region. Variations across countries with reference to cancer incidence and mortality have been associated with the Human Development Index (HDI) [58]. HDI is a measure of development that concerns not only the standard living but also the education and health status [58]. It is known that characteristics such as, age, risk factors prevalence, use of preventive strategies, early detection tests and high-quality treatment contribute to these differences. However, the level of development could also significantly influence these factors and cancer variations among regions. We should mention that countries with low HDI are characterized by higher cancer incidence attributable to causing infections, such as Helicobacter pylori (H. pylori). Across the world regions, 15% of cancer cases are attributed to infections, while in countries with medium and low HDI this percentage is higher (25%) [59]. It is noteworthy, that the most common cancer, across the world, related to infection is stomach cancer which is followed by liver and cervical cancers [6, 7]. Moreover, according to the most common types of cancer in each geographic region for males and females, separately, the most usual cancers in men other than lung and prostate include liver in several countries in Western Africa and South-



Cases

Males		Females	
All sites	8,818,700	All sites	8,218,200
Lung, bronchus & trachea	1,368,500	Breast	2,088,800
Prostate	1,276,100	Colon, rectum & anus	823,300
Colon, rectum & anus	1,026,200	Lung, bronchus & trachea	725,400
Stomach	683,800	Uterine cervix	569,800
Liver	596,600	Thyroid	436,300
Urinary bladder	424,100	Uterine corpus	382,100
Esophagus	399,700	Stomach	349,900
Non-Hodgkin lymphoma	284,700	Ovary	295,400
Kidney	254,500	Liver	244,500
Leukemia	249,500	Non-Hodgkin lymphoma	224,900

Deaths

Males		Females	
All sites	5,347,300	All sites	4,142,600
Lung, bronchus & trachea	1,184,900	Breast	626,700
Liver	548,400	Lung, bronchus & trachea	576,100
Stomach	513,600	Colon, rectum & anus	396,600
Colon, rectum & anus	484,200	Uterine cervix	311,400
Prostate	359,000	Stomach	269,100
Esophagus	357,200	Liver	233,300
Pancreas	226,900	Pancreas	205,300
Leukemia	179,500	Ovary	184,800
Urinary bladder	148,300	Esophagus	151,400
Non-Hodgkin lymphoma	146,000	Leukemia	129,500

Figure 2.7 The estimations of new cancer cases (top) and deaths (bottom) worldwide for leading cancer sites [8, 57].

Eastern Asia. For women, the most common disease type except breast is cervix and liver in Mongolia [57]. The financial costs of cancer management (i.e. treatment, care and rehabilitation) are increased substantially. These costs are direct and indirect influencing the economic status of both the patients and their families. Direct costs are related to treatment, care and resilience expenses, whereas indirect costs concern the morbidity costs due to lost productivity and the mortality costs (early death). Costs related to health insurance services and nonmedical charges are also included within the latter category. In Europe, the estimated annual direct costs for cancer in 2014 was 83€ billion [60]. The increased number of cancer cases and deaths, as well as of cancer therapies will result in higher rates of the global cost [61].

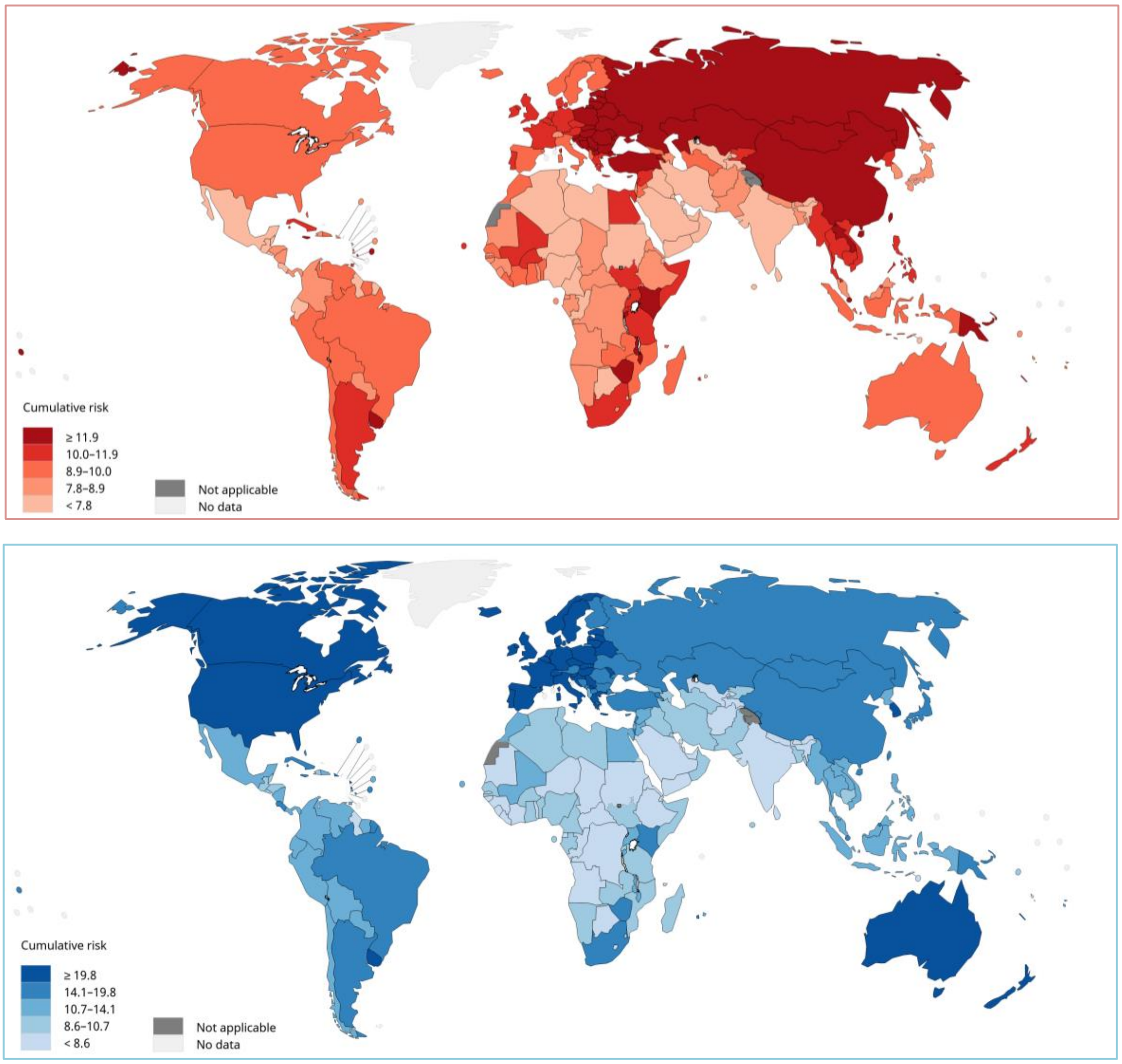


Figure 2.8 Global maps presenting the worldwide cumulative risk estimations of mortality (top) and incidence (bottom) in 2018 [6, 8].

Towards cancer prevention and control, a considerable number of cancer types could be prevented by reducing tobacco use and unhealthy lifestyle. Based on [8, 62], in 2015 around 20% of cancer deaths were caused by tobacco use, worldwide. Moreover, according to the World Cancer Research Fund estimates [63] about 15%-20% of malignant tumors are relevant to excess body weight, physical inactivity, and/or poor diet for health and growth.

Several approaches have been considered for controlling cancer including: (i) prevention, (ii) early detection, (iii) diagnosis and treatment and (iv) efficient therapy, such

Table 2.3 The estimated number of new cancer cases and deaths by world area according to GLOBOCAN 2018. The numbers for both males and females are given for each region [6, 8].

	Cases			**Deaths**		
	Male	**Female**	**Overall**	**Male**	**Female**	**Overall**
Eastern Africa	126,400	198,400	324,900	92,900	134,400	227,300
Middle Africa	40,500	53,500	94,000	30,200	37,600	67,800
Northern Africa	132,300	146,800	279,100	95,600	81,000	176,600
Southern Africa	47,400	61,500	108,900	29,700	31,300	61,000
Western Africa	87,200	136,900	224,200	62,100	88,300	150,400
Caribbean	54,900	51,700	106,600	34,000	28,500	62,400
Central America	109,900	135,600	245,500	56,800	60,800	117,600
South America	480,600	511,400	992,100	250,900	234,800	485,600
Northern America	970,100	926,000	1,896,100	363,900	329,100	693,000
Eastern Asia	3,090,600	2,497,300	5,587,800	2,129,600	1,315,100	3,444,700
South-Eastern Asia	470,900	504,900	975,800	342,400	283,400	625,800
South-Central Asia	848,200	871,000	1,719,200	614,600	545,000	1,159,600
Western Asia	204,400	186,200	390,600	128,900	90,800	219,700
Eastern Europe	595,200	607,800	1,203,000	381,700	310,800	692,500
Northern Europe	326,600	296,800	623,400	145,400	126,800	272,200
Southern Europe	479,200	393,000	872,200	244,900	174,300	419,300
Western Europe	658,700	554,100	1,212,700	305,900	239,800	545,800
Australia/New Zealand	87,700	76,100	163,800	32,700	25,600	58,300
Melanesia	6,400	8,200	14,600	4,200	4,700	8,900
Micronesia	500	500	1,000	400	300	700
Polynesia	800	700	1,500	500	400	900

as palliative care [64]. The World Health Organization (WHO) has emphasized on these approaches and suggest that countries across the world should create national strategies based on their economic status for controlling and preventive cancer. Until today, several national policies for cancer control have been founded and raise awareness about the risk factors and their minimization, the adoption of healthy lifestyles and the early disease detection [65].

Prevention concerns with the reduction or elimination of exposure to cancer causes, such as UV radiation, tobacco use, unhealthy lifestyle and diet and other environmental risk factors. In order to achieve the long-term control of the disease and the public health potential primary prevention should be adopted. Early detection would allow clinicians and patients to diagnose tumors at early stages allowing the timely diagnostic follow-up and the more effective medication. Organized and opportunistic screenings by physicians to individuals are strategies for early proven detection tests. Cancer diagnosis and treatment is among the most important steps in the disease management. Once the clinical and pathological assessments are carefully assigned (early diagnosis) then the most appropriate options and therapeutic protocols could be determined and prescribed. Among the initial agents for cancer treatment, surgery, chemotherapy, radiotherapy, hormone therapy, immune therapy, and targeted therapy are the most suitable. However, even if the developing countries assisted by WHO could integrate radiotherapy into sustainable cancer programs, there are low HDI countries that could not afford radiotherapy facilities and enough cancer centers, such as central Africa. Surgery, chemotherapy, and radiotherapy are among the most valuable modalities of palliative care [66]. Palliative care corresponds to the effective pain management and monitoring. Patients diagnosed with advanced cancer stage, especially in low income countries, are in the need of inexpensive pain relief medications, ranging from aspirin to opiates [67].

According to the GCO and the summary statistics for cancer mortality and incidence in Greece on 2018, the number of new cancer cases and deaths were estimated to 67,401and 33,288 for both sexes, respectively. The risk of developing cancer before the age of 75 years was equal to 33,1% and 22,8% for males and females, respectively. Similarly, the risk of dying from cancer before the age of 75 years was 15,6% and 7,8% for both males and females. Among the most frequent cancers excluding non-melanoma skin cancer for both sexes the top five ranked are: (i) lung, (ii) breast, (iii) colorectal, (iv) prostate and (v) bladder cancer [7]. Regarding the age-standardized incidence and mortality rates in Greece, for the top 10 cancers, breast and lung cancers have the highest incidence and mortality rates, respectively (Figure 2.9). Furthermore, Figure 2.10 depicts the respective numbers of new cancers cases in 2018 for all ages for both males (top) and females (bottom), respectively.

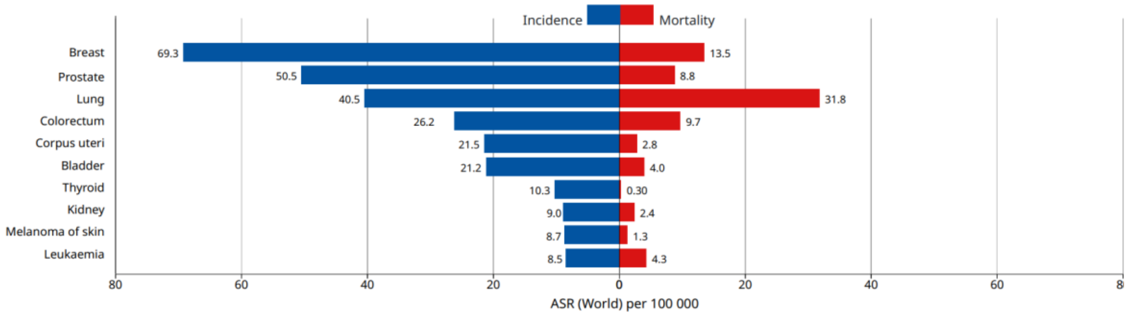


Figure 2.9 Estimated numbers of new cancer cases in Greece in 2018 for males (top) and females (bottom) [7].

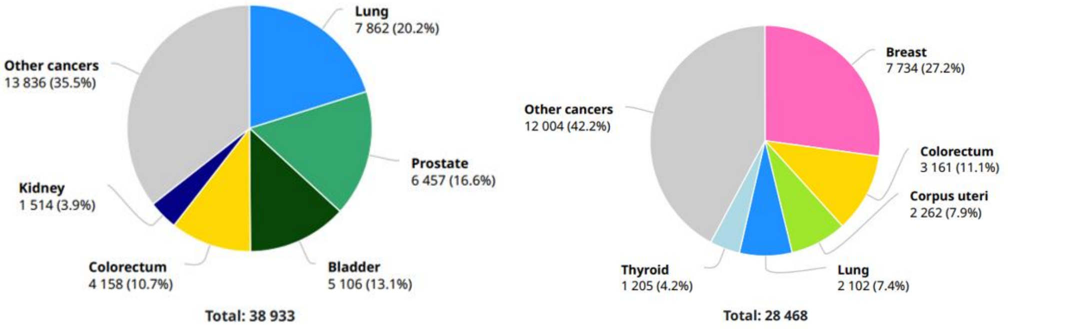


Figure 2.10 Bar chart of top ten cancer age-standardized (ASR) incidence and mortality rates per 100,000 individuals [7].

2.4 Biomedical data used for cancer diagnosis, prognosis and treatment

With the continuous improvement and the greater availability of new powerful and low-cost research technologies, the volume of biological and clinical data during biomedical studies has exploded. More specifically, technologies like next-generation sequencing [21, 68], high-throughput screening [69] and mass spectroscopy [70] have contributed to the ways biomedical data are collected and stored. In the era of big data, large biomedical data repositories, such as the Cancer Genome Atlas (TCGA) [71] and the 100,000 Genomes Project [72] have been created for improving prediction of cancer patients' prognosis and therapy. These public repositories consist of hundreds of matched histopathological imaging, genomic, and clinical data modalities providing the framework for a holistic and multi-modal integrative analysis of complex diseases, such as cancer.

The biomedical data availability and the insights they may provide into the biology of cancer, enhanced the possibility of making progress towards precision medicine and oncology. The molecular profiles of patients alongside the clinical findings allow for tailoring cancer prognosis, diagnosis and treatment.

Mining the large amounts of big data in order to decipher the molecular mechanisms of cancer and answering thereby complex biological questions, remains a big challenge. This open challenge will bring then precision medicine and oncology into the field of clinical care and management of diseases.

2.4.1 Big data for precision oncology

In the era of big data and with the advent of national and international electronic health records, the creation of comprehensive databases of multi-omics data as well as of initial patients' findings and treatment records have been enabled. Big data concerns the (i) high variety, (ii) high velocity, and (iii) large volume of information that can be acquired by different sources such as high-throughput technologies, electronic medical records, high-resolution imaging and omics approaches. The analysis and interpretation of these multi-modal data are progressing rapidly, remaining a challenge towards the evaluation of clinical endpoints.

The “high variety” characteristic reflects the different practices adopted to aggregate data into a single dataset for combined analysis. “High velocity” concerns the real-time compilation and analysis of generated data in terms of computational methods that lack human intervention. The definition of the “large volume” concerns the datasets that include orders of magnitude more observations and sample records than datasets previously reported and created [73].

In cancer research the acquisition, analysis and further exploitation of big data have been studied widely. Starting from DNA and RNA as well as from health information in digital format, researchers can elucidate the patient’s phenotype and genotype allowing the extraction of knowledge related to personalized selection of treatment. This knowledge can then be applied to prevent or decrease cancer incidence as well as improve the design of therapeutic protocols for reducing mortality rates.

The assembly of large consortia and the creation of patient cohorts have produced large amounts of biological and clinical data leading to the need for new solutions regarding storage, analysis and guidelines for sharing. The ultimate objective is to uncover hidden patterns and unknown collinearities within and between multiple sets of patient data gathered from heterogeneous sources. This knowledge can then be applied to enhance the health care of individuals separately, making personalized precision oncology a reality [74-78].

To accelerate progress toward a new era of precision oncology [79], researchers require access to curated datasets for empowering the identification of meaningful relationships among different samples. The establishment of infrastructure that could help clinicians and other relevant parties to store, analyze, integrate, access, and visualize large amounts of biomedical data is one of the main objectives of bioinformatics. Over the last years the access, integration and comprehensive analysis of big data within data “clouds” have become a basic requirement for optimizing the exploitation of heterogeneous data in precision oncology. Although difficult, the integration of raw data among different areas alongside their sharing, analysis and visualization will increase our understanding of the molecular basis of cancer at a personalized level.

Thinking of the complex nature of cancer and the multistep process of tumorigenesis, one can easily presume that not enough data can be obtained from single centers regarding cancer research. Moreover, the main goal of precision medicine in cancer is to tailor diagnosis, progression and therapeutic protocols of patients according to the overall status at both the phenotype and genotype levels. Therefore, the implementation of the FAIR (Findable, Accessible, Interoperable and Reusable) data principles in precision oncology have emerged and good practices in data sharing have been clearly defined [80]. “Findable” implies data that can be searched and found online in search engines for instance. The term “Accessible” concerns the data that can be retrieved and extracted directly or after a “request and approve” procedure. “Interoperable” means that data to be shared follow specific standards and finally, “Reusable” implies that the metadata produced by the analysis of raw datasets can be further exploited to produce the same results and/or be integrated with other datasets in different analysis pipelines [81].

Sharing of heterogeneous biomedical data among research groups is of utmost importance. In recent years, researchers express their increase willingness to share biomedical data and embrace data sharing practices. Towards this direction, reuse and reproducibility of research have been increased with reference to cancer prevention, disease follow-up, targeted medicine and treatment management [82, 83]

Based on this knowledge, specific challenges appear when data sharing in medicine and in precision oncology is considered. These challenges involve mainly the privacy and ethical issues as well as the way data are collected and stored [77, 84, 85].

The future of big data in precision medicine lies in the current and/or second next generation sequencing technologies. These technologies produce large amounts of short reads, while they are sequenced multiple times. The data extracted are characterized by technical and analytical errors and a large proportion cannot be aligned to known regions within the genome. To address these problems, the third-generation sequencing technologies will be introduced. DNA and RNA molecules are sequenced and reads up to 100,000 base pairs are produced providing deeper insights into the human genome and its reconstruction [86]. Regarding precision oncology, this is of great importance since the reference genome would be more complete and more accurate as the read alignments and calling variants will become more precise [74]. The sequencing of other biological components such as the proteome, epigenome and non-coding RNA would enhance the correlation between the patients’ genotypes and phenotypes leading to the identification of altered pathways and the suggestion of new targets [87-90].

2.4.2 High-throughput technologies for cancer research

The expression profiling studies using cDNA microarrays [91] and the recent development of DNA and RNA sequencing technologies [92], have revolutionized biomedical research and clinical practice [93]. From Sanger sequencing almost 40 years ago [94] to the introduction of next-generation sequencing (NGS) techniques in 2005 [21] we can thereby answer complex biological questions by aggregating high-throughput data.

During the last decade, high-throughput sequencing technologies have enabled the accumulation of large and different types of biological data (omics data) alongside the respective medical data in digital format (i.e. electronic health records - EHRs) for monitoring cancer onset and progression based on genomic changes at the molecular level. Exploring health and disease at the omics scale would enhance the processing, analysis and sharing of data that would empower targeted therapies and hence the era of personalized precision medicine. With the rise of omics data, integrative approaches have been proposed towards their analysis and the elucidation of the underlying molecular pathways and processes within cells [95, 96].

2.4.2.1 DNA microarrays and gene expression

Array technologies can quantify interactions among a set of molecules (i.e. DNA fragments and proteins) based on molecular probes that have been previously defined. The most commonly used of these technologies is DNA microarrays which enable the simultaneous measurement of the messenger RNA (mRNA) levels of certain genes within cells [91].

DNA microarrays is the most well-known method for detecting and quantitating gene expression levels. Its relative low cost, ease of use and the rapid technological progression constitute the main reasons why DNA microarrays have been used extensively. Most experiments conducted towards the study of expression levels of genes in parallel have been conducted through cDNA (complementary DNA) arrays [97-99].

The general principle of DNA microarrays refers to the physical property of hybridization of DNA molecules and the scientific progress in nanotechnology which enables the immobilization of many molecules on coated glass microscope slides with extremely high precision. A DNA microarray experiment contains therefore single-stranded DNA fragments on a solid surface. Through hybridization, these segments (probes) will find subsequently their complementary single-stranded molecules (cDNA) passing over the chip.

More specifically, genes of interest are obtained and following purification and quality control are printed on a surface (i.e. coated glass). mRNA from both the control and the test samples is labelled with fluorescent dyes (i.e. Cye3 and Cye5 dyes, respectively). Then, the targets which are also fluorescent are pooled in order to hybridize to the probes on the array. Through laser excitation of the targets, an emission is yielded measured using a laser microscope. Data derived from a DNA microarray experiment can be viewed as a normalized ratio (Cye3/Cye5) which corresponds to (i) no change on the levels of gene expression (no deviations from 1), (ii) increased expression levels (>1) or decreased (<1) expression levels relative to the reference RNA samples. It is noteworthy that the extracted information depends on the type of organism under study, the number of genes whose expression levels someone would like to measure, and the quantification ability of the microarrays used. In any case, these are high-throughput experiments which can be translated into a few thousand or tens of thousands of gene expression values. A reference example of the results of a DNA microarray experiment is shown in Table 2.4. The rows correspond to the quantified gene expression values for each gene (probe code number) identified (first column). The measurements of gene expression imply the fluorescence measurement for the given probe for each sample. We should mention that this type of data can be further exploited towards the identification of DEGs between the control and the test samples.

Generally, a DNA microarray experiment contains four specific steps: (i) mRNA isolation from the corresponding sample, (ii) generation of complementary DNA (cDNA) by reverse transcription, (iii) labeling of the cDNA with fluorescent dyes and (iv) hybridization of cDNA to the microarray and fluorescence measurement [91, 97].

2.4.2.2 Next Generation Sequencing

NGS technologies have been introduced the last decade due to the high demand for technologies that deliver fast, low-cost and accurate information with reference to the human genome. The main advantage of these technologies over conventional approaches, such as the array technologies, is the extraction of large amounts of sequencing data cheaply [21, 92].

In the realm of NGS techniques the primary order of millions or tens of millions of sequences is identified at a time and in parallel. Hence, gene expression microarrays are now being replaced by sequencing methods that can provide information about transcripts of a gene and/or of sequence variants [100, 101].

In the case of quantifying gene expression with NGS, the experiment is conducted based on the sequencing of sample mRNA which is converted to cDNA as in the microarray experiments. The main difference of RNA sequencing technologies concerns initially the fragmentation of the cDNA in order to be around 300-500 bases. Amplification via polymerase chain reaction (PCR) is performed and then sequencing through synthesis is applied. New cDNA clones are created based on the sample clones and synthesis is conducted nucleotide-nucleotide [101]. These steps result in a file that contains a very large

Table 2.4 An output file with the gene expression results of a DNA microarray experiment. The first column contains the probe code number that can be assigned to a gene or region within the genome. The following values in columns 2-6 correspond to the fluorescence measurement for the given probe for each of the fifteen different samples. These values are given without any preprocessing steps regarding their normalization.

ID_REF	GSM764749	GSM764750	GSM764751	GSM764752	GSM764753
1007_at	10,513	10,007	99,586	98,068	8,145
1053_at	75,275	83,168	48,133	79,584	79,279
117_at	70,429	74,713	59,696	5,016	56,357
121_at	25,882	26,992	25,738	34,532	28,001
1255_at	2,234	22,433	22,281	22,601	22,519
1294_at	39,417	52,873	58,897	40,192	4,762
1316_at	40,741	34,632	34,305	34,302	35,546
1320_at	26,283	26,366	2,597	26,686	26,532
1405_at	91,945	77,194	81,731	26,686	82,065
1431_at	33,865	35,316	2,702	32,732	24,578
1438_at	57,451	64,705	68,193	52,343	4,206
1487_at	7,392	74,031	72,347	70,048	69,781
1494_at	25,065	2,358	23,378	23,722	2,363
1552256_at	63,473	52,559	6,521	60,042	75,942
1552257_at	76,159	88,312	93,935	90,271	87,538

amount of short sequence reads which can be analysed subsequently based on biomedical and bioinformatics approaches. The quantification of the results can be achieved according to three main steps: (i) the quality control of the sequences and removal of those that do not meet specific reliability criteria, (ii) the mapping of the sequence reads to the reference genome to detect accurately how many times did each base of the genome under study have been read and (iii) the quantification of number of reads per transcript.

Considering the main applications of NGS technologies one can argue that the resequencing of the human genomes could unveil new genetic abnormalities that affect individual's health in terms of complex and multistep diseases, such as cancer. Moreover, NGS strategies can reveal genetic variants as they allow the sequencing of targeted regions and the whole genome of an individual. Hence, rare and common genetic variants within coding regions can be screened and determined.

2.4.3 Omics data for cancer research

Recent advancements in high-throughput technologies enabled the acquisition of large biological datasets (i.e. different types of omics data). Towards this direction, cancer research is progressing drastically with reference to the disease monitoring and management. Furthermore, the advent of these techniques permitted the characterization and quantification of the main classes of the biological molecules, namely (i) DNA, (ii) RNA and (iii) proteins. Hence, omics data could be generated independently at various genome levels (i.e. epigenome, transcriptome, proteome, metabolome and microbiome layers) [95] allowing thereby their global integrative analysis in cancer research as regards to multi-omics approaches.

Based on this knowledge, public comprehensive repositories have been developed containing relevant information for (i) gene expression measurements (Gene Expression Omnibus database [102]), (ii) phenotypes and genotypes (dbGaP database [103]), (iii) proteomics (ProteomeXchange [104]), (iv) metabolomics (MetabolomeXchange [105]) as well as (v) genome wide association studies of Single Nucleotide Polymorphism (SNP) - trait associations (GWAS catalogue [106]).

Among the several omics approaches that have been studied so far, in the current thesis we place special interest in (i) the genome and genomics, (ii) transcriptome and transcriptomics, (iii) proteome and proteinomics, (iv) epigenome and epigenomics, (v) metabolome and metabolomics and (vi) microbiome and microbiomics, concerning the recent studies in cancer diagnosis, prediction and treatment [95, 107]. Figure 2.11 depicts the multiple omics data types as proposed in [95] among the respective biological layers for disease research. Each data layer reflects both the genetic basis and environment within cells, whereas interactions between the different layers can be observed (black arrows between layers). Genome and phenotype levels are presented implying that from the genome level the starting point can be the genome features and genetic variants, while from the phenotype level we can start an experimental study at any layer of interest.

Genomics concerns the study of an organism's whole genome. The human genome consists of 3 billion DNA base pairs. These base pairs encode approximately 20,000 genes for the cells' functioning. This information reflects the coding regions which is approximately 1-2% of the whole genome, whereas the remaining percentage corresponds to non-coding regions (i.e. structural and functional annotations) [107, 108]. Because cancer is a genetic disease, the elucidation of patients' genetic background is of great importance

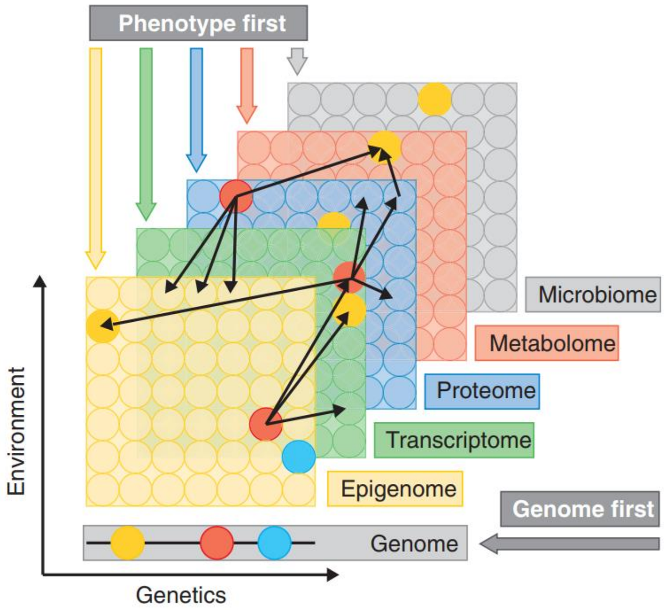


Figure 2.11 The levels of omics data types. Omics data are represented as circles collected under an experimental study of certain molecules. Each data layer concerns both the environment and the genetic status, except the genome level. Interactions among the different data types are depicted as black thin arrows. From the genome level one can start by studying the genomic alterations and genetic variants, whereas from the phenotypic level the starting point could be each of the individual data layers [95].

for identifying the genetics changes between different phenotypes. Many genetic variants exist within the genome. Most of them are harmful and increase the risk for a disease while others are protective and benign. Variants are discriminated into (i) the single nucleotide variations (SNVs), such as small insertions and/or deletions, and (ii) the structural variations (SV), such as copy number variants and inversions [109]. The most common SNVs are the single nucleotide polymorphisms (SNPs). Variants that are found in coding regions may affect protein sequence and their function, while SNVs or SVs found in non-coding regions may impact the expression of genes and other biological processes. Sanger sequencing, DNA microarrays and NGS methods are among the technologies that are used to capture genetic variants.

The transcriptome is the total amount of RNA transcripts within a cell or a population of cells. It consists of coding and non-coding regions, such as mRNA, transfer RNA (tRNA), micro RNA (miRNA) etc. The most common technique for gene expression profiling is DNA microarrays and recently the NGS methods. Changes in the transcriptome can impact health and disease; thus, analyzing mRNA transcripts could reveal the absence or presence of transcripts and assess also the impact of genotype on gene expression using alleles information. GEO [102], ArrayExpress [110] and Expression Atlas in EBI [111] are among the main repositories that store, analyze and share gene expression profiling studies.

Proteome refers to the entire set of proteins in each cell or biological sample. The thorough study of proteomics, protein-protein interactions and structural proteomics is known as proteinomics. Although sequencing technologies have emerged and widely used, the sequencing of proteins cannot be currently performed considering that incompleteness and inaccuracy of the respective sequence databases. Mass spectrometry (MS) [112] is mainly used in proteomics for the quantification and qualification of the proteome. Within proteomics, protein structure analysis and protein-protein interactions are studied in terms of identifying diagnostic molecules and common functions among proteins that co-localize or interact [112, 113].

Among genomics, transcriptomics and proteomics other fields at the omics scale have started to gain attention for their contribution to the understanding of disease conditions. Although these new areas have not yet achieved the level of complexity, depth and resolution like the main omics scales, their improvements could provide better explanations for the disease causes, its prevention and the design of effective and targeted treatments. We herein describe in more details the new omics areas which include the metabolomics, epigenomics and microbiomics. It has been found that the newly introduced approaches could enhance the potential of cancer diagnosis, prognosis and therapy. However, more effective inter-disciplinary efforts are needed for integrating multi-omics data towards (i) deciphering the underlying molecular basis of the disease, (ii) predicting patient outcomes and (iii) supporting treatment decisions [114].

Metabolomics implies the comprehensive catalogue of metabolites in an organism's cell [115]. Metabolites are produced during biochemical reactions and their alterations influence the genetic background of an individual. They are involved in metabolic reactions while they are essential for the proper function of cells and their growth. Hence, the study of metabolomics in cancer research could potentially improve the diagnosis and discovery of the underlying molecular pathways that characterize certain phenotypes.

The molecules that impact DNA metabolism after their binding are studied in the area of epigenomics. Epigenomics deals with the role of epigenetic modifications of DNA or DNA-associated proteins, while their importance in biological processes and disease development is evident based on epigenome-wide association studies [116].

The field of microbiomics is a fast-growing research area in which all the microorganisms of a given community are investigated together. The human microbiome is very complex considering the microorganisms that are colonized in the human skin, mucosal surfaces, and the gut. Profiling the human microbiome will unveil variations in its constituents (i.e. microbiota) allowing for finding correlations of microbial species with disease or other phenotypes [117, 118].

2.4.4 Imaging data for cancer research

Together with the multi-omics data and the health content obtained by clinicians in digital formats (i.e. EHRs), imaging data can be also considered as a valuable resource of information in cancer research. Biological image datasets contain quantitative measurements of cell, tissue and organism's processes and structures. They are of large volumes and provide illustrations showing tissues at the subcellular level. Metadata can be also acquired related to the imaging protocols used, the biological system under study and the output quantitative results of the image data. Obviously, the integration of cancer image data along with the respective medical and omics datasets within added-value platforms could support computational preprocessing and reanalysis for better disease outcomes [119-121].

In clinical oncology, imaging data have a critical impact on the evaluation of treatments, the extraction of imaging biomarkers, the design and planning of new therapeutic protocols as well as for diagnosis and staging of cancer. In the field of precision oncology, a new direction focusing on (i) the extraction of imaging features in terms of high-throughput technologies and (ii) the relationships among imaging phenotypes and genomics data have emerged in recent years for improved patient outcomes [122]. Radiomics and radiogenomics have gained increasing attention as they can facilitate cancer research in terms of precision diagnosis, assessment of prognosis and risk prediction and design of targeted therapies [123]. Radiomics concerns the detection of imaging biomarkers and important image features through high-throughput technologies. Several steps are required in order to extract significant imaging signatures including image acquisition, tumor segmentation, feature detection and predictive modeling [122]. Radiogenomics allows for the integration of imaging phenotypes with the related omics profiles towards precision medicine. The main idea is to identify correlates among genotypes and phenotypes of specific tumors; thereby, elucidating the biological and molecular processes underlying tumor development. The identification of novel imaging biomarkers and significant correlates of tumors molecular profile with the relevant phenotypes, could therefore improve tumor classification to major subtypes as well as patient stratification [123, 124].

2.4.5 Sensor data for cancer research

In the era of personalized oncology and medicine, the acquisition of patient health data from mobile apps and wearable devices is also considered of utmost importance for disease monitoring and for curing complex diseases like cancer [125]. Mobile health (mHealth) apps and wearable health devices have become increasingly popular as they provide accurate information about the patient’s treatment history, daily lifestyle and behavior. Collecting sensor data for diseases like cancer, enables their integration with other potential data sources facilitating thereby patient monitoring and accurate assessment of health status [75, 79]. Due to the multistep process of tumor development and considering the biology of cancer disease, the accumulation of heterogeneous data from different sources could potentially improve the decision making in disease management.

Based on this knowledge, the present thesis considers cancer disease and its progression for the elucidation of the underlying molecular mechanisms in terms of computational methods for modeling purposes. Data from cancer genetics studies as well as from transcriptomic repositories were utilized with the development and analysis of novel algorithms in the field of DBNs and other well-established ML approaches. Considering that cancer is a genetic disease and a multi-stage process we aimed at extracting knowledge from the pathway level along with gene expression changes between different phenotypes for modeling the dynamic behavior of the molecular mechanisms with reference to cancer diagnosis and prognosis. Cancer burden is going to be increased the next few years in terms of incidence and mortality rates, worldwide. Concerning the epidemiological studies that have been conducted and the bioinformatics analysis results, a huge amount of data is now available for deciphering the complex nature of cancer related to the mutations and metastases. We herein employed high-throughput and biomedical data coming from genotype studies and from genomic analysis for modeling cancer progression and prediction. Classification and predictive modes were obtained based on multidimensional techniques within the field of biomedical engineering and bioinformatics. Therefore, we could conjecture about the improved disease management in clinical practice while the acceleration of decision making in precision oncology can be anticipated.

CHAPTER 3 DATA SCIENCE AND ARTIFICIAL INTELLIGENCE IN CANCER

3.1 Data science in cancer research

With the emergence of data science, new meaningful insights can be obtained from large biomedical and high-throughput databases towards helping cancer researchers for better disease management [126]. Harnessing raw data, i.e. big data, in public health and especially in cancer research would allow clinicians tailoring medical treatment to the individual profiles of each patient. Data science covers a range of computational approaches and methods for the extraction of actionable knowledge from large, complex, multidimensional, and diverse data sources.

In the era of big data and precision oncology, technical challenges like (i) sharing, (ii) accessing and (iii) analyzing large biological and clinical datasets, as well as extracting meaningful outcomes from these data sources, could be overcome in terms of data science. On this basis, leveraging big data to their fullest extent could improve cancer care and decision making of patients' monitoring. A representative example is the Big Data for Breast Cancer initiative (BD4BC) [127] that highlights the need of data science and big data for advancing research, discovering new therapies and improving breast cancer patient care and outcomes; thus, optimizing precision medicine in cancer research. The initiative offers opportunities to researchers and data scientists to create a workforce with researchers that will understand breast cancer risks, onset, and progression and can apply data science methods to answer the challenges faced by breast cancer patients.

Accelerating research with reference to improved personalized treatments, will help cancer patients and the general public towards understanding the sharing of data and data science tools in order to deliver patient-oriented care and save lives. The promise of data science for advancing cancer research will decrease the percentage of population risk for cancer incidence and the rate of cancer deaths worldwide [126].

In general, the benefits that data science can bring to healthcare systems and especially to cancer research and management may be summarized to: (i) the better forecasting of population trends, (ii) the delivery of more preventive patient care, (iii) the selection of personalized therapies according to each patient profile and (iv) the cost-effective actions taken across a healthcare system and hospitals.

Data science can help to predict more accurately the trends on cancer patients and on the disease costs. The identification of high-risk patient groups for developing the disease and the selection of targeted therapies for prevention could be achieved. Additionally, data science could expand the research towards the identification of novel risk factors and biomarkers for cancer prevention. Hence, research is not only focused on the treatment of patients but to interventions that could prevent disease occurrence or reoccurrence. In precision oncology, tailoring therapies to the characteristics of each individual profile would increase life expectancy by considering the most effective and appropriate treatment pathways. Using the results of large population studies and integrating information from the different omics levels could allow the delivery of further personalized treatments.

In a data-enabled healthcare system, data science and the computational tools and methods it covers could optimize the productivity and costs across a hospital. Diagnostic tools and prescription options could be faster and cheaper providing more accurate results than the current practice. It has been shown that AI and ML has contributed to the automation of data-driven tasks within a healthcare system for extracting new clinical knowledge towards disease monitoring [130]. In addition, the use of data-enabled ML applications would improve the risk assessment, the plan of optimal treatment and the extraction of optimal research outcomes towards cancer prevention.

3.2 Artificial intelligence and machine learning for precision oncology

In the last decades, the potential of Artificial Intelligence (AI) to transform cancer research and thereby patient care and disease management have emerged. AI concerns the representation of problems and logic in a human-readable fashion [128] implying a very good idea that human thinking and machine computing might be "radically the same” [129]. As a field of computer science, AI aims at developing computational methods that can perform and execute analytical and predictive tasks.

In general, AI have expanded into three different areas in medicine for potential application, such as: (i) clinical practice, (ii) translational medical research and (iii) basic biomedical research [130]. In these areas indicative list of current and potential AI applications in medicine and healthcare systems has thereby been proposed [130]. In clinical practice, automated data collection, prediction of transcription factor binding sites and text mining are among the several applications of AI. In a similar way, in translational medicine the biomarker and drug discovery as well as the genetic variant identification have been elucidated by using AI. In clinical practice, disease diagnosis, patient monitoring, patient stratification and treatment selection are among the most indicative examples of AI usage.

The application of AI in clinical oncology corresponds to five main use cases concerning disease diagnosis and management [131]. In the first case of preclinical research an example of AI usage could be the automated analysis of multi omics datasets and the prediction of potential side effects of cancer. In the case of process optimization an indicative example of AI applications is the analysis of patient’s experience and resilience after disease diagnosis or follow-up treatment. In the third case of AI applications in clinical pathways, the analysis of clinical prognostication and of digital imaging could be indicative examples. Concerning the patient-facing applications for symptoms checking and disease monitoring, AI could be also exploited. Last but not least, AI applications at the population level correspond to the prediction of infectious disease outcomes and the identification of risk factors of complex diseases such as cancer, among others [132].

The potential applications of AI in medicine and precision oncology could augment patient stratification and offer new challenges and opportunities in clinical practice and decision making. Toward this direction, we herein discuss the main key challenges in AI for precision oncology which correspond to: (i) multi-view data integration, (ii) insufficient cancer datasets, and (iii) interpretable and explainable AI in cancer research.

With the rise of omics data and the large amount of clinical data generated from different sources, their integration in a learning process is a key step for the successful analysis through AI techniques [133]. Several methods have been proposed that incorporate different data views at the model's input level for obtaining better results related to disease outcomes and progression. Moreover, the combination of features that are extracted from heterogeneous data sources and of predictions made by different AI approaches have been studied thoroughly [134]. Towards enhancing the prediction of patient outcomes or specific phenotypes in precision oncology, the integration of multi-omics data with imaging features and respective clinical and medical data could enhance the predictive capabilities in AI while addressing the main challenge.

Dealing with different distributions of the training and test datasets in a classification problem is a great challenge that should be faced through AI approaches, such as transfer learning [135]. Despite the recent advances in high-throughput technologies, the availability of large, matched and fully annotated datasets is not always guaranteed. Hence, approaches like transfer learning have been studied and further applied in precision oncology for classification, regression and clustering where informative features detected in a specific domain can also be applied to a different but related domain. Transfer learning strategies has been adopted for prediction in learning problems to enhance the predictive accuracy of patients' risk stratification.

Achieving interpretable and explainable AI models in biomedical and healthcare research is one of the most challenging tasks. The evaluated performance of predictive models should reveal the causality that supports explanation and understanding of the problem related to patient-specific predictions. On this basis, explainable AI systems could provide details about the predictive features that are assigned to specific phenotypes or clinical outcomes.

In the era of AI, the problem of the black box models concerns the existence of complex models that are characterized by their insufficient interpretability (i.e. how they came to the prediction) [133]. Making black box models explainable, by providing explanation for their outputs, can be achieved at different levels when developing a predictive model, i.e. from the preprocessing steps to the solution of a classification problem and how predictions are made. The potential of interpretability for obtaining explanations and contributions of the variables for making prediction have been elucidated based on attribution methods [136, 137]. Moreover, improvements on models' interpretability relevant to precision oncology have been tested through "visible approaches" [138]. These new approaches concern the use of prior biological knowledge of the cells to be integrated within predictive models for the elucidation of the underlying biological processes. This knowledge will guide therefore new hypothesis to be made toward the progression on predictive performance of AI models.

Recently, promising efforts have been made toward interpreting predictive models in healthcare [139, 140] and especially in precision oncology [141]. However, these efforts are at their early stages and thus collaborative work would allow the understanding and explanation of how the results are obtained in terms of patient-oriented predictions alongside with their meanings. Generally, the three main steps for understanding AI decision making through modeling approaches are: (i) the explainability (i.e. understanding the reasoning behind any decision), (ii) the transparency (i.e. understanding of AI model decision making) and (iii) the provability (i.e. certainty behind decisions through mathematical and other computational approaches).

Below, we further describe the field of ML, a successful and well-known branch of AI, which have contributed to cancer diagnosis, prognosis and treatment by means of descriptive and predictive modeling. Several established algorithms are described along with the newly ones, such as DL techniques. Concerns about the interpretability and reusability of ML models and their results are also discussed explicitly.

3.2.1 Machine learning applications in cancer diagnosis, prognosis and treatment

Machine learning, a branch of Artificial Intelligence, relates the problem of learning from data to the general concept of inference [142, 143]. Every learning problem consists of two phases: (i) the estimation of unknown dependencies in a system from a given dataset and (ii) the use of these dependencies to predict new outputs of the system. The contribution of ML approaches in biomedical research and healthcare is apparent due to the existence of large amounts of biological and clinical data [144]. The applications of ML methods in biomedical engineering can be summarized to (i) sequence annotation, (ii) disease gene annotation, (iii) drug discovery and (iv) patient risk stratification.

According to the learning problems, there are two main categories that ML approaches belong to: (i) the supervised learning and (ii) the unsupervised learning categories. In supervised learning a labeled set of training data is used to estimate or map the input data to the desired output. On the other hand, under the unsupervised learning methods no labeled examples are provided and there is no notion of the output during the learning process. As a result, it is up to the learning scheme or algorithm to find patterns or identify groups according to the input data.

In supervised learning this procedure can be thought as a classification problem. The task of classification refers to a learning process that categorizes the objects to one of several predefined categories. The other two common ML tasks are regression and clustering. In the case of regression problems, a learning function maps the data (i.e. independent variables) into a real-valued variable (i.e. dependent variable). Subsequently, for new unseen data the value of a dependent variable can be estimated, based on regression analysis. Clustering is a common unsupervised task and implies the identification of categories or clusters (groups) that describe the data. Based on this process each new sample can be assigned to one of the identified clusters concerning the similar characteristics that they share. Suppose for example that we have collected medical records relevant to breast cancer and we try to predict if a tumor is malignant or benign based on its size. The ML question would concern the estimation of the probability that the tumor is malignant or not (1=Yes, 0=No).

Another type of ML method that have been also applied is semi-supervised learning, which is a combination of supervised and unsupervised learning. It combines labeled and unlabeled data in order to define and construct an accurate learning model. Usually, this type of learning is used when there are more unlabeled datasets than labeled.

When defining a learning problem, data samples constitute the basic components. Every sample is described by several attributes with different types of values. Furthermore, knowing *apriori* the specific type of the attributes allows the selection of appropriate algorithms that can be used in ML-based analysis. Data-related issues refer to the quality of the data and the preprocessing steps in order to prepare them for a ML problem. Data quality issues include the presence of (i) noise, (ii) outliers, (iii) missing or duplicate data and (iv) data that is biased, hence unrepresentative. When data quality is improved, typically the quality of the resulting analysis is also improved. A few different techniques and methods have been proposed, relevant to data preprocessing that focus on modifying the data for better fitting in a specific ML algorithm. Among these techniques the most important approaches include (i) the dimensionality reduction (ii) the feature selection and (iii) the feature extraction. There are many benefits regarding the dimensionality reduction when the datasets are described by many features. ML algorithms work better when the dimensionality is lower [142]. Moreover, the reduction of dimensionality can eliminate irrelevant features, reduce noise and can produce more robust learning models due to the involvement of less but informative features. In general, selecting a subset of features which are representative of the initial feature set based on dimensionality reduction techniques, can be characterized as a process in the feature selection approach. Three common approaches exist for feature selection namely embedded, filter and wrapper approaches [142]. In the case of feature extraction, a new set of features can be created from the initial set that captures all the significant information in each dataset. The creation of new feature sets allows the gathering of benefits from the dimensionality reduction techniques.

However, the application of feature selection techniques may result in specific fluctuations concerning the creation of predictive feature sets. In the literature, the phenomenon of (i) lack of agreement between the predictive gene lists discovered by different groups, (ii) the need of thousands of samples in order to achieve the desired outcomes, (iii) the lack of biological interpretation of predictive signatures and (iv) the dangers of information leak recorded in published studies has been thoroughly discussed [142, 143].

The main objective of ML techniques is to develop a model that can be further used to perform a classification, prediction, or estimation task. The most common task in a learning problem within biomedical research is classification. As mentioned previously, the learning function classifies the data item into one of several predefined classes. When a classification model is developed, by means of ML techniques, training and generalization errors can be produced. The former refers to misclassification errors on the training data while the latter on the expected errors on testing data. A good classification model should fit the training set well and accurately classify all the instances. If the test error rates of a predictive model begin to increase even though the training error rates decrease, then the phenomenon of model overfitting occurs. This situation is related to model complexity meaning that the training errors of a model can be reduced if the model complexity increases.

Obviously, the ideal complexity of a model not susceptible to overfitting is the one that produces the lowest generalization error. A formal method for analyzing the expected generalization error of a learning algorithm is the bias-variance decomposition. The bias component of a learning algorithm measures the error rate of that algorithm. Additionally, a second source of error over all possible training sets of given size and all possible test sets is called variance of the learning method. The overall expected error of a classification model is constituted of the sum of bias and variance, namely the bias-variance decomposition [142].

Once a classification model is obtained using one or more ML algorithms, it is important to estimate the classifier's performance. The performance evaluation of each predictive model is measured in terms of metrics such as sensitivity, specificity, accuracy and AUC. Sensitivity is defined as the proportion of true positives that are correctly observed by the classifier, whereas specificity is given by the proportion of true negatives that are correctly identified. The quantitative metrics of accuracy and AUC are used for assessing the overall performance of a classifier. Specifically, accuracy is a metric related to the total number of correct predictions. On the contrary, AUC is a measure of the model's performance which is based on the ROC curve that plots the tradeoffs between sensitivity and 1-specificity.

The predictive accuracy of the model is computed from the testing set which provides an estimation of the generalization errors. In order to obtain reliable results regarding the predicting performance of a model, training and testing samples should be sufficiently large and independent while the labels of the testing sets should be known. Among the most commonly used methods for evaluating the performance of a classifier by splitting the initial labeled data into subsets are: (i) the holdout method, (ii) the random sampling method, (iii) the k-fold cross-validation method and (iv) the bootstrap approach. In the holdout method, data samples are partitioned into two separate sets, namely the training and the test sets. A classification model is generated from the training set while its performance is estimated on the test set. Random sampling is a similar approach to the holdout method. In this case, in order to better estimate the accuracy, the holdout method is repeated several times, choosing the training and test instances randomly. In the third approach, namely k-fold cross-validation, each sample is used the same number of times for training and only once for testing. As a result, the original data set is covered successfully both in the training and in the test set. The accuracy results are calculated as the average of all different validation cycles. In the last approach, i.e. bootstrap, the samples are separated with replacement into training and test sets; hence, they are placed again into the entire data set after they have been chosen for training.

We herein, describe some of the well-known ML techniques that have been applied widely in the literature for the case study of cancer diagnosis, prognosis and treatment. Artificial Neural Networks (ANNs), Decision Trees (DTs), Support Vector Machines (SVMs) and Bayesian Networks (BNs) are presented along with their main characteristics [142]. We identify the trends regarding the types of ML methods that are used, the types of data that are integrated as well as the evaluation methods employed for assessing the overall performance of the methods used for cancer prediction or disease outcomes.

ANNs handle a variety of classification or pattern recognition problems. They are trained to generate an output as a combination between the input variables. Multiple hidden layers that represent the neural connections mathematically are typically used for this process. Even though ANNs serve as a gold standard method in several classification tasks [142] they suffer from certain drawbacks. Their generic layered structure proves to be time-consuming while it can lead to very poor performance. Additionally, this specific technique is characterized as a “black-box” technology. Trying to find out how it performs the classification process or why an ANN did not work is difficult to understand and interpret.

DTs follow a tree-structured classification scheme where the nodes represent the input variables and the leaves correspond to decision outcomes. DTs are one of the earliest and most prominent ML methods that have been widely applied for classification purposes [142]. Based on the architecture of the DTs, they are simple to interpret and “quick” to learn. When traversing the tree for the classification of a new sample we can conjecture about its class. The decisions resulted from their specific architecture allow for adequate reasoning which makes them an appealing technique.

SVMs is a well-known method of ML approaches applied in the field of cancer prediction/prognosis. Initially SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into classes. The marginal distance between the decision hyperplane and the instances that are closest to boundary is maximized. The resulting classifier achieves considerable generalizability and can therefore be used for the reliable classification of new samples. It is worth noting that probabilistic outputs can also be obtained from SVMs [142]. The identified hyperplane can be thought as a decision boundary between the clusters. Obviously, the existence of a decision boundary allows for the detection of any misclassification produced by the method.

BNs classifiers produce probability estimations rather than predictions. As their name implies, they are used to represent knowledge coupled with probabilistic dependencies among the variables of interest via a directed acyclic graph. BNs have been applied widely to several classification tasks as well as for knowledge representation and reasoning purposes.

The potential of AI in biomedicine and precision oncology has become apparent recently with advances in new ML technologies for computer-aided diagnosis [145]. These new technologies are capable of being integrated in clinical practice for improving patient outcomes and accelerating clinical decision making [146]. Since the early 2000s, DL approaches, a branch of ML, have advanced rapidly with applications in bioinformatics and biomedical engineering [147, 148]. Based on big data and the parallel and distributed ML frameworks for their analysis, DL architectures have emerged and are categorized into four groups: (i) the deep neural networks (DNNs) [149, 150], (ii) the convolutional neural networks (CNNs) [151, 152] and (iii) the recurrent neural networks (RNNs) [153, 154]. Basically, DL architectures correspond to artificial neural networks of multiple non-linear layers. Different types of DL architectures have been presented according to the type of the input data and the research objectives [155]. The main characteristic of DL is that the feature layers are learned from data using a general-purpose learning procedure and are not thereby constructed by the user. The DL applied research can be categorized into three main research topics, including the (i) omics, (ii) biomedical imaging and (iii)biomedical signal processing fields, among others.

In cancer, several DL architectures have been applied for the classification and / or detection of cancer types [156]. More specifically, a single deep CNN has been utilized for the classification of skin lesions using images and the disease labels (malignant and benign) as inputs to the algorithm [157]. The evaluation performance of the CNN algorithm showed that DL application on cancer prognosis outperforms other conventional ML techniques. DL frameworks have been also developed and further utilized for cancer diagnosis and classification based on gene expression profiles [158, 159]. Concerning cancer prognosis and treatment, DL methods have been also proposed to tackle the problem of predicting the drug response in certain cancer types [160].

Reinforcement learning (RL) [161], a distinctive class of ML, has also found applications in cancer research in terms of finding the optimal treatment policies and computer-aided disease diagnosis [162, 163]. In RL, an agent (i.e. the physician) learns from the interaction with his/her environment to achieve a goal based on the outcome that he/she wants to optimize (reward function). The learning process of an agent in a typical RL cycle is a continuous procedure. The interaction with the environment occurs at discrete time points. Once an environment's state is received the agent selects a certain action to interact with it. The environment responds then to the action and the reward that the agent will or will not receive is finally determined [161]. The corresponding applied DL and RL approaches to biological data include research avenues ranging from protein structure prediction to cancer prediction and risk stratification [147].

Furthermore, network-based ML approaches have attracted considerable attention in precision oncology and especially in network-based analysis of patient genomic profiles and drug repositioning [164]. The integration of genomic data with molecular networks in network analysis empowers the detection of network-based features and the prediction of cancer phenotypes. Moreover, the identification of cancer driver genes can be achieved when oncogenic alterations are detected based on the genomic profiles and the given molecular networks. Toward the network-based drug repositioning, the graph connectivity measures allow the prediction of any drug-target interaction. Network-based classification approaches extract all the topological features based on the drug-drug and target-target interaction networks. Subsequently, ML classifiers can be used to predict new targets in the test set of the drugs list. Recent and forthcoming application of DL, RL and network-based approaches to clinical oncology highlight their potential in health care and their significant impact in decision making with reference to cancer prediction and therapy.

The challenges on the reproducibility of ML models in healthcare and the concerns of how well new findings and results can be validated and reused by different research teams have also emerged the past few years [165]. A research study can be considered reproducible if given the employed dataset and the source code an independent researcher can obtain the same results with the initial study. Additionally, a study is replicable if the same conclusions can be reached by an independent group that studies the same clinical problem and perform the same experiments and analysis procedure after collecting new data.

Reproducibility does not imply that the results of a certain study are correct and validated. With the advancement of ML prediction models in clinical practice several obstacles and challenges should be addressed to consider these tools valid and safely deployed. The improved understanding of the underlying mechanisms of a complex disease such as cancer and its better clinical management, are the main outcomes of reproducibility and replication; thus, any limitation that hampers the use of reproducible ML research results should be addressed. The use of big data and ML algorithms in precision oncology gives the ability to analyze diverse data types and further integrate them into predictions for cancer diagnosis, prognosis and appropriate therapeutic protocols [166]. Examples of ML applications in oncology care delivery [74, 130, 133, 167] reveal the importance of using ML trained models toward accelerating progression in health care decisions and further help doctors to optimize complex clinical problems. As ML models improve patient outcomes and further influence the clinical decision making, they should firstly be reproduced and replicated before they are deployed in the clinical domain.

In the present thesis, we developed and applied novel analysis algorithms and multidimensional approaches in the field of ML and AI aiming at exploring the molecular basis of cancer by integrating high-throughput data, such as transcriptomics and genomics. To this end, we demonstrated the potential usefulness of ML methods and algorithms and anticipated that the combination of information from different levels could contribute to the elucidation of cancer prognosis and risk prediction.

CHAPTER 4 LITERATURE OVERVIEW

4.1 Introduction

In the era of personalized precision oncology, modeling cancer diagnosis, prognosis and treatment based on multidimensional computational approaches, such as machine learning and data mining, have become apparent. Cancer classification and prediction have been studied extensively in the literature by means of novel analysis and algorithms aiming at providing clues for the underlying complex molecular mechanisms of the disease.

In this chapter, we present a literature overview on modeling approaches which deal with cancer classification and prediction. Concerning the availability and the significance of high-throughput data related to cancer progression, we first present two well-known probabilistic methods, i.e. BNs and DBNs, that have been used widely in several studies for modeling gene expression data. Several research works are described that have contributed to cancer classification and prediction by developing probabilistic models that decipher the relationships within regulatory networks. Next, we present indicative studies which employ classification-based methods for cancer prediction. These studies utilize different types of biomedical data for developing and evaluating their models concerning cancer progression.

Finally, a comprehensive overview of ensemble-based techniques that have been developed the last few years for cancer risk prediction is given. Ensemble methods which are related to conventional machine learning techniques and provide integrative models at both the feature and the decision levels are described for modeling cancer progression. In the last section, we go one step beyond the-state-of-the-art and present the contribution of the current thesis towards the research that have been undertaken in modeling cancer diagnosis and prognosis based on computational approaches in the field of ML.

4.2 Modeling high-throughput data using probabilistic methods

Probabilistic approaches enable the representation and manipulation of uncertainty as regards to models and predictions, and they play a central role in machine learning and artificial intelligence [168]. Using probability theory to express any form of uncertainty we can further compute the distributions for representing all the unobserved quantities in a model and for finding their relation to the data [169]. Subsequently, the unobserved quantities given the observed data are inferred using the basic probability rules.

Furthermore, high-throughput technologies, such as DNA microarrays, allow the measurement of expression levels of many genes simultaneously, as they change over time. The amount of such experimental data can be further exploited in order to provide an overview of how genes interact with each other forming thereby a network and allowing the integrative analysis of biological systems [170, 171]. Gene networks can be defined as a graph over a set of nodes (genes or gene activities) and several edges that may represent different kinds of relationships. Modeling gene expression data in order to identify the structure of the underlying causal network that generates the observed data and further perform cancer classification in terms of tumor types is very appealing [172-174]. The combination of microarray data, and biological knowledge, such as protein-protein and protein-DNA interactions, toward estimating gene networks by using a Bayesian network model has been studied in the literature for determining the direction of gene regulation [173, 175]. Moreover, the inference of temporal transitions and changes between nodes in biological networks may unveil crucial aspects of the molecular processes in a complex disease, such as cancer [176, 177].

The prevalent example in machine learning over the last years for representing probabilistic models include the directed graphs, i.e. the BNs also known as belief networks, among others (i.e. undirected graphs). BNs is a well-studied approach for analyzing gene expression patterns. They represent the dependencies among several interacting quantities, such as the expression levels of genes. Biological processes that involve locally interacting components, for example components whose values directly depend on the value of other relative components, can be described through BNs. Several algorithms have been used towards their application in learning the structure of BNs from observations providing models of causal influence [176, 178-181]. Although BNs are characterized by probabilities and conditional independencies the notion of causal influence can be also defined [182]. These causal connections on gene expression data can be therefore deciphered depending on several assumptions of the nature of biological systems.

BNs refer to the general class of graphical models in which nodes and the edges between them denote the assumptions on their conditional dependence [183]. In a BN, causal reasoning (from known causes to unknown effects) and/or diagnostic reasoning (from known effects to unknown causes) can be deduced.

In addition, DBNs, an extension of BNs, enable (i) the modeling of stochastic phenomena such as the expression of genes within time into a cell, (ii) the incorporation of prior knowledge and (iii) the handling of hidden variables [184-186]. They have been utilized for modeling and classification of time series microarray data and for discovering how a random variable X evolves over time during a stochastic process [186]. The conditional probability distribution of future states of this variable has the Markov property which states that future events are independent of past events given the present.

The last decades, both BNs and DBNs approaches have been employed in the field of cancer research for modeling purposes; especially by applying the respective algorithms towards progression in cancer evolution and therapeutic strategies [187-190]. The proposed models can handle stochastic events in the context of probabilities accounting for noise and loss of independence which correspond only to strong interactions among the observed data (i.e. the expression values of certain genes). Below, we present the computational studies that have been published in the literature regarding the use of BNs and DBNs in the management of cancer disease. Gene expression data and other relevant types of biological data, such as epigenetics and mRNA expression profiles, regarding cancer progression were utilized aiming at developing network models based either on BNs or DBNs methodologies.

4.2.1 Bayesian network models for cancer classification and prediction

Several studies in the literature have proposed computational methods with reference to BN modeling aiming to infer gene network models from multiple sources of biological data, such as gene expression data. To this end, in [191] an integrative inference model is described for the reconstruction of gene regulatory networks in ovarian cancer. Differences in the network topologies have been identified which may reveal the regulatory mechanisms associated with different cancer subtypes. Additionally, in [180] a framework for learning the structure of gene networks from experimental data is proposed. Bayesian networks with the integration of external knowledge are used in order to extract gene interaction information for pairs of genes. In a similar manner, the inference of gene regulatory networks using BNs is studied in [192]. The authors showed that the integration of gene expression and epigenetic data can improve the identification of gene regulatory interactions in terms of accuracy.

Recent studies have proposed network analysis methods towards accelerating the application of BNs models in predicting cancer subtypes. In [187] the approach of network analysis through BNs was selected for the detection of subtle but coordinated changes in expression of interacting and functionally related genes. The gene expression profiles of Acute Myeloid Leukemia (AML) and Myelodysplastic Syndrome (MDS) were analyzed using topological analysis for improving the classification of these two malignancies. BNs were employed to model the interactions between thousands of genes. The proposed predictive model deciphers the association between gene modules (i.e. certain genes that are related to biological pathways) and the disease type.

In [189, 193] cancer progression and survival have been studied by BN models based on their structure and parameter learning. The first study contributes to the monitoring of myeloid leukemia progression through a causal BN model. Hence, based on the model’s predictions, the authors demonstrated that possible mechanisms related to the disease progression can be elucidated from the chronic phase to blast crisis of the disease. In the second study changes in mRNA expression profiles and the cell-cycle progression in breast cancer cells that follow inhibition of the MEK signaling network across time were assessed based on BN modeling. In [194], ensembles of Bayesian networks were developed to investigate and identify novel MEK-dependent regulatory molecules of the cel-cycle which may connect with the NFkB network.

In the study of [188] the basic motivation was to test the viability of developing BN models towards improving the clinical decision support with reference to survival prediction and therapy selection in lung cancer patients. The authors propose a decision support tool in terms of BN models for the accurate personalized estimation of patients' survival as well as of the selection of treatment recommendations. A large national lung cancer patient dataset was exploited, and the obtained results were promising for accelerating the clinical decision support in lung cancer care.

In a similar study [195], a substantially larger patient cohort has been employed for developing a BN for modeling the survival prediction for colon cancer patients. Relevant BN applications for survival prediction and local failures have been conducted aiming at predicting the life expectancy and the disease progression in cancer patients [196, 197]. On this basis, the use of BNs as a predictive tool in clinical practice and cancer care has been studied thoroughly in the literature with encouraging outcomes for disease prediction and treatment management. The prediction of patient survival outcomes has been achieved through certain BN approaches alongside the exploitation of many patient records including gene expression measurements.

4.2.2 Dynamic Bayesian network models for cancer classification and prediction

Building stochastic models empowers the automated identification of the structure of underlying causal networks that are generated from observed data. DBNs have been proposed years ago for modeling stochasticity, integrating prior knowledge, and handling hidden variables and missing data [184]. Apart from that, it is known that gene expression is an inherently stochastic phenomenon in living organisms; therefore, DBNs can be used for modeling time-series data. DBNs methods concern the development of directed graphical models that capture time which flows forward [179]. Within a DBN model the arcs can be directed or undirected, since probability correlations are studied. The term "dynamic" in a DBN model corresponds to the modeling of a dynamic system and it does not mean that the structure of the graph changes over time.

DBN models have been developed and implemented in the field of cancer care for modeling gene expression data and further predict the disease status. Based on this knowledge, personalized predictions for patients at low/high risk for tumorigenesis can be obtained in the field of precision medicine.

Relevant approaches have been proposed in the literature for the inference of gene regulatory networks by utilizing time series microarray data [185, 198, 199]. These methods integrate time course data and implement DBN algorithms for learning the structure of gene regulatory networks accurately. They further discuss the application of these methods for deriving gene regulatory networks with their transition nodes; thus, deciphering the biological networks under study. A slightly different algorithm has been also presented in the literature that scores regulatory interactions between genes using DBNs [186]. The authors tested their method in mRNA time series data from breast cancer cells and identified co-expressed genes accurately.

The exploitation of clinical as well as longitudinal data for learning the structure of a gene network allows for the identification of gene interactions that can be proven crucial for cancer prediction and outcomes. Furthermore, the integration of genomic data with network knowledge allows for the identification of biomarkers not only as individual genes but as functional hubs as well.

In [200] DBN models have been developed for lung cancer screening. More specifically, the models were developed based on longitudinal data (i.e. patients screening information at follow-up period) for identifying high-risk lung cancer patients at an early stage and improving their survival. Demographics, smoking status, cancer history, family lung cancer history, exposure risk factors, comorbidities related to lung cancer, and screening related information were employed for building the DBN predictive models. The results presented are comparable to clinical experts' decisions with DBN models outperforming other conventional statistical and ML techniques. The performance evaluation of the proposed methodology concerns the generalization ability of the methodology to model new unseen data with reference to lung cancer incidence prediction.

A multiscale and multiparametric approach based on DBNs has been studied in [190] for modeling the onset and progression of oral squamous cell carcinoma (OSCC) after remission. Gene expression microarray data from circulating blood cells throughout the follow-up period in consecutive time-slices were collected and analyzed in order to model the temporal dimension of the OSCC. To this end, a DBN method was implemented to capture the underlying mechanism dictating the disease evolvement. The proposed model was further employed for monitoring the status and prognosis of the patients after remission.

DBN models have been increasingly used with further applications in decision making related to cancer diagnosis, prediction and treatment. In [201] the assessment of increased cervical cancer risk has been performed in terms of a DBN-based approach able to handle the inherent temporal nature of screening observations over time and further minimize cancer risk through more frequent test.

4.3 Modeling biomedical data using classification techniques

The last two decades a variety of ML techniques and feature selection algorithms have been widely applied to cancer prognosis and prediction [202-205]. Most of these studies employ ML methods for modeling cancer progression and survival and identify thereby informative features that can be employed for classification purposes. Generally, these studies utilize high throughput data, clinical findings as well as histological parameters for developing predictive models in terms of ML algorithms. The successful disease prognosis depends on the quality of a clinical diagnosis; however, a prognostic prediction should consider more than a simple diagnostic decision. When dealing with cancer diagnosis and prognosis (i.e. prediction) one is concerned with three predictive tasks: (i) the prediction of cancer susceptibility (risk assessment), (ii) the prediction of cancer recurrence and (iii) the prediction of cancer survival. In the first two cases one is trying to find the likelihoods of tumorigenesis and of developing again a type of cancer after complete or partial remission. In the last case, the prediction of a survival outcome such as disease-specific or overall survival after cancer diagnosis or treatment is the main objective. The prediction of cancer outcome usually refers to the cases of (i) life expectancy. (ii) survivability, (iii) progression and (iv) treatment sensitivity.

Most of the studies in cancer literature make use of one or more ML algorithms and integrates data from heterogeneous sources for the detection of tumors as well as for the prediction of cancer development in terms of risk assessment. A growing trend in the last decade concerns the use of conventional supervised learning techniques, namely SVMs, DTs and BNs, as well as of DL algorithms towards cancer prediction and treatment. The respective ML algorithms have been extensively used in a wide range of applications in healthcare systems with regards to precision oncology [133]. With the advent of genomic, proteomic and imaging technologies, molecular information related to cancer onset and progression can be obtained. Molecular biomarkers, cellular parameters as well as genes that are differentially expressed among phenotypes have been proven informative indicators for cancer diagnosis, prognosis and treatment. High throughput technologies nowadays have produced huge amounts of cancer data that are collected and are available to the medical research community. However, the accurate prediction of a disease outcome is one of the most interesting and challenging tasks for physicians. As a result, ML methods have become a popular tool for medical researchers. These techniques discover and identify patterns and relationships between them, from complex datasets, while they can effectively predict future outcomes of a cancer type. Additionally, the implementation of feature selection approaches alongside the selected ML algorithm enables the detection of significant features that are informative and contribute the most to cancer patients' risk stratification.

In the literature, a significant number of relevant ML-based studies integrate data from heterogeneous sources in order to predict the desirable outcome. To this end, the studies that have employed certain ML techniques, such as SVMs, for cancer susceptibility and recurrence prediction are presented in Table 4.1 along with their relevant publications and findings. We further discuss some of the most recent studies related to cancer prognosis and survivability with their respective results obtained based on their methods (Table 4.2).

Table 4.1. Relevant publications with regards to the use of ML applications in cancer diagnosis and prediction. The cancer type, the total number of samples, the data types along with the methods and the obtained results are presented for each separate study.

Publication	**Cancer Type**	**Dataset**	**Type of Data**	**Methods**	**Results**	**Validation**
Waddell M. *et al.,* [212]	Multiple Myeloma	80 patients	SNPs	SVM	Acc. = 0.71	Leave-one-out cross validation
Listgarten J. *et al.,* [213]	Breast cancer	174 patients	SNPs	SVM	Acc. = 0.69	20-fold cross validation
Kim W. *et al.,* [210]	Breast cancer	679 patients	Clinical, Pathologic, Epidemiologic	SVM	Acc. = 0.89	Hold-Out
Tseng C.-J. *et al.,* [211]	Cervical Cancer	168 patients	Clinical, Pathologic	SVM	Acc. = 0.68	Hold-Out
Lu H. *et al.,* [214]	Breast cancer	82,707 records	Incidence and population	Genetic Algorithm	Acc. = 0.88	3-fold cross validation
Vasudevan P. *et al.,* [215]	Glioblastoma	215 patients	Genomic	ANN	Acc. = 0.89	10-fold cross validation
Sepehri S. *et al.,* [216]	Lung cancer	396 patients	Clinical, radiomics	LR	Acc. = 0.76	Random sampling

Table 4.1. *continued*

Publication	**Cancer Type**	**Dataset**	**Type of Data**	**Methods**	**Results**	**Validation Method**
Yu K. H. *et al.*, [206]	Lung cancer	9,879 features	Image features	SVM	Acc. = 0.85	10-fold cross validation
Lu T. P. *et al.*, [207]	Ovarian cancer	575 genes	Gene expression microarray	SVM	log-rank test, p = 0.015	Leave-one-out cross validation
Zhang S. *et al.*, [208]	Prostate cancer	43 genes	Genetic features	SVM	Acc. = 0.66	5-fold cross validation
Ahmad L. G. *et al.*, [217]	Breast cancer	1189 patients	Clinical variables	ANN	Acc. = 0.94	10-fold cross validation
Exarchos K. P. *et al.*, [218]	Oral cancer	41 patients	Clinical, imaging and genomic	BN, SVM	Acc. = 0.69	10-fold cross validation
Chang S.W *et al.*, [219]	Oral cancer	31 patients	Clinicopathologic and genomic markers	Adaptive neuro-fuzzy inference system	Acc. = 0.74	5-fold cross validation

Table 4.1 presents indicative studies published in the field of cancer diagnosis and prognosis prediction. The relevance of each publication was assessed based on the keywords of the predictive tasks found in their titles and abstracts. Specifically, we selected those publications that study two of the three foci of cancer prediction (i.e. cancer susceptibility prediction and cancer recurrence prediction). The third foci, i.e. cancer survival prediction, is discussed afterwards in Table 4.2 where the relevant studies are depicted.

The cancer type, the data sources employed, the methods and the results obtained are also depicted. Several types of cancer are studied for making accurate prognosis including breast cancer, prostate cancer, thyroid and ovarian cancer among others. Most of the studies utilized the well-known ML technique, namely SVMs which outperforms other well-established techniques. In the studies of [206-213] the prognostic performance of SVMs was evaluated on external datasets yielding promising results for predicting and differentiating cancer subtypes as well as phenotypes.

In [210] an SVM-based model for the prediction of breast cancer recurrence, called BCRSVM has been proposed. The authors support the idea that the classification of cancer patients into high-risk or low-risk groups allows experts to adjust a better treatment and follow-up planning. In this work, the development of a predictive model regarding the breast cancer recurrence within five years after surgery was achieved. SVM, ANN as well as Cox-proportional hazard regression were employed for producing the models and find the optimal one. The authors claimed that after comparing the three models based on their resulted accuracies, they found that the proposed model outperformed the other two. From the initial set of 193 available variables in their data, only 14 features were selected based on their clinical knowledge. These data refer to clinical, epidemiological and pathological variables of 733 patients considered out of 1.541. In the final stage of the feature selection process, Kaplan-Meier analysis and Cox regression were applied which resulted in 7 variables as most informative. These features were then entered as input to the SVM and ANN classifiers as well as to the Cox regression statistical model. In order to evaluate the performance of the models, the authors employed the hold-out method, which splits the data sample into two sub-sets, namely training and testing set. Likewise, in most studies, accuracy, sensitivity and specificity were calculated for the estimation of the models' performance in terms of generalizability and robustness. Based on their results, the authors claimed that the proposed model outperformed the ANN and Cox regression models with accuracy 84.6%, 81.4% and

72.6%, respectively. Comparison among the performance of other previously established recurrence prediction models revealed that BCRSVM has superior performance. According to their findings, the authors propose that for each of the three predictive models, the most significant factor regarding the prediction of breast cancer recurrence was the local invasion of tumor. Similarly, in [207] the authors utilized SVMs to develop a model for the chemo-response. The data exploited were from the Cancer Cell Line Encyclopedia (CCLE) while TCGA and the GSE9891 datasets were utilized for evaluating the predictive model. Moreover, based on their findings the resulted 10-gene predictive model demonstrated that a longer recurrence-free survival was observed in the high response group. These patients had also good response and favorable prognosis. The presented studies constitute indicative examples of the use of ML approaches to cancer prediction. Several other studies utilized well-established ML methods, such as ANNs, BNs and DTs for predicting cancer diagnosis and progression by integrating heterogeneous data sources to obtain accurate models [214-219].

Recently, several studies utilized techniques related to the DL approach aiming at exploiting many imaging data and further improve the decision support on cancer prediction. In [220-222] deep and convolutional neural network models were applied for diagnostic purposes in cancer care. The authors aim at improving the diagnostic accuracy of cancer by analyzing imaging as well as multi-dimensional data. The ability of DL methods to accelerate cancer diagnosis, prognosis and treatment has been elucidated in the literature with efficient algorithms that can capture feature representation in a general manner for classification purposes [167].

Specifically, in [220] multi-dimensional data were exploited, and a novel multimodal deep neural network approach was proposed for cancer prognosis prediction. The predictive model consists of an input layer, multiple hidden layers and an output layer, while heterogeneous data such as clinical, gene expression and copy number variations were considered. The proposed method achieved an overall better performance compared to conventional ML techniques like SVMs and RFs. The authors also presented the usefulness of exploiting different data sources along with multimodal deep neural networks for cancer prognosis prediction. In [221], a multicohort diagnostic study was performed using ultrasound images sets and deep convolutional neural networks. Based on the obtained results the predictive model achieved high performance towards identifying thyroid cancer patients. In [222], a deep convolutional neural network was trained on whole-slide images derived from TCGA to classify accurately cancer and normal tissues in non–small cell lung cancer. The performance of the proposed method achieved an average AUC equal to 0.97

Similarly, the research work of [223] demonstrates new strategies based on deep learning-based predictive models. In this study a multi-model ensemble method that exploits RNA-seq datasets of different cancer types is proposed based on DL. Feature selection techniques were applied while their classification method was evaluated in terms of cross-validation. Their results reveal that more accurate results can be obtained from their ensemble methodology rather than from single classifiers. Breast cancer risk prediction has been also studied in [224]. Deep learning models were designed and developed based on screening mammograms to assess breast cancer risk within 5 years. The final hybrid DL model incorporates both traditional risk factors and mammograms showing significant higher results compared to the single models that utilized only individual risk factors or mammograms. Other related works consider the cancer risk prediction in terms of mammography-based models [225, 226]. Conventional ML-based approaches and DL methods were applied for early cancer detection based on the utilization of mammography images and clinical findings.

Concerning cancer survival prediction, several works have been done and published the last decades towards assessing disease survivability through ML techniques. *Table 4.2* shows indicative studies that utilized well-established ML algorithms for predicting life expectancy. The types of cancer, the data used alongside the methods and the obtained results are also presented. The relevance of each publication was assessed based on the keywords of the predictive task found in their titles and abstracts. Specifically, we selected those publications that study cancer survival in terms of ML-based methods.

In [227], a deep learning-based method was developed for predicting survival across many cancers. Histopathology images were used from around 5,000 cases alongside their slides. Based on this cohort and on multivariable Cox regression analysis the proposed method was associated with disease specific survival (hazard ratio of 1.58, 95% CI 1.28-1.70, p<0.0001). Deep learning methods have been also utilized in [228] towards improving survival prediction in colorectal cancer. This multicenter study exploited hematoxylin–eosin (HE)–stained tissue slides from patients and further investigated whether there are prognostic features that can predict their survival. A convolutional neural network was trained and after the validation of the model on an external dataset a nine-class accuracy equal to 0.94 was obtained. Colorectal survival prediction was also studied in [229]. A two-stage model was developed based on ensemble ML techniques to predict survival time on a monthly basis; hence, improvement on the decisions regarding the treatment options could be achieved. A classification and a regression model were proposed by exploiting data from the surveillance, epidemiology, and end results program (SEER) [230]. Data preprocessing was performed, and the class imbalance problem was addressed based on under sampling techniques. According to the obtained classification results an overall better performance and generalizability were achieved when class imbalance handling was considered in the training phase. Breast cancer survivability prediction was studied in [205, 231] in terms of ML models which were trained and evaluated after integrating different data sources. Promising results were obtained, and the authors demonstrated the potential of ML principles in the progression of cancer prediction. Related studies that utilized conventional but robust ML methodologies, such as SVMs and ANNs approaches, have been also published recently addressing the survival prediction in patients with gastric, glioma, lung and bladder cancer [232-235]

Table 4.2. Relevant publications with regards to the use of ML applications in cancer survival prediction. The cancer type, total number of samples, data types along with the exact methods and the obtained results are presented for each study.

Publication	**Cancer Type**	**Dataset**	**Type of Data**	**Methods**	**Results**	**Validation Method**
Wulczyn E *et al.*, [227]	Multiple cancer types	9,086 image slides	Clinical, imaging	DL	hazard ratio [CI 95%]: 1.58 [1.28-1.70], p<0.0001	External validation
Park K. *et al.*, [205]	Breast Cancer	433,272 patients	SEER	DT	Acc. = 0.93	External validation
Zhu *et al.*, [232]	Gastric cancer	289 patients	Clinical, histological, laboratory	ANN	Acc. = 0.85	Cross-validation
Wang Y. *et al.*, [229]	Colorectal cancer	158,483 patients	SEER	Tree-based approach	Acc. = 0.70	10-fold cross validation
Kather J. N. *et al.*, [228]	Colorectal cancer	> 100,000 image patches	Imaging features	CNN	hazard ratio [CI 95%]: 1.99 [1.27-3.12], p = 0.0028	External validation

Table 4.2 *continued*

Publication	Cancer Type	Dataset	Type of Data	Methods	Results	Validation Method
Papp L. *et al.*, [233]	Glioma	70 patients	Medical imaging and demographics	Genetic algorithm	AUC = 0.90	14-fold cross-validation
Kate R. J. *et al.*, [231]	Breast cancer	174,518 patients	SEER	LR	AUC = 0.85	5-fold cross-validation
Hasnain Z. *et al.*, [234]	Bladder cancer	3503 patients	Clinical	SVM	Sens. = 0.70 Spec. = 0.70	Outer 10-fold cross validation Inner 3-fold cross validation
Lynch C. M., [235]	Lung cancer	10,442 patients	SEER	SVM	RMSE: = 15.05	10-fold cross validation

4.4 Cancer risk prediction using ensemble learning techniques

Recently, multi-modal fusion strategies applied at both the feature and the decision levels based on AI and ML techniques have emerged as a promising framework towards cancer prognosis prediction. Integrative AI prognostic modeling methods have been developed and deployed through ensemble-based ML principles. Although significant progress has been made towards precision medicine most cancer patients still don't receive individualized cancer treatment. Hence, many patients do not receive the necessary treatment affecting thereby both quality of life and clinical outcomes. A plethora of information is available with modern imaging and sequencing techniques (i.e. radiomics and genomics) which could increase the diagnostic potential and further optimize personalized treatment plans in cancer progression. To this end, integrating data from multiple sources into a high-precision clinical support system using ensemble ML techniques is an unmet clinical need due to the lack of different cancer data views, the heterogeneity of the data and the complexity of integrative AI frameworks. Fusion of different model priors that have been developed based on several data sources, could be achieved considering an ensemble fashion which may include both bagging and boosting [236]. Moreover, fusion strategies at the decision level have been proposed in terms of sparse ensembles and multi-modal learning scheme providing a framework for integrating different data modalities as well as different base classifiers [134, 237, 238].

In the literature this crucial clinical need is addressed through integrative diagnostics and prognostics approaches that empower personalized precision medicine in cancer patient stratification through the integration of the most common data sources (i.e. imaging data, clinical findings, histological and omics data). From the clinical perspective, the first advances are related to radiogenomics correlations like MRI-gene expression combinations for cancer patient survival [239]. Similarly, integrative multiscale analysis for assessing cancer progression by concatenating DNA methylation in conjunction with imaging data has been also studied in the literature [240]. However, there is still lack in such approaches due to the absence of comprehensive multi-modal fusion methods towards precision cancer decision support systems.

The emergence of deep learning methods and multi-dimensional data allowed the design and development of multimodal deep neural networks architectures for cancer diagnosis, prognosis and treatment. In [220] a novel and robust methodology is proposed with reference to DL architecture and to data integration with reference to breast cancer prognosis prediction. Specifically, gene expression data, copy number alteration profile and clinical variables were exploited after integrating them. The authors demonstrate the superiority of the proposed methodology and anticipate its robustness in the cancer clinical practice since it outperforms other conventional ML algorithms. They also confirmed the effectiveness of their study by comparing their results with experiments on separate data sources. The overall performance of the multi-modal deep neural network alongside the fusion of several data sources achieved an AUC value of 0.84.

A similar methodology that considers the use of multi-modal learning for predicting disease-genes associations have been proposed [241]. The method has been tested in several diseases, including cancer showing promising results in finding relationships among disease-genes by utilizing multimodal deep belief networks. In comparison with other well-known algorithms the proposed method was evaluated in terms of 5-fold cross-validation on a set of curated disease–gene associations achieving an AUC value of 0.96. Based on these results, the proposed method could accurately predict gene associations that are responsible for cancer progression. Furthermore, concerning the computer-aided diagnosis in breast, in [242] the authors have demonstrated the usefulness of dealing with multiple predictive models and thereby with the intelligent combination of their prediction probabilities. Data fusion of several SVM models was accomplished using generalized regression neural network (GRNN). Cross validation was performed for evaluating the performance of the model and an overall AUC of 0.81 was obtained.

Several studies in the literature have utilized ensemble methods at both the feature and the decision levels for improving the clinical prediction of cancer and thereby contributing to treatment de-escalation. Although the results proposed seem promising in terms of cancer prediction, most of these studies combine the predictions from multiple models on the same dataset based on machine learning techniques. Furthermore, stacked generalization [243], a method for combining estimators to reduce their biases has been proposed many years before for combining the output of several base classifiers. It is a schema for minimizing the generalization error rate of one or more generalizers. Stacked generalization works by deducing the biases of the generalizer(s) with respect to a provided learning set. The predictions of each individual estimator are stacked together and used as input to a final estimator to compute the prediction. In [244] a diagnostic model was developed for prostate cancer with both high accuracy and good interpretability. Clinical and demographic information was used for implementing the ensemble method.

Ensemble ML methods, such as DTs, RFs and GB algorithms could enhance the classification accuracy by aggregating the predictions of multiple base classifiers [142]. The rationale for ensemble methods is that the error rate during a classifier's performance is considerably lower than the error rate of the base classifiers, considering that the base classifiers are not identical but independent. Based on this knowledge, the integration of multi-omics data and the subsequent exploitation of ensemble techniques for integrative diagnosis, prognosis and treatment could contribute drastically in the application of robust approaches in the cancer clinical practice for better management of the disease and the decision making.

Multiple Kernel Learning (MKL) technique that permits data fusion in terms of kernels' integration has also emerged in the last decade with applications in cancer prognosis [134, 245]. In [246, 247] MKL transforms the data integration to kernel integration in the sample space including various data types such as omics data, clinical, treatment, histological and biomolecular data as well as individual gene sets. More specifically, in [246] omics data are utilized distinctly within multiple kernels for the classification of breast cancer subtypes. The proposed framework encompasses both SVM and MKL to accelerate the classification accuracy of breast cancer samples. The fusion of the heterogenous data occurs in the sample space; hence, the identification of informative features was achieved. In [247], the authors proposed a MKL framework for discriminating early-and late-stage cancers using gene expression profiles. Based on the predictive performance of this method, an improved understanding of the underlying molecular mechanisms was achieved which might have affected cancer progression.

Additionally, SimpleMKL [248] was employed which is a very efficient multiple kernel machine algorithm and it has been proven that outperforms many other algorithms. $L_2 - norm$ MKL methods are very efficient approaches and simpleMKL is based on a weighted $L_2 - norm$ regularization. Based on this knowledge, the authors combined SVM optimization and kernel fusion processes to one standard SVM optimization problem which allows to improve the glioblastoma clinical treatment. MKL is an intermediate fusion technique that firstly computes similarity matrices for each data view separately and then integrates these matrices to produce a final kernel learner for further use in a ML model.

4.5 Contribution of the Thesis

4.5.1 A DBN-based model for the prediction of oral cancer recurrence

As discussed in sections 4.1.1 and 4.1.2, modeling gene expression data through BNs and DBNs have been studied widely in the literature for identifying DEGs and pathways associated with cancer onset and progression. Network analysis using microarray data have been also introduced for investigating different network topologies between two disease subtypes and further suggest treatment options for improving cancer therapy and management. Moreover, the exploitation of pathway knowledge in terms of enrichment analysis could provide details about the interacting molecules; thus, significant pathways that are related to certain tumors may be unveiled. Several works have studied the network topology of significant genes to elucidate interactions that may contribute to mutations and thereby to cancer onset.

In the current thesis, we analysed gene expression microarray data available from open repositories for modeling the progression of cancer between different phenotypes. Specifically, we employed DBNs to incorporate both the significant genes that have been found to contribute to cancer recurrence along with the knowledge from the pathway level based on the enrichment analysis. Therefore, the most significant molecules were extracted and fed as input to the DBN model for the prediction of a disease relapse. The identification of highly connected genes that participate in the most overrepresented pathways, along with the DEGs determined in a previous work [13], compose the training set for the interaction network models. The contribution of the present thesis in the analysis of gene expression data and the inference of a DBN model for cancer prediction has been presented in [249]. The proposed method takes into consideration time series gene expression data in order to predict a disease recurrence. Subsequently, we can conjecture about the causal interactions between genes in consecutive time intervals.

4.5.2 Cancer classification from time series microarray data through regulatory DBNs models

As discussed in section 4.1.2, gene regulatory networks concern the deciphering of the relationships between transcription factors and their target genes. In cancer genomics the modeling of the molecular and cellular processes during tumorigenesis, by inferring networks of genes regulation, is of paramount importance. In the literature, the employment of biological data and the conduction of network analysis for the reconstruction of gene regulatory networks and thereby the extraction of disease subtypes has been studied thoroughly. Canonical correlation analysis [186] and DBN approaches [180, 185] have been applied to time series gene expression data for the inference of validated gene regulatory networks. To this end, these approaches with the inherent ability to represent longitudinal data allow the exploitation of time series data with reference to DBNs for deciphering stochastic processes, such as gene expression. In addition, the integration of gene expression changes and their respective transcription factor binding sites (TFBSs) could further contribute to cancer classification in order to detect gene profiles that can accurately distinguish samples of different phenotypes.

In the present thesis, we developed gene regulatory networks from microarray time series gene expression data for cancer classification through DBNs. The interactions between significant genes, i.e. (i) DEGs, and (ii) their master regulators (MRs), were identified and the classification performance of the DBN-based algorithm was further evaluated. Going beyond the state-of-the-art, we studied the integration of heterogeneous data for cancer classification through DBN models. MRs were identified based on the identified DEGs. The proposed methodology suggests that the combination of both DEGs and MRs into a DBN-based methodology could yield better classification results than the combination of DEGs into a simple classification scheme. The contribution of the present thesis in cancer classification based on time series gene expression data and on TFBSs has been presented in [250]. We identified the genes that act as regulators and mediate the activity of transcription factors that have been found in all promoters of our differentially expressed gene sets. These features served as potential priors for distinguishing tumor from normal samples using a DBN-based classification approach.

4.5.3 Predicting lymphoma development by exploiting genetic variants and clinical findings in a machine learning-based methodology with ensemble classifiers

Several clinical, serological and histopathological variables have been found as predictors for lymphoma development. These adverse risk factors have been found to be correlated with the aggressive behavior towards lymphoma development relative to the genetic background of patients. On this basis, lymphoma risk prediction in the context of Sjögren’s Syndrome have been studied widely in the literature in terms of statistical analysis and prediction rules based on clinical and biological predictors. As discussed in sections 4.2 and 4.1.3, data mining algorithms and ML-based methodologies have been also used for the identification of patient subgroups and the prediction of lymphoma as regards of features’ importances.

In the present thesis we proposed a robust ML-based pipeline which incorporates a list of sequential estimators for predicting lymphoma development. We aim at assessing the contribution of combined clinical, serological and histopathological features with genetic variants in predicting lymphoma. Special emphasis is given in the ML-based ensemble framework which encapsulates both RFs and GB algorithms to compare their performance and further identify the most accurate overall performance. The evaluation of the proposed models’ performance reveals the potential usefulness of integrating information from the genetic background of patients for predicting the risk for cancer development. The contribution of the current thesis in developing and implementing robust predictive models for cancer risk prediction has been presented in [251]. We highlighted the potential usefulness of genetic variants and clinical findings in predicting lymphoma development in Sjögren's Syndrome patients based on ensemble methods.

CHAPTER 5 DYNAMIC BAYESIAN NETWORKS FOR THE PREDICTION OF ORAL CANCER RECURRENCE

5.1 Introduction

It is generally argued that cancer is a disease characterized by abnormal cells growth that invades healthy tissues in the body. The last decade, rapid advances in cancer research community revealed that cancer is a complex disease with fluctuations in gene expression process at the molecular level. Oral Squamous Cell Carcinoma (OSCC) constitutes one of the most frequent neoplasm in humans [5, 57] and its mortality rate is known to be very high. It can be detected in any part of the oral cavity or oropharynx; thus, it may be referred to any malignancy that has been initiated in the head and neck region. Due to locoregional recurrence in cancer, early identification of a disease relapse can be crucial for the patient’s prognosis and treatment. Furthermore, extended investigation of the underlying molecular mechanisms and the disease progression may offer a crucial impact on the disease management and outcomes.

With the advent of high-throughput technologies, such as DNA microarrays, the expression levels of many genes can be measured simultaneously. The amount of such experimental data can be further analyzed in order to identify long lists of individual genes with their expression values which provide knowledge regarding the condition being studied. Furthermore, by grouping and analyzing, at a functional level, smaller sets of related genes allow us to identify: (i) groups of genes that function on the same and/or different pathways, and (ii) pathways that differ between two phenotypes [252]. Methods for the analysis of pathways and significant gene sets have been developed in the literature to gain insight into the functional mechanisms of living cells. Therefore, pathway analysis at the functional level is very appealing.

Lately, several studies in the literature have applied pathway analysis methods to microarray transcriptomic data, aiming to identify significant gene sets that are representative for a given pathway, as well as to explore the pathways which are related to a specific phenotype. Specifically, in [177] the authors have utilized Kidney Renal Clear Cell Carcinoma (KIRC) patients' sequencing data in order to identify a set of DEGs and pathways associated with the disease. After performing pathway and network analysis they suggested that distinct disease subtypes are correlated with different biological processes. In the same context, in [191] they proposed a network analysis method for angiogenesis in ovarian cancer. They identified different network topologies between the two disease subtypes and suggested possible therapeutic improvements for the treatment of ovarian cancer. Finally, in [174] the authors introduce a new method that scores for each pathway and tumor a pathway deregulation value; thus, each sample is characterized by its score and the stratification of the disease is performed in terms of pathway-based variables.

Modeling transcriptomic data in order to reveal how genes interact with each other and form an interaction network may provide more insights into the molecular processes and the progression of a complex disease. To this end, computational methods, such as DBNs, have been proposed in the literature for the inference of the underlying structure of gene networks [185]. These approaches integrate genome-wide expression data for the inference of gene regulatory networks through the utilization of DBNs.

In the current study, transcriptomic data were exploited in order to perform gene-based pathway enrichment analysis of OSCC patients. More specifically, a set of DEGs among the two groups of patients in the dataset, i.e. (i) patients that have suffered a disease relapse after complete remission, and (ii) patients that have not suffered a relapse after complete remission of the disease were employed. Moreover, pathway enrichment analysis was performed for predicting the disease recurrence through the utilization of DBN models. Compared to a previous study [253], we have updated several parts aiming to combine the knowledge from the pathway level and the DBN modeling methodology for the prediction of a disease relapse. We also conducted an overrepresentation analysis and detected the pathways which are enriched in our gene set. The extraction of highly connected genes that participate in the most overrepresented pathways, along with the DEGs determined in [190] compose the training set for the interaction network models. The derived results indicate that the combination of the specific gene set with the highly connected nodes from the Pre-NOTCH Expression & Processing (PNEP) pathway can provide the most accurate prediction of oral cancer recurrence.

5.2 Materials and Methods

The proposed methodology consists of three main steps. In the first step, transcriptomic data is analyzed in order to identify a subset of the most differentially expressed genes among the two groups of patients. In the second step, pathway enrichment analysis is performed for the specific gene list aiming to identify the most significant pathways in terms of overrepresentation. Finally, highly connected genes that participate in the most significant pathways are extracted. This gene list along with the genes determined firstly as differentially expressed, constitute the training set for the development of the interaction networks regarding the prediction of OSCC recurrence.

5.2.1 Transcriptomic Dataset

In the current study, transcriptomic data from 23 patients that have been diagnosed with OSCC and had reached complete remission, were considered [190]. For each patient, data from circulating blood cells have been collected and the 4x44K oligo-RNA human genome array, from Agilent Technologies was utilized. The measurements of the gene expression values were conducted during the baseline state and the follow-up period of each patient; thus, time course data are used in each step of the proposed methodology. During the follow-up period and for a 24-month time span, blood genomic data were collected from each patient regularly, during scheduled visits planned in consecutive time intervals. During this time, the possibility of a disease relapse was studied.

According to the occurrence or not of a disease relapse during the follow-up study, patients have been discriminated into two groups, respectively. The first group includes the 12 patients out of 23 that have suffered a disease recurrence, while the second one consists of the remaining 11 patients that have not suffered a disease relapse during the follow-up study.

Since the initial transcriptomic data file consists of expression values of many entries, some basic filtering steps were applied upon the raw data according to [190]. These steps refer to the removal of duplicate and control features, as well as genes of low quality or high rates of missing values. Duplicate features correspond to genes that are printed in the array more than once in random positions, thus, they are excluded from the dataset. Subsequently, an algorithm for microarray analysis is employed in order to extract a subset of the most differentially expressed genes between the two groups of patients. This gene list is then exploited aiming to perform pathway enrichment analysis.

5.2.2 Dataset Formulation

The initial genomic data file consisted of 45,015 expression values for each patient. After the filtering steps, the output was a set of 33,491 entries which were then fed as input to the Significance Analysis of Microarrays (SAM) statistical technique [254], aiming to identify a limited subset of the most DEGs. SAM searches for genes that differ significantly in terms of their expression during the follow-up period. The final gene list was determined upon the False Discovery Rate (FDR) [142] of the gene expression values between the two groups of patients. More specifically, the Wilcoxon statistical test was performed which identifies those genes that are mostly differentially expressed between the groups. The threshold for the fold change between the two groups of patients was set to 1.8 according to [190].

5.2.3 Pathway Enrichment Analysis

The subset of genes that were found to be most differentially expressed between the two groups of patients, were further considered along with their expression values in order to perform pathway enrichment analysis based on the Reactome database information [255]. Reactome is a curated, peer-reviewed database of human pathways and processes. The exploitation of the Reactome tools allowed us to perform pathway analysis of our datasets and further explore the overrepresentation of pathways in the submitted data. After the assignment of specific pathways to the supplied list, we studied whether the genes that represent OSCC risk associated genes are assigned by chance within specific pathways or refer to significantly disrupted pathways correlated with the disease in terms of the p-value and the FDR score that were calculated automatically by the analysis tool. The criteria considered before choosing the most enriched pathways for further analysis in the current work were: (i) the selection of the first sub-pathways of any single pathway, instead of large pathways or reactions according to the pathway hierarchy panel, and (ii) the selection of those pathways that were found enriched with the most entities of our submitted gene list. These criteria were selected in order to identify the most critical genes within the pathways found enriched.

Additionally, in order to identify the highly connected nodes of the most significant pathways and proceed to the next step of the proposed methodology, the functional interaction network of each pathway was derived. The Cytoscape software platform [256] was utilized and the Reactome plugin was applied. Thus, we were able to extract all the information regarding the network and the connected nodes.

5.2.4 Dynamic Bayesian Networks Models

The approach of DBNs has been widely used for the inference of gene interaction networks from time series microarray data; thus, they constitute a suitable choice for modeling oral cancer recurrence, as well.

DBNs are an extension of BNs which encode the joint probability distributions over a set of random variables $X = \{x_1, ... x_n\}$. A BN is a pair B=(G, Θ). The first component G is an annotated directed acyclic graph and the second one Θ represents the parameters that quantify the network. Given G and Θ, a BN, B, defines a unique joint probability distribution over X given by:

$$P_B(x_1, .. x_n) = \Pi_{i=1}^{n} P_B(x_i | pa(x_i)), \tag{5.1}$$

where $pa(x_i)$ denotes the parents of x_i, in G.

DBNs are tuned to model the stochastic processes of a set of random variables over time [184]. More specifically, they can describe causal interactions of stochastic processes between the state variables, thus, their application in complex systems may provide a better approximation of the interactions underlying the molecular processes. DBN theory is generally based on two assumptions. First, the process is Markovian in the set of variables X, i.e. $P(X[t+1]|X[0],\dots,X[t]) = P(X[t+1]|X[t])$. Second, it is assumed that the process is stationary, i.e. the transition probability $P(X[t+1]|X[t])$ is independent of t. To represent beliefs about the possible trajectories of the process, we need a probability distribution over random variables for all t. A DBN that represents the joint distribution over all possible trajectories of a process consists of two parts:

- a prior network B_0 that specifies a distribution over initial states $X[0]$, and
- a transition network, $B_{\rightarrow}$, over the variables $X[0] \cup X[1]$ that is taken to specify the transition probability $P(X[t+1]|X[t])$ for all t.

Given a DBN model, the joint distribution over $X[0],\dots,X[T]$ is:

$$P_B(x[0],..,x[T]) = P_{B_0}(x[0])\Pi_{t=o}^{T-1}P_{B_{\rightarrow}}(x[t+1]|x[t]). \qquad (5.2)$$

DBNs can be defined by a graphical structure and a set of parameters. Therefore, in order to construct a DBN we need to specify the intra-slice topology (connections within a slice), the inter-slice topology (connections between two slices) and the parameters for the first two slices. Figure 5.1 depicts a simple structure of a DBN model. Two time slices (t=0 and t=1) are illustrated, as well as the topology among the variables.

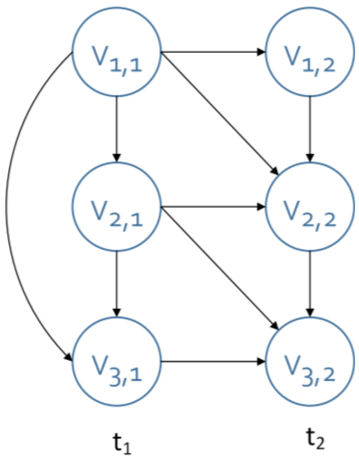


Figure 5.1 A simple DBN structure, where $V_{i,j}$ corresponds to the i_{th} variable in the j_{th} time-slice.

In the current work we employ the training dataset as determined above, in order to define the structure and the parameters of two DBN models related to the status of a specific patient, i.e. relapser or no-relapser. The parameters were specified according to the variables within the first time slice and across the first and the second time slice. They were represented as CPD (Conditional Probability Distribution) objects. The parameters of the models were specified as conditional linear Gaussian distributions [21]. Regarding the development of the DBN models, the junction tree engine, which is the source of all the exact inference algorithms, was used for inference [22].

We implemented our models in MATLAB, using the Bayes Net Toolbox (BNT) [178] and the Dynamic Bayesian Markov Chain Monte Carlo (DBmcmc) [178] packages. In addition, the Canonical Correlation Analysis-based (CCA) algorithm [186] was used aiming to compute potential interactions between genes. The results of CCA algorithm constitute the prior knowledge needed for the construction of the interaction network models.

5.3 Results and Discussion

5.3.1 Pathway Enrichment Analysis

According to the initial transcriptomic dataset and the application of the statistical SAM technique, we identified a subset of significant genes between the two groups of patients considered in the current study. Table 5.1 depicts these genes as pinpointed by the employment of the algorithm. This subset contains nine genes with their expression measurements which were further analyzed in order to perform pathway enrichment analysis through the utilization of the Reactome pathway database. Thus, gene identifiers from the submitted data were mapped to certain pathways. According to the analysis results per

Table 5.1 List of the differentially expressed genes after the employment of the SAM statistical technique.

Gene IDs		
AK023526	HMCN1	LEPRE1
NOTCH2	RGMA	STX6
THC2344152	THC2447689	TSC1

pathway, we observed that the pathways with the highest proportion to the input list are: (i) the Pre-NOTCH Expression and Processing (PNEP), and (ii) the Generic Transcription (GT) pathways. In a previous report [257], pathway enrichment analysis has been performed on the same transcriptomic data and six overrepresented pathways were identified and presented. In the current study, we only exploited two of the six enriched pathways, which are present in the Reactome pathway hierarchy and have the largest number of mapped gene IDs. Moreover, in comparison with the other enriched pathways, this pathway along with the PNEP were presented as first sub-pathways in the hierarchy panel. The PNEP pathway consists of the NOTCH gene family. Their transcription is developmentally regulated and is tissue specific, but little information exists on the molecular mechanisms of the transcriptional regulation [258]. Concerning the GT pathway and the gene transcription regulation in eukaryotic systems, the general principles and mechanisms by which cell- or tissue-specific regulation of differential gene transcription is arbitrated have been revealed [258].

Table 5.2 indicates the analysis results per pathway. The columns represent (i) the pathway name at the lower level of the Reactome pathway hierarchy and (ii) the number of submitted genes that map the pathway, respectively. It should be noted that the table contains only the first 10 enriched pathways with the number of entities that were found in each pathway. Since our criteria for selecting the pathways is based on the enrichment event, we only present here the pathway names along with the number of IDs from our submitted gene list.

Based on the results of the overrepresentation analysis and the selected enriched pathways, we proceeded with the construction of the interaction network models. As Table 5.2 depicts, the Pre-NOTCH Transcription and Translation pathway has also been found to be enriched to the submitted gene list and contains the same number of genes as the PNEP pathway. However, since this pathway is involved in the PNEP event according to the Reactome hierarchy and it is not represented as a first sub-pathway at the pathway hierarchy panel, was not considered for further analysis.

Table 5.2 Reactome analysis results per pathway.

Pathway name	IDs
Pre-NOTCH Expression and Processing	3
Pre-NOTCH Transcription and Translation	3
Notch-HLH transcription pathway	1
Signaling by NOTCH	2
Pre-NOTCH Processing in Golgi	1
Defective LFNG causes SCDO3	1
Pre-NOTCH Processing in the Endoplasmic Reticulum	1
Signaling by NOTCH4	1
Inhibition of TSC complex formation by PKB	1
Generic Transcription Pathway	2

In the next step, in order to identify the most significant nodes in each pathway, based on the number of connected interactors, we inferred their functional interaction network through the utilization of the Reactome FIViz plugin [256]. Hence, we were able to detect all the interacting partners and further select only the highly connected ones. Specifically, only the genes that had 10 or more neighbors were selected as highly connected. As reported in the literature [26], hubs are defined as nodes with connectivity greater than 5. In the current work, we considered a connectivity threshold equal or greater than 10 in order to

Table 5.3 Highly connected genes along with the number of neighbors in the PNEP pathway.

Gene IDs	Number of neighbors
NOTCH1	24
CCND1	15
NOTCH4	15
CREBBP	14
E2F1	12
E2F3	12
EP300	12
NOTCH3	12

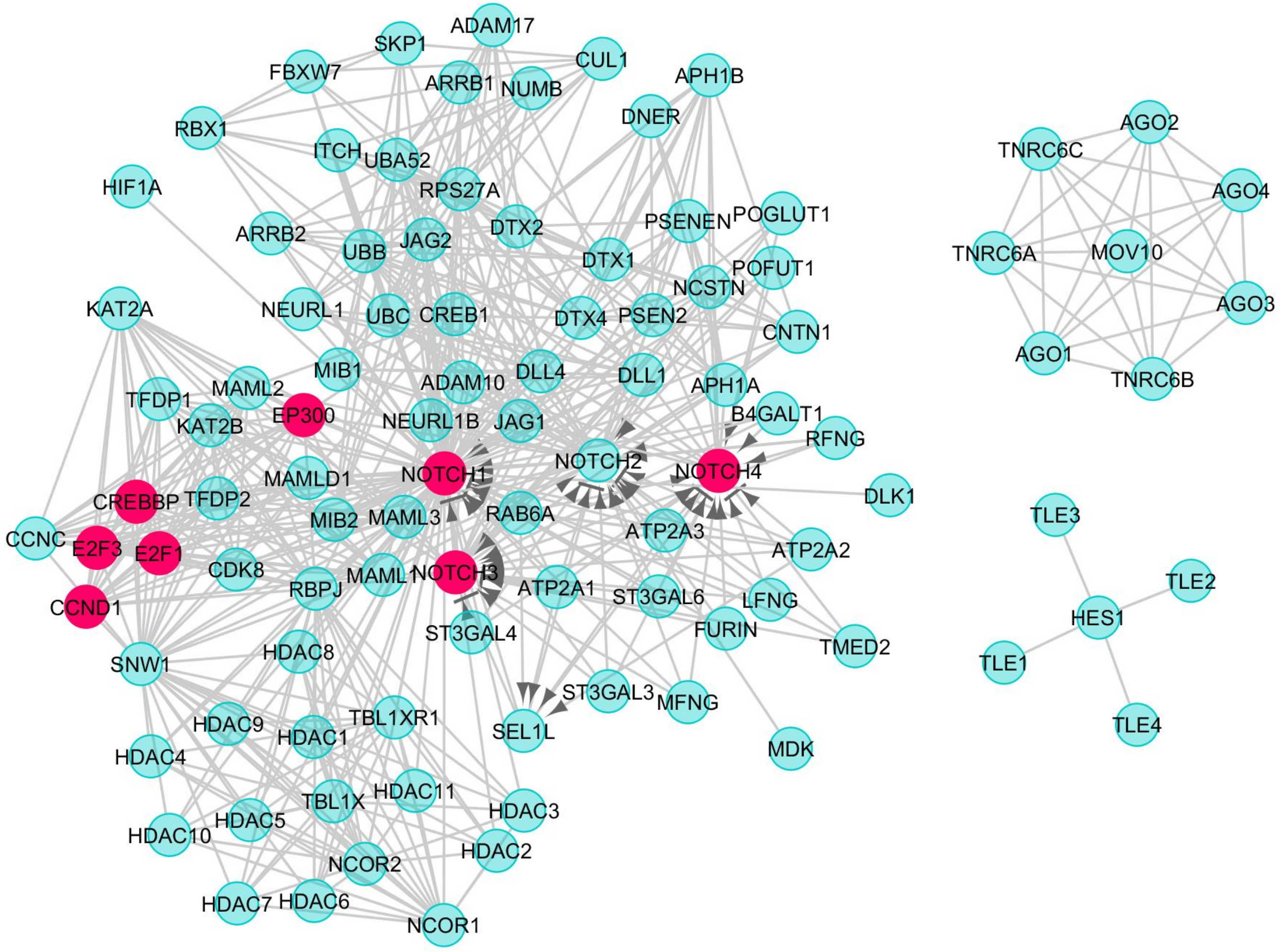


Figure 5.2 The functional interaction networks of the PNET pathway, in terms of the Reactome FIViz plugin. Circles correspond to genes with their respective edges/connections. Known functional annotations are presented as edges with arrow shape.

identify only the genes that are more plausible candidates to function as signaling centers. Table 5.3 depicts the highly connected genes in the PNEP pathway, which has the highest proportion to the input list, along with the number of neighbors for each gene. Figure 5.2 illustrates the interaction network underlying the functional relationships in PNEP pathway. Genes are connected based on the protein functional interaction network of the Reactome plugin in Cytoscape. They are depicted as nodes along with their respective interactions. The highly connected genes of the pathway are shown in red. Connections that have known functional annotations based on the Reactome plugin are illustrated with arrow shapes, whereas unknown interactions are presented as simple lines. It should be noted that the functional interaction network of the GT pathway was also derived (Appendix I). We herein present only the illustration of the PNEP pathway as it was identified as the most overrepresented pathway according to the submitted gene list. The extraction of the most significant genes which participate in both pathways along with the DEGs that were identified from the initial dataset, constituted the training set for the construction of the DBN interaction network models.

5.3.2 Prediction of OSCC recurrence through DBN models

Knowledge from the pathway level, regarding the transcriptomic dataset, was further exploited in order to build a model that can assess the prediction of a disease relapse. The gene expression values of the significant genes constituted the training set for the DBN algorithm. More specifically, the training set consists of the expression values of the DEGs in two time slices along with the expression values of the highly connected genes of the enriched pathway found in the analysis.

The leave-one-out cross validation technique was repeated 100 times with independent initializations in order to further reduce the bias introduced by the stochastic nature of the DBmcmc algorithm [178]. The 100 times were empirically selected in order to

Table 5.4 Results for each predictive model.

Dataset	Acc.	AUC	Sens.	Spec.	F-score
SAM & PNEP	81.8%	89.2%	76.9%	90.1%	83.4%
SAM & GT	54.5%	59.3%	54.5%	54.5%	54.5%

handle this bias. Due to the limited number of samples in our dataset, cross validation is a suitable technique for estimating accurately the performance of model prediction. The reported overall accuracy (Acc.) for the prediction of OSCC recurrence is 81.8%, the specificity (Spec.) is 90.1%, sensitivity (Sens.) is 76.9%, F-score equals to 83.4% and AUC is 89.2% (Table 5.4). In a same manner, we also estimated the prediction of a disease relapse through the utilization of the expression values of the highly connected genes that participate in the second most overrepresented pathway, namely GT pathway, along with the gene expressions of the most differentially expressed ones which constituted the training set for the DBN algorithm. Subsequently, the interaction network models were inferred for the relapsers and no-relapsers. The Acc. and AUC were 54.5% and 59.3%, respectively (Table 5.4), indicating that the employed criteria were enough since the second overrepresented pathway did not provide higher results. Figure 5.3 depicts the AUC curves for the DBN algorithm in the two training sets: (i) the DEGs along with the PNEP pathway's highly connected nodes (SAM & Pre-NOTCH Expression and Processing pathway) training set, and (ii) the DEGs along with the GT pathway's highly connected nodes (SAM & Generic Transcription pathway) training set. It should be argued that using the expression values of

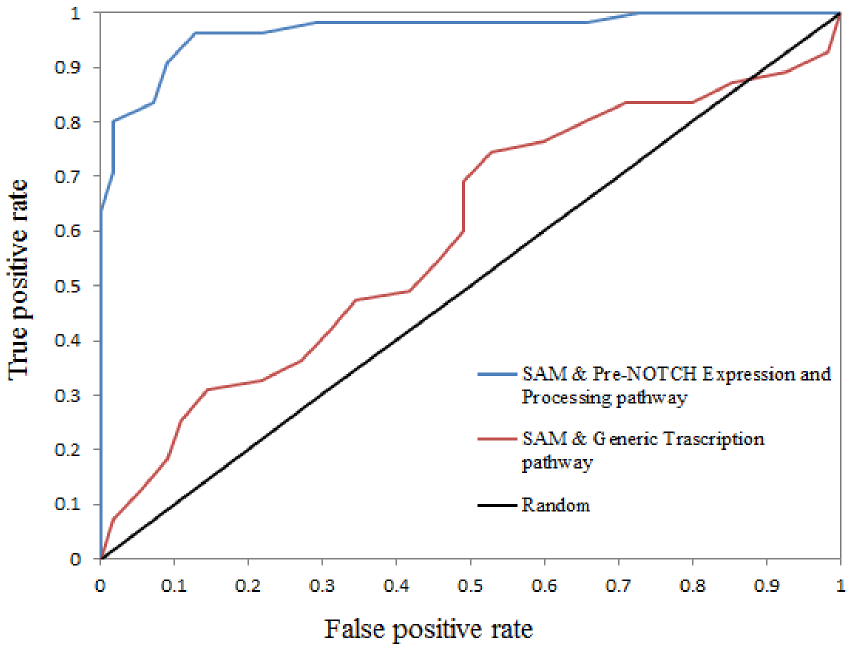


Figure 5.3 ROC curves for the performance of the predictive DBN algorithm in two training sets: (i) SAM & Pre-NOTCH Expression and Processing pathway training set, and (ii) SAM & Generic Transcription pathway training set.

Table 5.5 Most frequent interactions in the DBN model about relapsers.

RELAPSERS	
Intra-slice interactions	
GeneA	GeneB
THC2447689	NOTCH3
E2F1	E2F3
CCND1	E2F1
TSC1	E2F1
NOTCH3	CREBBP
Inter-slice interactions	
GeneA	GeneB
HMCN1	E2F3
AK023526	CREBBP
CCND1	EP300
STX6	EP300
NOTCH3	E3F3

the genes that are differentially expressed along with the expression measurements of the highly connected genes in the PNEP network, the model achieves much higher accuracy in terms of the algorithm’s predictive performance.

In the same manner, we also estimated the most frequent connections among genes in the intra- and inter-slice topology of the model for no relapsers. In addition, we further explored whether each one of the predicted interactions, for both groups of patients, have been validated experimentally in the literature regarding the OSCC progression and have been supported by the Human Protein Reference Database (HPRD) database [259].

Table 5.5 and Table 5.6 present the interactions between the nodes of the network models as computed by the DBN algorithm for the patients that had and had not suffered a disease relapse, respectively. The intra- and inter-slice interactions in the network model structure are shown in relation to the training set of significant genes. Specifically, in order to further evaluate the inferred interactions among genes, we calculated the number each interaction was observed after the 100 iterations of the learning simulation. Thus, we sorted the interactions according to their observations after all iterations, and we selected the most informative and frequent ones.

In addition, we further explored whether each one of the predicted interactions for both groups of patients, have been validated experimentally in the literature regarding the OSCC progression and have been supported by the HPRD database [259]. Considering the integration of knowledge from the pathway level for predicting cancer recurrence, we have utilized and tested certain pathway interactions separately through the utilization of the Reactome pathway database analysis tool. Thus, we were able to identify whether a given relationship exists in specific pathways and what kind of relationships occurs among the gene sets. Moreover, we also further explored the inferred interactions through the utilization of the National Center for Biotechnology Information (NCBI) Gene database [27]. From all the detected interactions presented in the current study, it is worth to note that CCND1 → EP300 was found to be overrepresented in significant Reactome pathways, such as the PNEP

Table 5.6 Most frequent interactions in the DBN model about no relapsers

NO RELAPSERS	
Intra-slice interactions	
GeneA	GeneB
NOTCH2	LEPRE1
TSC1	LEPRE1
NOTCH3	HMCN1
E2F1	LEPRE1
STX6	THC2344152
Inter-slice interactions	
GeneA	GeneB
NOTCH4	NOTCH2
EP300	AK023526
CREBBP	E2F3
AK023526	LEPRE1
EP300	THC2344152

and the Cell Cycle-mitotic pathways. This interaction has been validated experimentally and is supported by the HPRD source [259], as well. Specifically, it has been shown that Cyclin D1 (CCND1) gene associates with the C-terminal domain of the p300 transcriptional co-activator protein (EP300 gene) [260]. In addition, mutations, amplification and overrepresentation of the CCND1 gene alter the progression in cell cycle and are observed in a variety of tumors; thus, may be partly responsible for tumorigenesis. In [261] the authors explored the role of EP300 in tumor progression of OSCC. They suggest that the gene itself or one of its targets play a key role in the aggressive phenotypes of the specific disease, based on its overexpression and its association with clinical factors in patients suffering from OSCC. It should be highlighted that this interaction was identified accurately by the algorithm in the inter-slice topology of the model regarding the relapsers. Thus, it constitutes a true positive result which has been also verified experimentally and may provide better insights into the underlying molecular processes of the disease recurrence. The CREBBP → E2F3 interaction was also found overrepresented in the PNEP Reactome pathway. Both genes are involved in the NOTCH1 gene transcription. Subsequently, we detected that this connection is also supported by [28]. Although this finding has not been validated experimentally, it is generally argued that altered activity of the E2F3 transcription factor has been observed in several human cancers [27].

The interaction network prediction methodology was also compared with similar works in the literature [174, 177]. We observed that the proposed methodology exhibits promising results regarding the combination of gene expression profiles with pathway and interaction network analysis for: (i) the identification of altered interactions, and (ii) the prediction of a possible disease recurrence. Nevertheless, direct comparison could not be achieved since different datasets have been exploited on each work and different predictive algorithms have been utilized. Concerning the results of the current methodology, a remarkable advantage is the prior knowledge that has been integrated in order to construct the predictive model. Time series gene expression measurements from two consecutive time intervals were exploited and distinctly expressed genes between the two groups of patients have been identified. This knowledge was further analyzed in terms of pathway overrepresentation analysis which constituted subsequently the prior knowledge for the DBN algorithm. On the other hand, other approaches that have been proposed [174], take into consideration only knowledge from biological pathways in order to perform pathway analysis of expression data and further identify significant pathways related to a disease. The current work allows for the combination of knowledge from the transcriptomic and pathway levels in order to model the prediction of OSCC recurrence through the utilization of a DBN model. A future research direction is the enrichment of the dataset with more patient records as well as on better handling missing data in the new dataset with imputation techniques [142]. Moreover, the new dataset will contribute as a validation set on the currently developed predictive model.

5.4 Conclusions

In this study, we proposed a methodology that exploits transcriptomic data along with pathway knowledge aiming to predict OSCC recurrence through the employment of a DBN algorithm. The obtained results indicate that the integration of time series gene expression data and of distinctly expressed genes among the two groups of patients can provide better knowledge regarding the prediction of a disease relapse. Our methodology can be extended to other types of cancer as well as integrate pathway knowledge from various biomedical databases in order to further assess regulatory interactions in each biological network.

CHAPTER 6 CANCER CLASSIFICATION FROM TIME SERIES MICROARRAY DATA THROUGH REGULATORY DYNAMIC BAYESIAN NETWORKS

6.1 Introduction

The complex nature of the global cancer landscape unveils the continuous growth of the expected numbers of new cancer cases and deaths [8]. Improved understanding of cancer trajectories, from diagnosis to treatment, disease recurrence, late effects and comorbidities, could enhance the decision making in healthcare systems, towards a more precise and personalized patient management in almost every cancer type. In addition, the elucidation of the intertwined mechanisms at multiscale levels (i.e. molecular, cellular, etc.) through *in silico* medicine could empower cancer diagnosis at early stages and accurately predict any possible progression, leading to a better design of targeted therapeutic protocols [130].

In order to uncover the cancer biology across multiscale levels, detailed and comprehensive pathway and network-based descriptions with regulatory relationships should be considered [262]. Gene regulatory networks have been widely studied for deciphering the relationships between transcription factors and their target genes. Modeling the molecular and cellular processes during tumor progression, by developing networks of genes regulation, is of paramount interest in cancer genomics [262]. Several methods have been reported in the literature for network analysis and the reconstruction of gene regulatory networks with reference to the integration of biological data [171, 175, 186]. These studies employ gene expression data for inferring gene networks, while the identification of key regulatory elements (i.e. transcription factors) is also considered. Therefore, transcription factors and gene interactions that may be associated with the disease progression and subsequently may contribute to the design of patient oriented therapeutic protocols are retrieved. In [186], canonical correlation analysis and DBNs were applied to time series gene expression data for the inference of validated gene regulatory networks. In a similar manner [171], the construction of gene regulatory networks was achieved by using a Partial Least Squares (PLS) based feature selection algorithm. In [175, 180], Bayesian approaches were considered for network analysis by employing biological data and especially time series gene expression measurements. As aforementioned, among the different methods for modeling gene regulatory networks, BNs and DBNs have been systematically used for modeling gene expression changes over time [181, 184], in terms of regulatory network structures. These approaches with the inherent ability to represent the time-varying behavior of the underlying biological network allow for a better representation of spatiotemporal input-output dependencies. Therefore, the exploitation of time series data with reference to DBNs has been proven a valuable strategy for deciphering stochastic processes, such as gene expression.

With the advent of high-throughput sequencing technologies, multiple genome datasets related to the gene expression changes during disease onset and progression have been generated. Gene expression patterns extracted with reference to genomic and transcriptomic datasets could enhance the identification of differences between cancerous and normal cells. Hence, a better classification of cancer patients into distinct groups enables the development of more accurate diagnostic and/or prognostic models. In terms of tumor classification, several cancer types have been studied by employing microarray gene expression data [202]. These studies contributed to the identification of gene patterns and revealed the associations of gene expression differences with clinical outcomes. In addition to the exploitation of gene expression data for studying the underlying molecular mechanisms, the identification of transcription factors which actively regulate and mediate the expression of specific genes by their activity is also crucial for cellular processes in the biological systems [263]. It has been shown that a small number of these transcription factors can act as regulators. These regulatory molecules are considered as promising drug targets, since their alteration can influence the transcriptional mechanisms of the respective transcription factors. Subsequently, the combination of gene expression changes with transcription factor binding sites (TFBSs) could further contribute to cancer classification in order to identify gene profiles which accurately distinguish cancerous samples [264].

In the present study, gene regulatory networks were built from microarray time series gene expression data for cancer classification through DBNs. The linkages between important genes, i.e. (i) DEGs, and (ii) their master regulators (MRs), were identified and the classification performance of the DBN-based algorithm was further evaluated. Statistically significant genes, which have been identified after differential expression analysis, were considered for the identification of their regulatory molecules. The novelty of the current work pertains to cancer classification through DBN models based on MRs which have not been utilized so far for classification and prediction purposes. In terms of time series microarray data, our study suggests that the integration of both DEGs and MRs into a DBN-based methodology could yield better classification results than the combination of DEGs into a simple classification scheme. In addition, the exploitation of molecules from the pathway level enables researchers to gain better insights into the underlying complex molecular processes of cancer. The presented DBN-based classification models demonstrate high discrimination and predictive power of cancer and non-cancer samples. The impact of investigating the temporal dependencies among genes from time series microarray data is also revealed in the current work. As DBNs enable the inference of temporal relationships between state variables, they can be considered as a better approximation of the actual stochastic process.

6.2 Materials and Methods

6.2.1 Transcriptomic datasets

Microarray time series gene expression datasets were retrieved from the NCBI's GEO functional genomics public repository [102]. In the current study, the selection query of the most appropriate datasets in the current study was based on three keywords, i.e. (i) "gene expression profiling", (ii) "time course" and (iii) "cancer type". The datasets with the highest number of samples and recent published results were finally selected. The R/Bioconductor software package GEOquery [265] was utilized to download the indicative GEO Datasets, which represent curated collections of biologically and statistically comparable GEO samples. The three different cancer types selected to apply our computational workflow were: (i) Pancreatic Ductal Adenocarcinoma (PDAC), (ii) Colon Cancer (CC), and (iii) Breast Cancer (BR). The respective gene expression profiling studies selected from GEO were GSE14426, GSE37182 and GSE5462. These datasets correspond to measurements of gene expression in cancerous and non-cancerous tissues and blood samples at different time points during the follow-up period. More specifically, in the GSE14426 study a pancreatic stellate cell line in plastic culture wells was treated with all-trans retinoic acid (ATRA) for 5 timepoints: 30 mins, 4 hours, 12 hours, 24 hours and 168 hours, to evaluate the post-treatment genes' expression changes. In total, 30 control and ATRA RNA samples were used and 48,701 Illumina identifiers were measured in terms of microarray analysis. In the GEO study GSE37182, RNA was extracted from sample specimens of 14 patients at four post-surgery time-points. These samples were further analyzed for gene expression changes. A total of 48,803 Illumina identifiers were studied and a mixed-effect model was used to identify the probes with different expression means across the four different time points. In the third dataset (GSE5462), sequential biopsies of the same cancers from a group of postmenopausal women with large operable or locally advanced breast cancer before and after 10 to14 days of treatment with letrozole were taken. In a total number of 58 patients, microarray analysis was performed and the expression of 22,283 reference IDs was measured. In order to explore the distribution of the values for the samples we have selected (i.e. the values of the original submissions), the GEO2R interactive web tool [102] was utilized. Assessing the distribution is important to determine the suitability of the selected samples for any comparison and for the identification of DEGs. Generally, median-centered values are indicative that the gained data are normalized and cross-comparable.

6.2.2 Time course differential expression analysis of microarray studies

Differential expression analysis was performed using the R/Bioconductor limma package [266] which is an appealing choice to analyze data from experiments involving microarrays. Limma operates on a matrix of gene expression values or other genomic feature, where each column corresponds to an RNA sample. We extracted genes that have been differentially expressed over time in cancerous and normal samples in order to detect changes of gene expression during disease progression or their association with the applied treatment protocol. Specifically, transcriptomic datasets were selected for time course differential gene expression analysis. Limma operates by fitting linear models to a matrix of gene expression values to handle differences in variability among genes and samples. A linear model can be considered as:

$$E[y_g] = X\beta_g. \tag{6.1}$$

For each gene g, we have a vector of gene expression values y_g and a design matrix X which relates these values to some coefficients of interest β_g. Except gene-wise analysis, limma also empowers higher-level analysis of gene expression profiles by gene-wise independence or interaction and the decomposition of gene signatures into distinct molecular pathways. The utilization of the limma approach facilitated the analysis of genomic experiments selected as a whole; thus, any correlation that may exist between the samples could be revealed.

The detection of DEGs in a statistically rigorous manner was achieved by setting the FDR threshold to 0.05 [267]. *p-values* were calculated to detect the statistically significant changes in gene expression. For all the datasets considered, the *p-value* threshold was set to 0.01. Additionally, to infer the most up- and down-regulated genes the *log2 fold change* cut-off was set to 1.5 (for GSE14426) and 0.5 (for GSE37182 and GSE5462 datasets). Different cut-off values for the *log2 fold change* were used since the proposed DBN methodology suffers from computational complexity for large number of input genes. Keeping the input gene list to a minimum should be a requirement for the efficient design of the proposed DBN-based algorithm. Therefore, a list with the most statistically meaningful genes could be retrieved.

6.2.3 Upstream analysis for regulatory molecules

To enable a promoter-pathway interpretation of the identified genes according to the differential analysis workflow, a pathway analysis strategy which is implemented in the GeneXplain platform [268] was adopted. MRs that are known to be linked with cancer onset and progression were identified. GeneXplain is a commercial online workbench which provides several bioinformatics and systems biology functions and enables several standard statistical and systems biology analyses. In the upstream analysis a list of DEGs is considered for performing the promoter and pathway analysis. In the first step, potential TFBSs were identified in all promoters and enhancers of the DEGs of the experiment under study and in a negative control set, as well. Hence, transcription factors which are characteristic of the gene set under study and have potentially regulated these DEGs were identified. The comprehensive matrix library TRANSFAC was used for the sequence analysis [269]. Step 1 resulted in several transcription factors possibly responsible for the differential regulation of the identified DEGs set. In the next step, signaling pathways which have been shown to be activated by the hypothesized transcription factors were reconstructed by employing the TRANSPATH database [270]. TRANSPATH contains information about all relevant signaling cascades that regulate the activity of the transcription factors. Subsequently, the molecules which converge in the selected pathways were considered as master regulators. The algorithm for TFBS enrichment analysis, called F-Match, has been described in [270]. The aim of this algorithm is to find nodes in the global signal transduction network that may potentially regulate the activity of the set of transcription factors found in the first step of the analysis. These nodes could be considered as most promising drug targets, since any influence on a respective node may switch the transcription mechanisms of several DEGs which are regulated by the related transcription factors.

As mentioned previously, two databases were utilized in the analysis workflow for performing the promoter and the pathway analysis, i.e. the TRANSFAC and the TRANSPATH databases. The negative control set for each experiment was the set of genes which are required for the maintenance of basal cellular functions (i.e. “Housekeeping genes”) [270]. The TRANSFAC profile “all Human 1 in 10k base” was also selected. A profile is a set of matrices and their cut-offs designed for function-driven searches within regulatory regions of genes whose function is partially known. The start locus for detecting the promoters was set to -1000 and the end to 100 as annotated in the Ensembl genome database [271]. Important regulators were discovered in signal transduction pathways on the network of the TRANSPATH database with a default cutoff for *Score* at 0.2, for FDR at 0.05 and for *Z-Score* at 1.0 [270].

The *Score* value of each master regulatory molecule reflects how well this molecule relates to other molecules in the database, and how many molecules from the input list are present in the network of this master molecule. The score is computed for each potential master regulator and it reflects a certain balance between sensitivity and specificity of signal transduction from this master node to the downstream effector transcription factors [272]:

$$S(k) = \sum_{k=1}^{k=max} \frac{M_k}{\left(1+k*\frac{N_k}{N_{max,k}}\right)* M_{max,k}}, \tag{6.2}$$

where k is the radius of pathway steps (i.e. maximal radius) from the master node to the effector nodes, M_k is the number of input transcription factors reached by a signal from the master node within k steps, and N_k is the total number of all potential steps from the master node to the effector nodes, M_k is the number of input transcription factors in the database reached by a signal from the master node within k steps. $M_{max,k}$ and $N_{max,k}$ are the highest values among all possible master regulator nodes which help to normalize the score in the (0,1) interval. The higher this score, the more sensitive and more specific this master regulator is for the set of input transcription factors. The parameter k is a user-defined penalty (default value is 0.1) [272]. The FDR corresponds to the expected proportion of false positives (Type I errors). The *Z-score* value reflects how specific each master molecule is for the input list. The higher the *Z-score* value for a molecule, the more specific this molecule is for the input list, and the lesser is the probability to find such a molecule as master regulator in another analysis. *Score* and *Z-Score* reflect separate characteristics of the suggested master regulators.

Overall, according to the output of the above upstream analysis, the annotated regulatory molecules, were characterized by four metrics: (i) *Score,* (ii) *Z-score,* (iii) FDR and (iv) *Ranks Sum* [270]. The *Ranks Sum* score corresponds to a sorting approach which combines the *Score* with the *Z-Score*. Finally, each identified molecule was ranked upon the *Ranks Sum* score. The lower the *Ranks Sum,* the more promising in terms of candidate master regulator the molecule is.

6.2.4 Dynamic Bayesian Networks for modeling time series microarray data

DBNs enable the modeling of stochastic phenomena, the incorporation of prior knowledge and the handling of hidden variables. The conditional probability distribution of future states within a DBN structure implies the Markov property which states that future events are independent of past events given the present.

As discussed in CHAPTER 5, DBNs, an extension of BNs, encode the joint probability distributions over a set of random variables $X = \{x_1,\ x_2, \dots,\ x_n\}$. To define a DBN, the graph structure and the conditional probability distributions at each node should be computed. Learning the structure of a DBN corresponds to the specification of the intra-slice and the inter-slice topologies. It should be noted that the Gaussian distribution is the most commonly used approach for defining the probability distribution of a node given its parents (with mixture of Gaussians to approximate other continuous distributions being another possible choice). There are two different approaches to structure learning: (i) constraint-based, and (ii) search-and-score [178]. In the constraint-based approach, we start with a fully connected graph, and remove edges if certain conditional independencies are measured in the data. This approach has the disadvantage that repeated independence tests lose statistical power. In the more popular search-and-score approach, we perform a search through the space of possible directed acyclic graphs, and either return the best one found (a point estimate) or return a sample of the models found. Since the number of directed acyclic graphs is super-exponential in the number of nodes, we cannot exhaustively search the space, so we either use a local search algorithm (e.g., hill climbing) or a global search algorithm (e.g., Markov Chain Monte Carlo) [273]. As Figure 5.1 illustrates, a typical DBN structure with two time slices ($t1$ and $t2$) consists of the intra and the inter-slice topology which are represented between the variables $v_{i,j}$, where i is the number of the variable and j the exact time slice.

In the proposed study, DEGs and MRs are the features that were further exploited for the inference of DBN models related to the classification problem under study (i.e. distinguish tumor samples from normal samples). Gene expression values of both DEGs and MRs constituted the training data for our learning algorithm. The classification model was implemented in MATLAB using the BNT toolbox [178] and the Dynamic Bayesian Markov Chain Monte Carlo (DBmcmc) package. Moreover, the CCA [186] was employed to compute the prior knowledge for our DBN modeling approach. This algorithm computes potential interactions among targets (i.e. DEGs) and regulators (i.e. MRs) in terms of weight vectors which maximize the canonical correlation between two genes. We selected this algorithm due to its ability to score potential regulatory relationships in a set of genes. These scores are then employed as prior information for the DBN-based algorithm. Concerning the classification task, we had to discriminate the cancerous from non-cancerous samples. We employed the training dataset (i.e. the DEGs and MRs) to define the structure and the parameters of the DBN models with reference to patients and control subjects, respectively. The parameters are represented as CPD (Conditional Probability Distribution) objects and are specified as conditional linear Gaussian distributions. Leave-one-out cross validation technique was repeated 100 times with independent initializations towards the decrease of bias introduced by the stochastic nature of the DBmcmc algorithm [178]. Therefore, we utilized as much data as possible for the training phase, while the tests sets were kept mutually exclusive covering the entire dataset effectively. Concerning the bias that may be introduced during the training of the DBN models and the variance of the estimated performance that may be high due to the leave-one-out cross validation approach, the selection of both DEGs and MRs was done based on the whole dataset. Hence, the computational complexity was kept low when repeating the whole procedure 100 times.

For each case (i.e. sample) the log-likelihood of both DBNs models (cancerous/non-cancerous) was estimated using the junction tree inference engine [178]. Cases were classified as cancerous when the corresponding DBN model had higher log-likelihood given the specific case evidence, compared to the non-cancerous one (i.e. the case had higher probability to be generated from the cancer probability distribution model). The AUC of the classification method was estimated using the vector of classifier predictions scores (log-likelihood of the two DBN models) given the true class labels (actual patient status).

6.3 Results

6.3.1 Time series differential expression analysis

The list with the most statistically significant genes that have been expressed differentially among the groups selected for each dataset was determined according to the *p-value* and *log2 fold change* cut-offs. Figure 6.1 depicts the volcano plots for each microarray dataset. Genes are represented as circles. Red circles correspond to the DEGs with *p-value* < 0.01, whereas genes with lowest or highest *log2 fold change* and lowest *p-values* at the same time, are depicted as green circles. The dashed green and blue lines indicate the *p-value* and *log2 fold change* thresholds, respectively. For GSE14426 the *log2 fold change* cut-off was set to 1.5, while for GSE37182 and GSE5462 to 0.5 to depict the most informative gene set. The *FDR* cut-off was set to 0.05 to detect the most statistically significant genes.

Based on the literature results of the GSE14426 dataset, two specific DEGs that were identified by the authors as statistically significant genes during the disease progression after treatment, were also present in our gene list extracted by the limma package. sFRP4 and RARβ are the transcripts that have been studied in the literature for their repression and progressive increase, respectively. According to the GSE37182 study, the probes that have

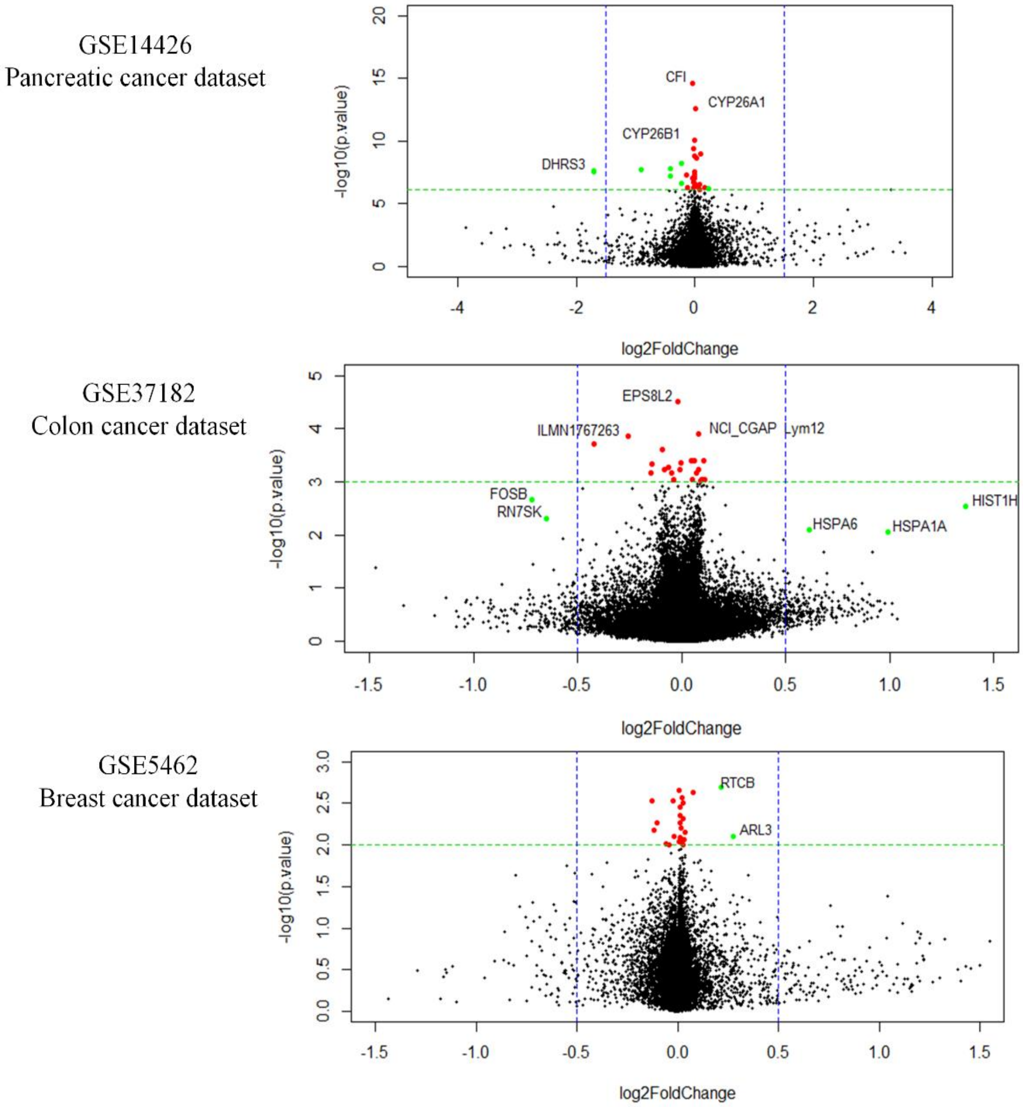


Figure 6.1 Volcano plots for the three microarray dataset considered. Red and green circles indicate the most differentially expressed genes and the genes with the highest/lowest log2 fold change, respectively. Green and blue dashed lines indicate the p-value and log2 fold change thresholds, respectively.

exhibited different means of expression across four time points for both normal and tumor samples are the histones: (i) HIST1H1D, (ii) HIST1H1E, (iii) HIST1H4E, and (iv) HIST4H4. These results are complementary to our findings with the histone protein HIST1H2BF involved in our gene list with changes in its expression during the follow-up. According to the differential expression analysis of the dataset GSE5462, the nine genes that were detected in the literature with significant differential expression are: NUSAP1, KIAA0101, TPBG, ZWINT, MLF1IP, CDC2, CCNB1, HMGB2 (downregulated), and COLEC12. Based on our results the genes that were found as statistically significant were only two: RTCB and ARL3, according to their fold change.

6.3.2 Promoter and pathway analysis of microarray data

Promoter analysis was also performed for potential TFBSs identification in combination with a knowledge-based analysis of the upstream pathway that may control the activity of these transcription factors. The activity of these transcription factors has been shown to lead to hypothetical master regulators. We applied this strategy separately to the gene lists obtained from the previous step (differential expression analysis) in the current pipeline. The significant regulatory molecules were identified by setting the maximal distance of the search for MRs equal to 10 steps upstream of the input DEG list. This selection gives a good chance to find regulators that are quite distant in the network by considering upstream direction from the identified transcription factors. For the DEG set in the pancreatic dataset, we found 219 master regulators. For the colon dataset we identified 164 regulators annotated with reference to the input gene set and for the third dataset (i.e. breast cancer) 200 MRs were selected. To further exploit the large number of MRs annotated for the three gene lists, as well as to avoid overfitting during the predictive modeling approach we employed a relatively small set of MRs for modeling our data. The output tables were sorted according to the *Ranks Sum* values. This metric suggests molecules with a balance between their well-studied status and high connectivity to the selected profile (reflected by *Score*). The computed *Z-Score* reflects the molecule’s novelty and specificity concerning the input DEGs set. The top-ranking master regulators obtained from the upstream analysis for the genes considered in each dataset are provided in Appendix II. Particularly, the regulators with the highest-ranking scores for datasets GSE14426, GSE37182 and GSE5462 were: (i) cyclin B1 (CCNB1), (ii) RSK2 and (iii) NR1, respectively. Figure 6.2 depicts the identified MRs for the lists of up- and down regulated genes in these three microarray datasets. The MRs are depicted at the top-most position of the schematic overview and illustrated in the pink rectangle. They connect molecules of up to 10 steps upstream which are represented with green rectangles, starting from the identified by the geneXplain platform transcription factors sites (blue rectangles). Known complexes are also highlighted in Figure 6.2 (a)-(c), by the dark-green hexagonal frames. According to the STRING database that contains data on functional interaction networks of proteins [274], cyclin B1 (CCNB1) is essential for the control of the cell cycle at the G2/M (mitosis) transition and key predicted functional partners are cyclin-dependent kinases, cyclin A2, CDC28 protein kinase regulatory subunit

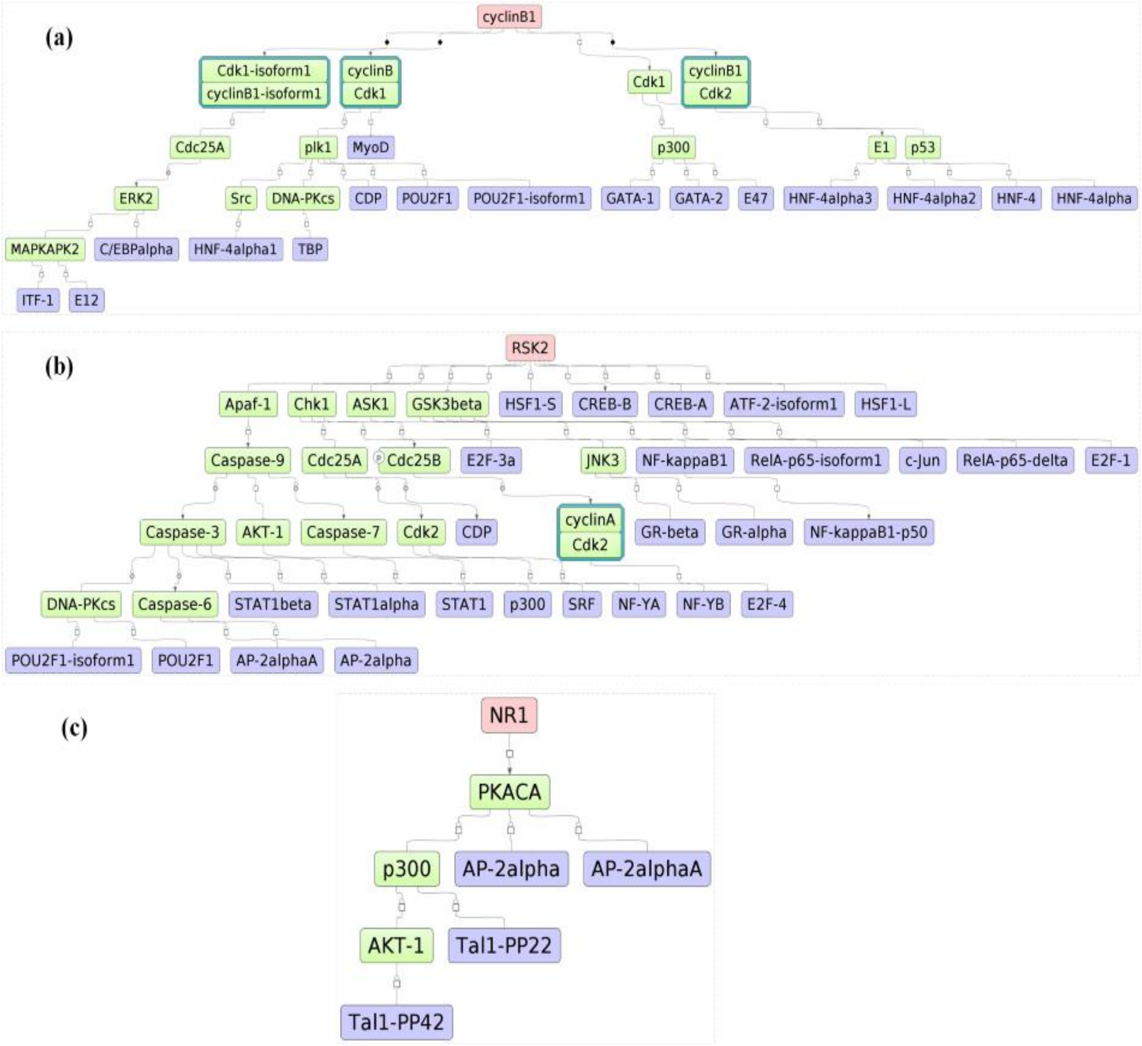


Figure 6.2 (a) cyclin B1 (CCNB1), (b) RSK2 and (c) NR1 diagrams as presented by the geneXplain platform. Known complexes and transcription factors are illustrated according to the identified connections with the master regulators. The master regulators are depicted in the pink rectangle. They connect molecules of up to 10 steps upstream (green rectangles), starting from the transcription factors sites (blue rectangles) identified. Known complexes are also highlighted dark-green hexagonal frames.

2 and polo-like kinase 1. The ribosomal protein S6 kinase (RSK2) mediates mitogenic activation of transcription factors, such as: (i) CREB1, (ii) ETV1/ER81, and (iii) NR4A1/NUR77 and operates downstream of ERK (MAPK1/ERK2 and MAPK3/ERK1) signaling pathway. The predicted functional partners include the cAMP responsive element binding protein 1, the Mitogen-activated protein kinase 1, the Tuberous sclerosis 1 and 2 among others. For the third identified master regulator, i.e. NR1, nucleotide binding proteins 1 and 2, BRCA1 interacting protein C-terminal helicase 1 and Regulator of telomere elongation helicase 1, are among the predicted functional partners according to the protein-protein interactions network of STRING [274]. Further investigation of the connected molecules will improve our understanding concerning the transcriptional pathways underlying each tumor development.

6.3.3 Classification with Dynamic Bayesian Networks

The identified MRs were further exploited by the DBN-based methodology in order to classify tumor samples. Leave-one-out cross validation was employed to evaluate the performance of the DBN models [142]. The conditional probability tables for all nodes in the intra and inter slices were computed for each DBN model. Based on the models' performance, we were able to classify as correct more than 90% of the test samples in one of the three datasets (GSE37182), while both accuracy and AUC metrics were high for all the three cancer datasets. Additionally, we used three subsets in each dataset to identify the gene list that yields the highest classification accuracy. Specifically, to avoid overfitting due to the high number of MRs identified and the small sized datasets considered [143], we compared the classification accuracy of the first 10, 15 and 20 MRs extracted from each dataset. Keeping the MRs list to a minimum, as in the case of DEGs, should be a requirement for our classification algorithm since it suffers from computational complexity for large number of input genes. Therefore, this step would preclude its potential success when the number of significant MRs is high.

Table 6.1 The classification results according to the DBN-based methodology in terms of DEGs, MRs and DEGs with MRs sets. Leave-one-out cross validation was used to evaluate the model's performance.

Input Set	**GEO Dataset**					
	GSE14426		*GSE37182*		*GSE5462*	
	Accuracy	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
DEGs	*65.00%*	*0.700*	*97.14%*	*0.994*	*65.76%*	*0.458*
MRs	*55.00%*	*0.516*	*97.85%*	*0.984*	*61.53%*	*0.499*
DEGs with MRs	*73.33%*	*0.822*	*98,57%*	*0.985*	*70,77%*	*0.562*

We found that in the case of the first 20 highly ranked regulators, highest classification accuracies can be achieved. In addition, we performed classification experiments based on: (i) the gene list identified during the differential expression analysis of each dataset, and (ii) the combination of DEGs along with their identified MRs. Table 6.1 shows the classification results (accuracy and AUC) with the indicative gene sets that could correctly classify the cancerous samples within the three microarray datasets. When the DEG sets and the MR sets were employed separately in our DBN algorithm lower classification results were retrieved. Figure 6.3 illustrates the ROC curves obtained using the same three subsets of each dataset. From Table 6.1 and Figure 6.3, we can conclude that the DBN-based algorithm achieved the highest classification accuracy when the list with the DEGs was enriched with the identified MRs. Leave-one-out cross validation technique was repeated 100 times to evaluate the proposed classification method. The overall results obtained for each case (i.e. (i) classification results with the DEGs, (ii) classification results with the MRs and (iii) classification results for both DEGs and MRs concerning the 3 MR subsets) for each microarray dataset are provided in Appendix III.

6.4 Discussion

In personalized medicine genomic profiling is becoming ubiquitous to elucidate the underlying molecular processes of cancer onset and progression. Several studies contributed to the identification of gene patterns and revealed the associations of gene expression differences for cancer classification. In the current study, we carried out cancer classification

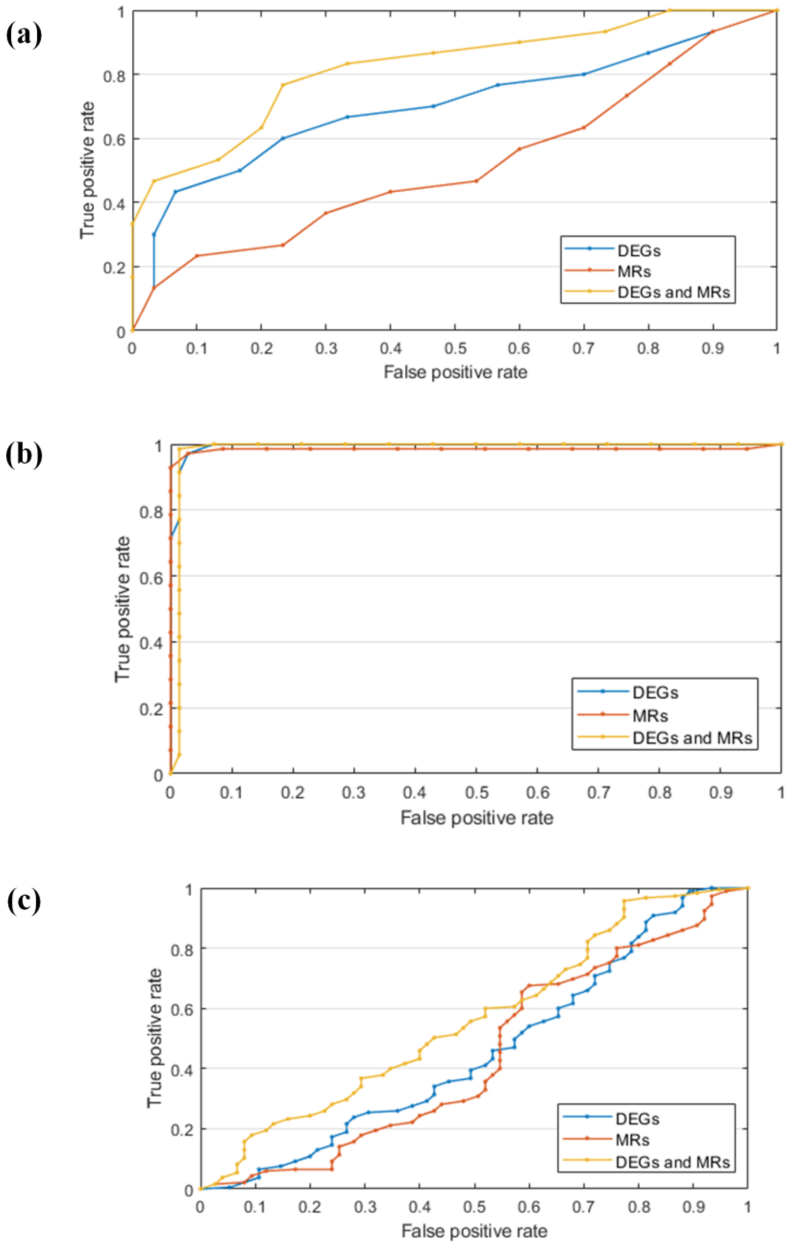


Figure 6.3 ROC curves for the classification performance of the DBN-based algorithm for datasets (a) GSE14426 (b) GSE37182 and (c) GSE5462, respectively. Three cases are considered, namely: (i) DEGs, (ii) MRs and (iii) DEGs and MRs.

on three different time series gene expression profiling research works by utilizing DBNs on both tumor and control samples. Based on the proposed methodology and on our results, the integration of different data types, from the pathway level (i.e. DEGs and MRs), into a DBN-based classification scheme could enhance the discrimination of cancer and non-cancer samples. We found out that, among the DEGs and their potential regulatory molecules, the combination of both associated data sources yields better classification results. Thus, our computational approach reveals the importance of integrating data from the pathway level for microarray-based cancer classification. By utilizing the information from transcription factors and their MRs, we improved the classification accuracy of cancer samples. Based on our results, the current study could be considered as a complementary work in terms of time series gene expression analysis and cancer classification through DBNs. We correctly classified tumor samples and distinguished them from control samples with high accuracy and AUC. Remarkably, for the GSE37182 microarray dataset we were able to classify more than 90% of the test set samples with AUC = 0.985. For GSE14426 and GSE5462 datasets the classification results yielded accuracies 73.3% and 70.8% and AUCs 0.822 and 0.562, respectively. We further compared the proposed classification DBN algorithm with state-of-the-art classification methods previously reported in the literature in terms of the top DEGs in each dataset. The follow-up measurements were considered for each DEGs list. We used the significant DEGs to compare the performance of the proposed methodology against well-known classification algorithms. Naïve Bayes (NB), SVM (polynomial kernel), RF and AdaBoost algorithms were implemented. The accuracies of the NB algorithm for the GSE14426, GSE37182 and GSE5462 datasets were 66.67%, 96.42% and 66.70%, respectively (Table 6.2). The performances of the SVM classification algorithm were slightly lower than NB with accuracies 62.50%, 94.64%, and 51.66%, in each case (Table 6.2). For RF and AdaBoost algorithms the results were slightly higher than NB and SVM for the GSE14426 and GSE37182 datasets. It should be noted that leave-one-out cross validation was also employed in these experiments to evaluate the performance of the classification algorithms and further compare their results with those of the DBN model. Here, we used only the DEGs as input to the four classifiers since the significant genes (i.e. DEGs) have been commonly used in most approaches for cancer classification.

In the current study, the novelty consists of the integration of both DEGs and MRs for discriminating cancerous from non-cancerous samples. Moreover, we should recall the inherent ability of DBNs to model time series microarray data which reveals their potential usefulness in modeling gene expression data.

Additionally, we performed experiments by utilizing both DEGs and MRs subsets within the competing algorithms (NB, SVM, RF and AdaBoost) for comparison reasons (Table 6.3). The results are slightly different from those presented in Table 6.2, with regards to the use of the input data. Although direct comparison with our method might be misleading, due to the static algorithms used in Table 6.2 and Table 6.3, we should note that slightly different results were achieved by the selected algorithms for the GSE37182 and GSE14426 datasets.

The identification of genes with differential expression which reflects different clinical profiles has been studied for the elucidation of the underlying biological mechanisms and disease molecular key pathways. Further exploitation of this type of information through computational approaches could provide the means for the classification of tumor samples. Towards this direction, the detection of genes which have been up- or down-regulated among individuals has been proven valuable for therapeutic protocols in personalized precision medicine [275]. We performed microarray differential expression analysis in three microarray studies pertaining to the identification of the most informative genes related to the diagnosis, prognosis and treatment of different cancer types. Based on our results, we found that our extracted lists with the most DEGs in each dataset were complementary to the ones identified in the literature. For the pancreatic dataset (GSE14426) we extracted the sFRP4 and RARβ genes with the lowest *p-values* and a *log2 fold change* > 1.5. These genes were also identified in the original study [276]. Differential expression analysis results of GSE37182 and GSE5462 showed that the identified significant genes in the current study, were also presented in the literature as differentially expressed in colon and breast cancer samples. Hence, our results are consistent with those from the respective studies. However, direct comparison could not be achieved due to the different parameters and methods utilized for the time course differential expression analysis. Concerning the new findings of the current analysis, we found that CFI (in GSE14426 dataset), EPS8L2 (in GSE37182 dataset)

Table 6.2 Comparison results between the proposed DBN-based classification method and the NB, SVM, RF and AdaBoost classifiers. For comparison reasons, only the significant genes were exploited by the classifiers. Leave-one-out cross validation was employed for performance evaluation.

GEO Dataset	**NB**		**SVM**		**RF**		**AdaBoost**		**Proposed method**	
	Accuracy	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
GSE14426	*66.67%*	*0.833*	*62.50%*	*0.625*	*70.83%*	*0.813*	*66.67%*	*0.785*	*73.33%*	*0.822*
GSE37182	*96.42%*	*0.990*	*94.64%*	*0.946*	*96.42%*	*0.929*	*98.21%*	*0.964*	*98,57%*	*0.985*
GSE5462	*66.70%*	*0.682*	*51.66%*	*0.517*	*53.34%*	*0.586*	*58.33%*	*0.608*	*70,77%*	*0.562*

and RTCB (in GSE5462 dataset) genes were identified with the lowest *p-values* as most statistically significant; thus, further analysis of these candidate genes may contribute to the elucidation of complex molecular processes underlying cancer progression.

Concerning the upstream analysis in the proposed workflow, we identified the factors that trigger the expression of a gene and may be responsible for its transcription. Towards this direction, better understanding about their interactions and the biological networks that these factors might be involved in, was achieved. We found the TFBSs that are enriched in the promoters of the DEGs. This knowledge was then exploited to search for regulatory molecules in signal transduction pathways upstream of the identified transcription factors. Our results are in accordance with those of other studies exploring genetic variants that may regulate gene activity and may influence transcriptional processes and mechanisms during cancer development.

Specifically, the most highly ranked MRs identified for each one of the pancreatic, colon and breast cancer datasets were: (i) cyclin B1, (ii) RSK2 and (iii) NR1, respectively. According to the biomedical and genomic information from the NCBI and the Online Mendelian Inheritance in Man (OMIM) database of human genes and genetic disorders, cyclin B1 (CCNB1) is a gene that encodes a regulatory protein involved in mitosis and it shows a broad expression in lymph nodes, testis and 21 other tissues (including pancreatic tissue). Its prognostic role has been studied extensively in solid tumors revealing the potential role to the invasiveness of adenomas. The ribosomal protein S6 kinase A3 (RSK2) encodes a member of the RSK (ribosomal S6 kinase) family of serine/threonine kinases. It consists of 2 non-identical kinase catalytic domains while phosphorylates a variety of

Table 6.3 Results obtained using the NB, SVM, RF and AdaBoost classifiers. Both the significant genes and the MRs were exploited by the classifiers. Leave-one-out cross validation was employed for performance evaluation

GEO Dataset	**NB**		**SVM**		**RF**		**AdaBoost**	
	Accuracy	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
GSE14426	*66.67%*	*0.882*	*79.16%*	*0.792*	*79.16%*	*0.875*	*79.16%*	*0.882*
GSE37182	*99.98%*	*0.1*	*100%*	*0.998*	*98.21%*	*0.1*	*98.21%*	*0.968*
GSE5462	*60.00%*	*0.672*	*50.00%*	*0.500*	*58.34%*	*0.568*	*68.34%*	*0.734*

substrates. The activity of this protein has been implicated in controlling cell growth and differentiation. It shows a ubiquitous expression in colonic and 25 other tissues. The third regulatory molecule identified from the microarray breast cancer dataset, the NR1 gene, has been found to encode an NADPH-dependent diflavin reductase. It has been determined that it is highly expressed in a panel of human cancer cell lines derived from ovary, breast, bladder, lung, colon, liver, and cervical carcinoma tissues.

Subsequently, we utilized the results from the classification algorithm to accurately distinguish patient samples into distinct groups (i.e. control and cancerous). The proposed DBN-based approach was developed in terms of the CCA algorithm to infer the potential regulatory interactions among a set of (i) DEGs, (ii) MRs and (iii) DEGs along with their MRs. This data was employed as prior knowledge for the development of the DBN models and for the inference of gene regulatory networks, as well. Our experiments showed that this methodology can accurately classify patients into distinct groups with high AUC. We found that the integration of both data sources (i.e. DEGs and their MRs) can yield better classification results (Table 6.1) in comparison to the results obtained when the data sources are employed separately. This is complementary to other published works [262], where the importance of deriving information from the pathway level is considered for disease classification. Our results extend the findings of previous studies related to cancer classification by utilizing the significant genes and their regulatory molecules. This knowledge can be employed by researchers in healthcare to better understand the underlying disease complex regulatory mechanisms, while the presented methodology could be further evaluated in the patient care in terms of classification of new patient samples.

In the present study, transcriptomics datasets were exploited for the identification of DEGs and their master regulatory molecules. However, DNA methylation and copy number alteration are also well-known regulators especially in cancer datasets. Based on this knowledge, in a future study we plan to exploit additional datasets to perform a more targeted upstream analysis. Moreover, the exploitation of data generated from next-generation sequencing (NGS) technologies could empower our findings related to the identification of gene expression patterns [68, 277] and to the classification problem addressed. The acquisition of count data from RNA-Seq analysis in oncology provides comprehensive knowledge about the high percentage of actionable mutations in cancer datasets. Hence, the information extracted from the sequence level could be translated into better clinical outcomes and personalized treatments. The proposed DBN-based methodology could be modified accordingly to increase the classification accuracy by adjusting the algorithm's parameters to new types of data. This would allow the extraction of more precise and robust results for the clinical practice.

6.5 Conclusions

We herein employed time series microarray gene expression data to identify DEGs with their potential MRs for cancer classification. Promoter and pathway analysis of the important genes revealed statistically significant regulatory molecules that may contribute to the transcription mechanisms of tumor development. We proposed a DBN-based approach which can model time series gene expression data for classification purposes. We classified more than 90% of test set samples from the GSE37182. Likewise, satisfactory results were achieved when the GSE14426 and GSE5462 datasets were utilized, which demonstrates that DBN-based models can accurately classify patient samples in terms of the computed interactions (conditional probabilities) between genes and their potential regulatory factors. Hence, the application of the current knowledge (meaningful and generalized cancer classification model) to the healthcare and cancer treatment domains could leverage predictive modeling and clinician's interpretability.

CHAPTER 7 LYMPHOMA DEVELOPMENT RISK PREDICTION THROUGH ENSEMBLE MACHINE LEARNING-BASED PIPELINE

7.1 Introduction

Sjögren’s syndrome (SS) is a chronic autoimmune disorder mainly manifested with symptoms denoting dryness of the internal mucosae as a result of exocrine gland involvement. Though SS is traditionally considered a disease of dryness, systemic features affecting internal organs commonly occur with lymphoma development being a major complication [278].

Over the last decades a large amount of data revealed several clinical (salivary gland enlargement, purpura, Raynaud, tongue atrophy), serological (RF, Ro/La autoantibodies, monoclonal gammopathy, low complement C4, serum BAFF) and histopathological features (extensive lymphocytic infiltration), as predictors for lymphoma development in the context of Sjögren’s syndrome. Of interest, these adverse risk factors are usually present early at disease onset implying that a distinct genetic background could rely behind the aggressive behavior towards lymphoma development [279].

On this basis, genetic variants of genes implicated in the regulation of chronic inflammation such as TNFAIP3 and LILRA3, B cell activation, type I IFN pathways like TREX-1 or epigenetic processes have been shown to increase SS related Non-Hodgkin Lymphoma (SS NHL) susceptibility especially when the disease onset starts before 40 years old, as evidenced by the higher frequencies of the BAFF-R, TNFAIP3 and LILRA3 variants in the young onset group.

Identifying novel biomarkers for elucidating the risk for lymphoma development still remains a clinical unmet need in SS. Lymphoma prediction based on clinical and biological predictors have been studied widely in the literature in terms of statistical analysis and prediction rules [280, 281]. Towards this direction, in [280] a predictive tool in clinical practice has been developed for SS-related lymphoma development. Based on the initial clinical, laboratory and histopathological variables of SS patients the probability score of lymphoma development reached 100% when all 7 risk factors were considered (i.e. salivary gland enlargement, lymphadenopathy, Reynaud phenomenon, anti-Ro/SSA or/and anti-La/SSB as well as RF positivity, monoclonal gammopathy and C4 hypocomplementemia). Data mining algorithms have been also exploited for the identification of patient subgroups and the prediction of lymphoma. The associations among patient's demographics, clinical and serological variables have been defined and a prediction model based on Artificial Neural Networks (ANNs) has been developed able to predict new unseen records with high sensitivity and specificity.

In this study, we aim at identifying the contribution of combined initial clinical, serological and histopathological features with genetic variants in predicting lymphoma development using a ML-based methodology with ensemble classifiers. We focused on the development of this methodology since it can classify accurately new patients according not only to their traditional clinical measurements but also to their genetic susceptibility as a critical factor that predispose SS complications, such as lymphoma. The proposed methodology is based on the GB and Random RF ensemble classifiers for developing the predictive models which are characterized by the ability to generalize their decision boundaries to regions where there are no available training examples. This type of classifiers was selected in terms of the variance and bias estimation which contribute to the expected error of a classification model. The novelty of the proposed ML-based methodology pertaining to the potential usefulness of genetics in predicting lymphoma development in SS patients. The classification results reported in our study are obtained from stratified 10-fold cross validation with the ensemble classifiers outperforming the single LR approach and the SVM classifier.

In the following sections, we first introduce the study cohort and the preprocessing steps followed towards the development of accurate predictive models. Next, the proposed ML-based methodology in terms of ensemble classifiers is presented. The formulation of the learning problem is given along with the background information on GB and RF estimators. Based on our results, we demonstrate that the combination of clinical phenotypes with genetic variants in SS could further improve the prediction performance of the ML models. We anticipate that the current work could provide new insights into the aggressive behavior of lymphoma development in SS patients.

7.2 Materials and Methods

7.2.1 Study cohort

Medical records of 143 primary SS patients (SS) without and 64 SS patients with a history or a current diagnosis of concomitant B-cell Non-Hodgkin lymphoma (SS NHL), fulfilling the revised European/American International classification criteria for SS, were collected (Table 7.1). DNA derived from whole peripheral blood of 207 patients with primary SS fulfilling the same classification criteria. The patients were genotyped for 13 single nucleotide polymorphisms (Table 7.2) which were subsequently extracted and stored at -20^{O}C upon use at the Department of Physiology, National and Kapodistrian University of Athens, Greece. Methods of DNA extraction and genotyping protocols are described in more details in [280]. Demographic, clinical and laboratory features were recorded after thorough chart review. Lymphoma diagnosis in the pSS-lymphoma group was based on the criteria outlined by the World Health Organization classification. This study was carried out in accordance with the recommendations of the Ethics Committee of the National and Kapodistrian University of Athens (approved No. 6337) with written informed consent from all subjects following the Declaration of Helsinki.

Information regarding the presence of glandular manifestations such as salivary gland enlargement was obtained. Systemic features such as Raynaud's phenomenon and lymphadenopathy were also recorded. Laboratory data included serological characteristics such as monoclonal gammopathy, autoantibodies (antinuclear antibodies, anti-Ro/SSA, anti-La/SSB antibodies, rheumatoid factor [RF], antimitochondrial, and anti-thyroid), cryoglobulins, and C3 protein levels. At the level of Minor Salivary Gland (MSG) tissue, germinal center formation and the presence of monoclonality (as described in [280]) were also recorded. Demographic and clinical characteristics such as: (i) gender, (ii) year of birth, (iii) year of disease diagnosis, (iv) age at SS diagnosis, (v) ≥ 40 age at SS diagnosis, (vi) lymphoma development and (vii) lymphoma type were also considered.

7.2.2 Data preprocessing and curation

Data preprocessing was performed by utilizing an automated framework for evaluating the data quality [282]. The main steps followed towards the dataset quality assessment are referred to: (i) the detection of missing values in an autonomous way, (ii) the detection and removal of outliers, and (iii) the identification of duplicate values and highly correlated distributions among variables.

The data curation framework enhanced our assessment related to the types of variables included in the raw dataset and their quality in terms of missing and duplicate measurements and outlier's detection. In the case of missing values handling, we excluded subsequently the clinical records that exhibited a percentage of missing values higher than 90%. The variables within the dataset that also exhibited a percentage of missing values higher than 80% were not selected for further analysis. In terms of outliers and duplicate values detection, the respective features were not considered in the cleaned dataset. To complete any missing value detected within the dataset after the data preprocessing and curation step, an imputation transformer was also developed in Python by utilizing the Sci-kit learn library [283]. The "SimpleImputer" transformer was adopted with strategy "mean" for replacing missing values using the mean along the continuous variables, and "most_frequent" for each categorical variable. After applying the preprocessing procedure, we concluded with 207 patient records consisting of 23 clinical, laboratory, demographic and genetic variables.

Table 7.1 and Table 7.2 present the four different categories of the preprocessed dataset exploited in the current study. We have removed the duplicate features, the patient records with high percentage of missing values and the features detected with outliers, as described above. The samples distribution of the categorical features within each class (class 0 = no lymphoma development; class 1 = lymphoma development) was described by the corresponding percentages, whereas for the continuous variables was described by their arithmetic mean and standard deviation values (SD) as well as their minimum and maximum values.

Table 7.1 The variables of the initial demographic, clinical and laboratory findings related to the patients' samples considered in the current study. The mean±SD values and the min/max values were calculated for continuous variables. The respective percentages were also calculated for the discrete variables. These values were computed for both classes (i.e. class 0 = no lymphoma development; class 1 = lymphoma development). The undefined percentages for categorical variables are also given.

Category	Variable	Class 0			Class 1		
		mean	SD	min / max	mean	SD	min / max
Demographic	Age at SS diagnosis (years)	50.93	13.38	15 / 74	50.67	14.27	24 / 81
Category	Variable	Class 0 (%)			Class 1 (%)		
		False	True	Undefined	False	True	Undefined
Clinical features	Salivary Grand Enlargement (SGE)	78.32	20.97	0.71	31.25	67.18	1.57
	Raynaud phenomenon	78.35	21.65	0.00	59.37	40.63	0.00
	Lymphadenopathy	86.00	14.00	0.00	54.68	45.32	0.00
	≥ 40 age at SS diagnosis	82.52	17.48	0.00	25.00	75.00	0.00
Laboratory characteristics	Monoclonal gammopathy	89.51	6.29	4.20	71.87	25.00	3.13
	Anti-Ro/SSA or/and anti-La/SSB positivity	74.12	24.47	1.41	11.00	89.00	0.00
	RF positivity	50.34	41.25	8.41	15.62	82.81	1.57
	Low C4	53.14	45.45	1.41	20.31	76.56	3.13

Table 7.2 The genetic variants (genes (IDs) and reference numbers (rs#)) related to the patients' samples considered in the current study. The percentages for common genotype (0), heterozygous (1), homozygous (2) and undefined SNPs within both classes are presented.

Category	GENE/ ID	rs #	Class 0 (%)				Class 1 (%)			
			0	1	2	Undefined	0	1	2	Undefined
Genetic variant	MTHFR/ 4524	rs1801133	39.80	46.20	14.00	0.00	39.06	2.18	8.76	0.00
	MTHFR/ 4524	rs1801131	55.64	32.16	11.20	0.00	53.12	5.93	10.95	0.00
	TNFRSF13C (BAFF Receptor)/ 115650	rs61756766	93.70	0.30	0.00	0.00	90.60	0.40	0.00	0.00
	TNFSF13B (BAFF)/ 10673	rs1224141	71.32	26.57	1.40	0.71	60.93	35.93	0.00	3.14
	TNFSF13B (BAFF)/ 10673	rs12583006	51.04	35.66	13.30	0.00	53.12	39.06	4.70	3.12
	TNFSF13B (BAFF)/ 10673	rs9514828	15.39	56.64	27.97	0.00	10.93	56.25	29.68	3.14
	TNFSF13B (BAFF)/ 10673	rs1041569	52.44	44.05	3.51	0.00	51.56	39.06	6.25	3.13
	TNFSF13B (BAFF)/ 10673	rs9514827	44.05	44.75	11.20	0.00	35.93	48.43	10.93	4.71
	TREX1/ 11277	rs11797	35.66	44.75	18.90	0.69	34.37	40.62	23.43	1.58

	TREX1/ 11277	rs3135941	69.23	23.07	7.00	0.70	78.12	17.18	3.12	1.58
	TNFAIP3/ 7128	rs2230926	89.51	9.80	0.00	0.69	93.75	6.25	0.00	0.00
	PTPN22/ 26191	rs2476601	89.51	10.49	0.00	0.00	89.06	10.94	0.00	0.00
	LILRA3/11026	deletion	79.00	14.70	0.00	6.30	87.50	6.25	1.56	4.69

7.2.3 Problem formulation

In a predictive learning problem, lymphoma development, L, $L \in \{L_0,\ L_1\}$, can be estimated as a non-linear parameterized function, F, of a confined set of features $x \in R^d$ such that $F(x) = L$, and $x = [x_1, \dots, x_d]$. We consider a set of training samples $Z = \{(y_i, x_i,)_{i=1}^N\}$ of known (y, x) values, where x corresponds to the random "input" or "explanatory" features and y to the "output" or the "response" variable. Each sample (y_i, x_i) associates the input vector $x_i \in R^d$ of each patient P_i with the actual classification of his/her lymphoma development status assessed by the clinicians. The main objective is to obtain an estimate or approximation of the function $F^*(x)$ mapping x to y, which minimizes the expected value of a loss function $L(y, F(x))$ over the joint distribution of all (y, x) values :

$$F^* = \underset{F}{\operatorname{argmin}} E_{y,x}\, L\big(y, F(x)\big) = \arg\min_F E_x\big[E_y\big(L(y, F(x))\big)|x\,\big]. \tag{7.1}$$

Commonly, a procedure to restrict $F(x)$ is the function to be a member of a parameterized class of functions $F(x; P)$, where $P = \{P_1, P_2, \dots\}$ is the set of parameters whose joint values identify individual class members. We consider "additive" expansions in the form:

$$F(x; \{\beta_m, a_m\}_1^M) = \textstyle\sum_{m=1}^M \beta_m h(x; a_m), \tag{7.2}$$

where $h(x;\ a)$ is a generic parameterized function of the input variables x characterized by parameters $a = \{a_1, a_2, \dots\}$ and $m = 1, \dots, M$, which denotes the m^{th} adaptive (parameterized) simple function (namely the base learner).

7.2.4 Cost-sensitive Random Forest Feature Selection and Ranking

The RF classifier was applied aiming at evaluating the importance of features with reference to the classification problem. The "balanced mode" of the RF estimator was selected in the current study to automatically adjust weights associated with the class frequencies in the training set. The identification of the most important predictor variables which contribute to accurate and unbiased predictions of the response variable was achieved. The maximum number of features selected after keeping the threshold disabled (i.e. threshold = $-\infty$) was also reported with reference to the feature ranking results.

According to the binary classification and regression trees and RFs background [284], the best split $s_t = s^*$ at each node t maximizes the decrease:

$$\Delta_i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R), \tag{7.3}$$

of some impurity measure i_t (e.g. the Gini index, the Shannon entropy or the variance of the response variable Y). t_L and t_R denote the left and right child nodes of node t following split s_t, respectively. The RF algorithm instead of searching for the best split at each node, selects a random subset of K variables and determines subsequently the best split over these features only. To evaluate the importance of a variable X when considering aggregation of randomized trees for predicting response Y, the weighted impurity decrease is used over all N_T trees in the forest [284]:

$$Imp(X_m) = \frac{1}{N_t} \sum_T \sum_{t \in T: v(s_t) = X_m} p(t) \Delta i(s_t, t), \tag{7.4}$$

where p_t is the proportion $\frac{N_t}{N}$ of training samples reaching node t and $v(s_t)$ is the feature used in split s_t. Equation (7.4) constitutes the Mean Decrease Impurity (MDI) importance. By sorting the importance scores $imp(X_j)$ in descending order, a ranked list of variables, $R = \left[X_{j'_1,\ldots,} X_{j'_d}\right]$ is obtained, where $J' = [j'_1, \ldots, j'_d]$, $j'_j \in [1, \ldots, d]$ and $imp(X_{j'_j}) \geq imp(X_{j'_{j+1}})$.

7.2.5 *Model training and parameter tuning with ensemble classifiers*

Ensemble methods enhance the classification accuracy by aggregating the predictions of multiple base classifiers [142]. During a classification task with ensemble methods a set of base classifiers is developed from the training data and the performance of the classification model is evaluated by voting on the individual predictions made by each classifier. The rationale for ensemble methods is that the error rate during a classifier's performance is considerably lower than the error rate of the base classifiers, considering that the base classifiers are not identical but independent.

Let D denote the original training data and T be the test set. A training set D_i is created from D, which size is kept identical with the original data while the distribution of records may be different. A base learner C_i is built from D_i, for $i = 1, \ldots, k$, which denotes the number of base classifiers. For each test record $x \in T$ to be classified, the predictions made by each base classifier $C_i(x)$ are then aggregated by taking a majority vote on the individual base learners predictions in order to obtain the class $C^*(x)$:

$$C^*(x) = Vote(C_1(x), C_2(x), \ldots, C_k(x)). \tag{7.5}$$

Ensemble methods achieve better classification results with unstable classifiers which are sensitive to minor perturbations in the training phase. Examples of such classifiers are the decision trees, the rule-based classifiers and the artificial neural networks. The proposed ML-based methodology enables the minimization of errors related to the variability of the training samples due to the utilization of ensemble algorithms. The bias-variance decomposition method is usually applied for the analysis of such types of errors concerning the predictions of a classification model [142]. In the current study, the GB and RF ensemble classifiers are considered and further implemented based on imbalanced datasets towards the development of predictive models with high generalization ability and less training errors.

7.2.5.1 Gradient Boosting (GB) classification model

Boosting is a known example of ensemble methods that manipulates the training samples for improving classification accuracy. The overall classification accuracy is obtained by aggregating the predictions of multiple base learners [285]. Boosting methods assign a weight to each training sample and at the end of each boosting round they may adaptively change the weight. GB for classification approximates a simple parameterized function and fits regression tree(s) sequentially on the negative gradient of the specified loss function. Furthermore, GB incorporates randomness into the function estimation procedures for improving the performance. Based on this knowledge, GB constitutes an appealing choice for solving classification problems and building predictive models from an input dataset.

According to (7.2), the parameters $\{a_m\}_0^M$ and the expansion coefficients $\{\beta_m\}_0^M$ are jointly fit to the training data in a stage-wise manner. Based on an initial guess $F_0(x)$, for each $m = 1,2, \ldots, M$ we have:

$$(\beta_m, a_m) = \arg\min_{\beta,a} \textstyle\sum_{i-1}^{N} L(y_i, F_{m-1}(x_i) + \beta h(x_i; a), \tag{7.6}$$

and

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m). \tag{7.7}$$

According to [285], GB solves equation (7.6) for differentiable loss function $L(y, F(x))$ with a two-step procedure. In the first step, the function $h(x;\ a)$ is fit to the current "pseudo"-residuals by least squares:

$$a_m = \arg\min_{a,\rho} \sum_{i=1}^{N} [y_{im} - \rho h(x_i; a)]^2, \tag{7.8}$$

where y_{im} is the current "pseudo"–residuals. Then, given $h(x; a_m)$, the optimal value of the coefficient β_m is determined based on:

$$\beta_m = \arg\min_{\beta} \sum_{i=1}^{N} L(y_i, F_{m-i}(x_i) + \beta h(x_i; a_m)). \tag{7.9}$$

Gradient tree boosting follows this approach considering that the base learner $h(x;\ a)$ is an L terminal node regression tree. In the current study, the ensemble model selection procedure reveals the efficiency of the gradient tree boosting classifier, which may well reduce model's variance, noise and bias; thus, minimizing model's training error. The proposed ML-based methodology was implemented in Python by utilizing the Sci-kit learn library and the imbalanced-learn toolbox for classification purposes. Gradient tree boosting algorithm was nested into an EasyEnsemble [286] random under-sampling scheme to handle the class imbalance problem in our dataset. The EasyEnsemble approach exploits the samples from the majority class that have been ignored by under-sampling. Given a set of minority class samples P, a set of majority class samples N, where $|P| < |N|$, the number of subsets T to sample from N and the number of iterations to train the gradient boosting ensemble s_i, EasyEnsemble method randomly samples a subset N_i, $|N_i| = |P|$ for learning an ensemble of balanced gradient tree boosting classifiers trained on different balanced bootstrap samples. The output of EasyEnsemble is a single ensemble with s_i gradient boosting classifiers ("ensemble of ensembles") which reduces the bias of model performance while improving generalization of the decision boundary.

7.2.5.2 Random Forest (RF) classification model

RF is a class of ensemble methods which encompasses multiple decision trees using random vectors from the original training data [142, 287]. Predictions made by multiple decision trees are combined via majority voting aiming at improving the classification accuracy. RFs constitute a combination of tree learners (predictors), where each tree is generated according to the values of an independent train set of random vectors. Internal estimates of the generalization error of the combined ensemble trees allow the control and monitoring of error, strength and correlation. Towards this direction, out-of-bag methods have been used aiming at improving internal estimates in relation to the classification accuracy [287]. Internal out-of-bag estimates have also applications to the understanding and measurement of variable importance.

For the k_{th} tree, a random vector Θ_κ is generated which is independent of the past random vectors $\Theta_1, \ldots, \Theta_{k-1}$, but with the same distribution. The training set and Θ_κ are used to construct the trees resulting into a classifier $h(x, \Theta_k)$, with x being the input vector.

According to the definition in [142], RF is a classifier consisting of a number of tree-based classifiers $\{h(x, \Theta_\kappa), k = 1, \ldots\}$, where $\{\Theta_\kappa\}$ are independent random vectors with the same distribution and each tree in the forest obtains a vote for the most popular class given input x. According to [142], it has been proven that the upper bound for the generalization error of RFs converges to the following expression, when the number of trees is sufficiently large:

$$Generalization\ error \leq \frac{\bar{\rho}(1-s^2)}{s^2}, \tag{7.10}$$

where $\bar{\rho}$ is the average correlation between the trees and s is a quantity that expresses the “strength” of the tree classifiers. The strength of a classifier implies its average performance, which is computed probabilistically in terms of the classifier’s margin:

$$margin, M(X,Y) = P\left(\widehat{Y_\theta} = Y\right) - \max_{Z \neq Y} P\left(\widehat{Y_\theta} = Z\right), \tag{7.11}$$

where $\widehat{Y_\theta}$ is the predicted class of X based on the random vector θ that builds the classifier. The higher the margin, the more likely the classifier predicts a new unseen record.

In the present study, the model selection procedure reveals that RF is also an appealing choice for ensemble-based classification purposes in order to learn imbalanced datasets and achieve better decision boundaries for the predictive model. The RF algorithm was nested into a random under-sampling scheme included in the imbalanced-learn toolbox for handling such datasets. Specifically, the balanced RF classifier was implemented which randomly under-samples each bootstrap sample to balance it. Each tree of the forest is provided with a balanced bootstrap sample resulting in an ensemble of samples including inner balancing samplers.

7.2.6 Performance evaluation and validation

To evaluate the classification performance of our proposed methodology six measures including balanced accuracy, which deals with balanced datasets, sensitivity, specificity, positive predictive value, negative predictive value and AUC were used for both GB and RF models.

An external stratified 10-fold cross validation was applied with reference to the feature ranking and the boosting classification scheme in each iteration, allowing for the reduction of the models' variance. The stratified 10-fold cross validator [283] returns stratified training folds by preserving the percentage of samples for each class within the dataset. We should note that an inner 5-fold cross validation was also applied in the proposed ML-based methodology for assessing the exhaustive grid search over specified parameter values for each classifier. Using grid search with a nested cross validation for parameter estimation ensures the optimization of model's parameterization.

7.3 Results

Table 7.3 and Figure 7.1 present the evaluation performance of the GB and RF ensemble classifiers. For the RF classifier both Gini and entropy criteria were applied in order to determine the best way to split the samples. These measures are defined according to the fraction of samples that belong to class i at a given node t. The best split is then selected according to the degree of impurity of the child nodes [142].

Three input cases were considered in the current study for comparison reasons and for assessing the models' performances. More specifically, the clinical phenotype of each patient along with the genetic data were considered (input case 1) for building the proposed predictive models and further evaluate their performance. For assessing the potential of combining the initial SS patient's medical features with genetic variants in predicting lymphoma development, we followed the same procedure for input case 2 (the clinical phenotype for each patient) and input case 3 (the genotyped data acquired for each patient) and evaluated the models' performances in terms of certain metrics and a hyper-parameter optimization criterion (i.e. balanced accuracy). For each prediction model (i.e. RF models and GB model) the mean value of each metric is presented along with the computed standard deviation.

We can observe that the combination of the initial clinical, serological and histopathological features with genetic variants (input case 1) result in the accurate prediction of lymphoma development in SS patients with considerable high balanced accuracy for RF Gini (0.7626±0.1787), RF Entropy (0.7590±0.1837) and GB (0.7780±0.1514) classifiers, respectively (Table 7.3). We should also report for input case 1, the high results obtained with reference to the sensitivity metric implying the high proportion of patients with lymphoma who have been predicted as positive by the classifiers (RF Gini classifier: 0.8000±0.3435, RF Entropy classifier: 0.8000±0.3435 and GB classifier: 0.8309±0.2594) (Table 7.3 and Figure 7.1).

As illustrated in the confusion matrices Figure 7.1, the GB model could predict more subjects as true positives (=104) and true negatives (=53) in comparison to RF Gini and RF models. The mean AUC of the models in terms of the sensitivity and specificity results are 0.7988±0.2186 (RF Gini classifier), 0.7995±0.1917 (RF Entropy classifier) and 0.8054±0.1570 (GB classifier) which constitute promising results for predicting lymphoma development (Figure 7.1). For input case 2 the GB classifier performed better with slightly higher mean AUC (0.8215±0.1534) in comparison to the mean AUC of input case 1 (0.8054±0.1570). The exploitation of only the clinical and laboratory patient records (i.e. input case 2) could be comparable with the combination of both genotyped data and the clinical phenotypes (i.e. input case 3) towards predicting lymphoma development. However, we can observe that the computed sensitivity, positive predictive and negative predictive values of the GB model for input case 2 are notably lower in accordance to the respective evaluation metrics for input case 1.

Concerning the exploitation of individual genetic variants for building the predictive models (input case 3) the results yielded by the proposed methodology are moderate with significantly lower balanced accuracy, sensitivity and specificity in comparison to input cases 1 and 2. Based on this knowledge, we can admit that the combination of both data sources (clinical and genetic profiles) could result in more accurate classification results by obtaining predictive models with reference to ML techniques.

Table 7.3 RF Gini, RF Entropy and GB classification results. Outer stratified 10-fold cross validation for evaluating the models’ performances and inner 5-fold cross validated grid search for hyper-parameter optimization have been performed for obtaining the classification results. Certain evaluation metrics have been computed regarding the models’ performance. The results are presented as mean±2SD. For each input case, balanced accuracy was selected as criterion within the grid search procedure. We highlighted in bold the best results obtained by either input case 1 (clinical and genetic data), input case 2 (clinical data) and input case 3 (genetic data), for each estimator in order to pinpoint the significant findings of the current study.

Classifier	input case #	Hyper-parameter optimization criterion	Balanced Accuracy		Sensitivity		Specificity		Positive Predictive Value		Negative Predictive Value		AUC	
			mean	2SD	mean	2SD	mean	2SD	Mean	2SD	mean	2SD	mean	2SD
RF Gini Ensemble classifier	input case 1	balanced accuracy	**0.7626**	**0.1787**	**0.8000**	**0.3435**	**0.7252**	**0.1702**	**0.5701**	**0.1795**	**0.8974**	**0.1638**	**0.7988**	**0.2186**
	input case 2		0.7626	0.1705	0.800	0.3188	0.7252	0.1577	0.5696	0.1856	0.8964	0.1569	0.8043	0.2019
	input case 3		**0.5350**	**0.1500**	**0.6880**	**0.3045**	**0.3819**	**0.2624**	**0.3352**	**0.1347**	**0.7389**	**0.1587**	**0.4986**	**0.1791**
RF Entropy Ensemble classifier	input case 1	balanced accuracy	**0.7590**	**0.1837**	**0.8000**	**0.3435**	**0.7180**	**0.1989**	**0.5662**	**0.1868**	**0.8960**	**0.1643**	**0.7995**	**0.1917**
	input case 2		07542	0.1907	0.7833	0.3598	0.7252	0.1577	0.5625	0.2006	0.8900	0.1713	0.8107	0.2040
	input case 3		0.5264	0.1823	0.6428	0.2651	0.4100	0.2907	0.3352	0.1722	0.7142	0.1812	0.4987	0.1557
GB Ensemble classifier	input case 1	balanced accuracy	**0.7780**	**0.1514**	**0.8309**	**0.2594**	**0.7252**	**0.2398**	**0.5921**	**0.2095**	**0.9099**	**0.1284**	**0.8054**	**0.1570**
	input case 2		**0.7509**	**0.1701**	**0.7833**	**0.2812**	**0.7185**	**0.1630**	**0.5591**	**0.1859**	**0.8844**	**0.1455**	**0.8215**	**0.1534**
	input case 3		0.5209	0.1792	0.6880	0.3392	0.3538	0.2872	0.3256	0.1441	0.7279	0.2634	0.4618	0.1996

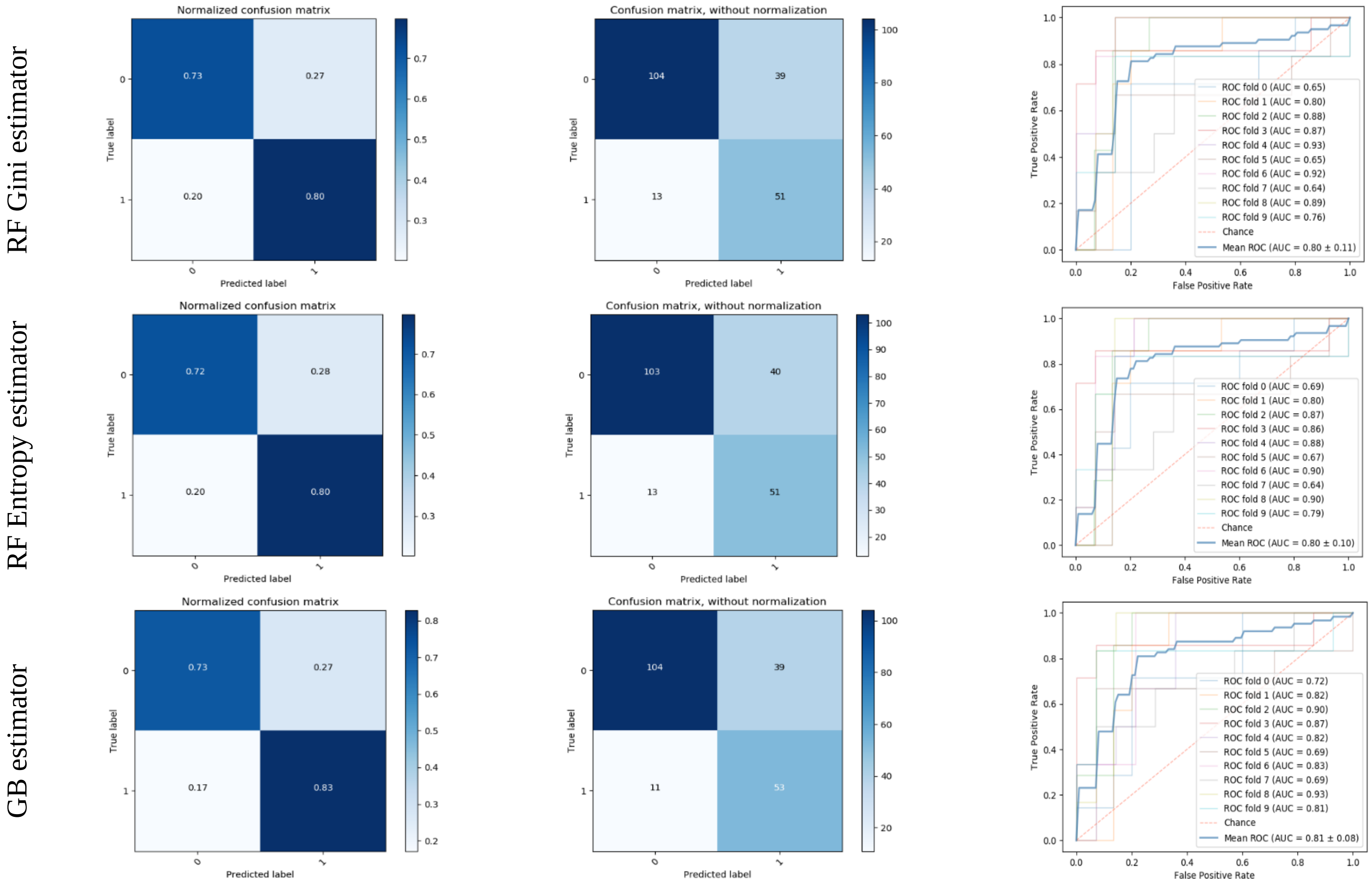


Figure 7.1 The normalized and non-normalized confusion matrices obtained for each classification model. The ROC curves after the evaluation of models' performance are also illustrated. Each row corresponds to the respective classifier's evaluated performance. In the upper side the classification performance of RF Gini estimator is depicted (confusion matrices and ROC curve). In the middle and lower side of the figure the classification results of RF Entropy and GB classifiers are presented, respectively. The ROC curves correspond to the mean ROC curves and AUC after applying the 10-fold cross validation procedure. The ROC curve in each fold is also illustrated for comparison purposes. In addition, the ± 1SD is also given with the mean ROC.

Figure 7.2 illustrates the boxplot with mean feature importances according to the feature selection and ranking procedure performed with RF selector. Hence, the most important features which contribute to accurate and unbiased predictions of lymphoma development were identified. We can observe that the 10 most informative features are SGE, age at SS diagnosis, low C4, lymphadenopathy, RF plus, BAFF snp2, TREX snp1, MTHFR677, TREX snp2 and BAFF snp6. We shall recall that the presented values refer to the mean importance rankings.

7.4 Discussion and Conclusions

Predicting the risk for lymphoma development remains a clinical unmet need in SS. The last decade significant progress has been made towards the understanding of key processes underlying B cell lymphomas occurrence through the identification of novel biomarkers and the development of prediction scores. However, main clinical and genetic aspects of this major complication need to be elucidated for providing a meaningful clinical impact and translational findings in the field.

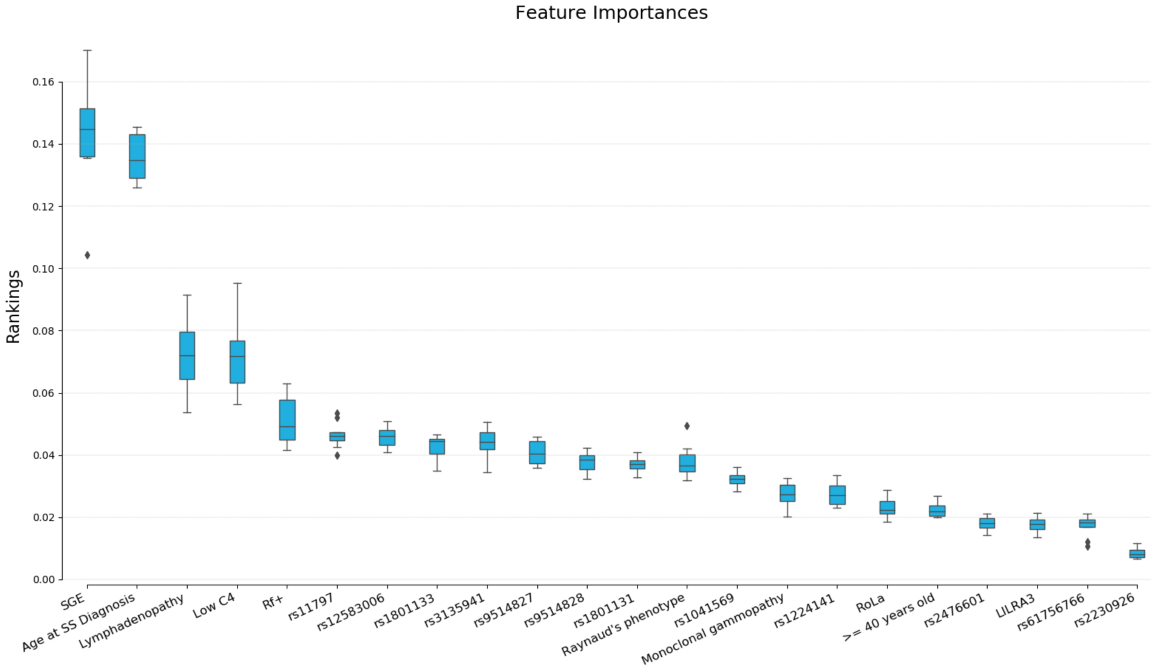


Figure 7.2 Boxplot with the mean feature rankings for each variable considered by the respective estimator (i.e. GB). RF feature selection was performed with threshold the “mean” for the computed importances and “max_features” equal to the max number of features in the dataset considered at the first experiment (input case 1clinical and genetic data).

In this study, we highlight the potential of combining the clinical, histological and serological parameters with the genetic profile of SS patients for the prediction of lymphoma development through a ML methodology consisting of ensemble algorithms. GB and RF classifiers were utilized to obtain accurate classification results based on their generalization ability and the minimization of errors in the training phase. The EasyEnsemble classifier from the imbalanced-learn toolbox was utilized due to the imbalanced dataset of the current study. Based on the selected estimators in the inner ensemble, the training phase was conducted on different balanced bootstrap samples while random under-sampling was considered. Feature selection and ranking was applied in terms of the RF selector based on importance weights. The threshold value used for feature selection and ranking was set to the maximum number of variables within our dataset. The number of features ranked by the estimator was 22, with SGE and age at SS diagnosis being the most important features that contribute to the classification of patients' samples (mean ranking of SGE = 0.1446, mean ranking of age at SS diagnosis = 0.1347). rs12583006 and rs11797 genetic variants are also included within the first 10 most informative features contributing to the prediction of lymphoma development (mean ranking of rs12583006 = 0.0462, mean ranking of rs11797 = 0.0460). The feature ranking results (Figure 7.2) confirmed the identification of SGE and lymphadenopathy as independent adverse predictors for NHL development. We should also note that the age of patients at disease diagnosis could be a potential predictor for lymphoma development. According to published results, mucosa-associated lymphoid tissue (MALT) lymphoma occurs in younger pSS patients [288] which indicates the severity of diagnosis at an early stage.

The reported classification results of the proposed methodology are high with balanced accuracies of 0.7626, 0.7590 and 0.7780 for RF Gini, RF Entropy and GB estimators, respectively. The respective mean AUC obtained by the classifiers are considerably high implying the accurate model prediction in terms of sensitivity and specificity as presented in Table 7.3 (RF Gini classifier: 0.7988, RF Entropy classifier: 0.7995 and GB classifier: 0.8054) and *Figure 7.1*. The mean ROC curves of RF Gini, RF Entropy and GB predictive models, with reference to input case 1, are depicted in Figure 7.3, pinpointing the variance of each curve based on the different subsets created when the training sets are splitted. The figures exhibit how the classifiers output is affected by changes in the training data and how different the subsets are from one another according to the cross-validation procedure. We can observe the low variance which is closely related to the robustness of our methodology.

We can also observe the high results achieved by the three classifiers in terms of the negative predictive value metric (Table 7.3). This constitutes a promising impact of our methodology in predicting accurately the patients that are found as negatives and have not been diagnosed with lymphoma during SS progression. As illustrated in Figure 7.1, the high sensitivity values were obtained when both initial findings and genetic variants are exploited. This reveals the ability of the developed classification models to predict at high proportion the patients who have lymphoma and are truly predicted as positive. To evaluate the predictions on the test sets, different scores were also applied besides the balanced accuracy criterion, such as the f1 score, the log loss metric and the recall. However, the results obtained were similar or with very slight differences in comparison to the balanced accuracy scoring parameter. The proposed methodology was also applied to different input cases (i.e. input cases 2 and 3) where the clinical and genetic variants were considered separately (Table 7.3). Obviously, the exploitation of the genotyped data from the patients result in moderate classification balanced accuracy related to the risk for lymphoma development.

On the contrary, individual clinical, serological and histopathological parameters have been identified in the literature as major predictors of B cell lymphomas. This is in accordance with the reported ML-based classification results (input case 2 in Table 7.3) revealing the superiority of collecting both the initial parameters and the genetic data on the disease onset. In the present work, we highlight the need for identifying risk clinical phenotypes in combination with the patients' genetic profiles for predicting the development of lymphomas which constitutes a major complication of SS. We show that the integration of both the patient's genetic background and the clinical phenotype could enhance the prediction accuracy of our ML models while improving disease diagnosis. Apart from the genotyped data coming from 13 genetic variants in the current study, the integration of new single nucleotide polymorphisms could clearly contribute to the development of more accurate predictive models related to lymphoma development with higher sensitivity and specificity results.

We further validated the methodology with other supervised learning methods used for classification, such as SVM and LR. Given the reported results based on the exploitation of both data types, we demonstrated that the proposed methodology with the ensemble classifiers outperforms the model performance based on SVM and LR. The reported balanced accuracy and AUC for SVM are 0.6395±0.2540 and 0.6934±0.2586, respectively. The evaluated performance for the LR predictive model resulted in a balanced accuracy of 0.7259±0.2087 and an AUC of 0.7962±0.2133.

Based on the scientific studies published in the field which deal with the underlying factors and mechanisms that predispose lymphoma occurrence, we could state that the proposed work constitutes a complementary one with considerable prediction results. Although novel biomarkers have been identified (i.e. BAFF and TNFAIP3 polymorphisms) and validated risk scores have been also developed in terms of clinical parameters, we showed that the combination of both data types and the application of ML-based frameworks could result in robust predictive models with impact in clinical practice. We should also note that the proposed study constitutes a stable methodology by exploiting ensemble classifiers and by addressing the class imbalance problem (Figure 7.3). Moreover, the reported classification results reveal the ability of the selected estimators to generalize their decisions to new unseen records with considerable accuracy and AUC values. The relatively small number of SS patients and the class imbalanced problem related to class 1 (i.e. 64 with either a history or a current diagnosis of SS NHL) are the main limitations of the current study. However, given the rates of unrecognized diagnosis of SS patients in the general population as well as the infrequency of SS initial findings in the healthcare sector, the dataset of the present study can be considered as one of the largest SS databases.

According to the reported classification results in the current study we could conjecture about the potential of exploiting the clinicogenomic profiles of patients for predicting lymphoma development during SS progression. Based on the proposed ML-based methodology we demonstrated that ensemble methods could obtain better classification results than conventional statistical methods and/or other supervised learning algorithms used for the development of predictive models in healthcare. Although lymphoma development presents an unmet clinical need in the research field of SS, the international efforts among groups and the conduction of SS prospective studies could provide a clinical impact to the disease management and the patients’ daily activity by integrating the genetic susceptibility profiles alongside the initial clinical findings.

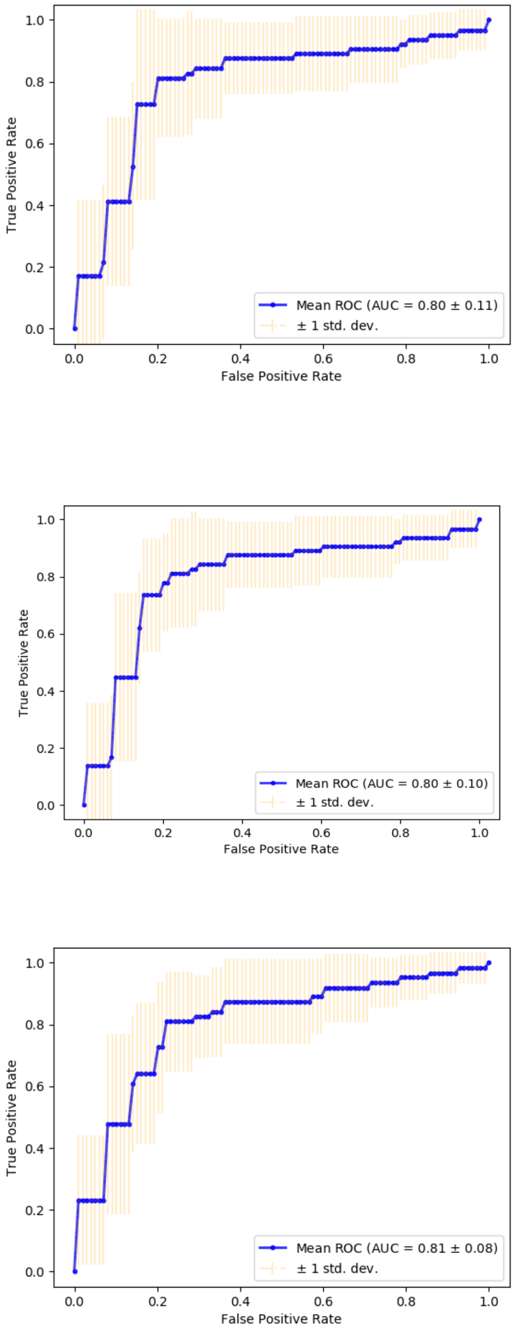


Figure 7.3 The calculated mean ROC curve and AUC, with the variance of each curve when the training set is split into 10 different subsets. This pinpoints how the estimator output is affected by changes in the training data, and how different the splits are from one another in 10-fold cross validation. The upper ROC curve corresponds to RF Gini estimator and the middle and lower ones to RG Entropy and GB classifiers, respectively.

CHAPTER 8 THESIS OUTCOMES AND PERSPECTIVES

8.1 Thesis Outcomes

In the realm of data science and precision oncology, AI-based approaches empower cancer researchers to extract new meaningful information from large clinical cohorts and molecular datasets. The diagnosis, prognosis and treatment of cancer through computational modeling have been widely studied in the literature and several ML-based frameworks have been developed and tested on high throughput datasets that could radically improve the disease management. DNA microarrays and modern sequencing techniques, such as NGS, allow the measurement of expression levels of many genes simultaneously, as they change over time as well as the identification of genetic mutations among phenotypes. Modeling gene expression data and further detect the interactions between genes and functional proteins within biological networks could reveal the molecular processes of cancer onset and progression. To this end, BNs and DBNs have been proposed for modeling gene expression data and inferring their regulatory networks that constitute a paradigm on how molecules interact; thus, forming a network of relevant causal interactions. These networks could be exploited to identify the changes of connections between genes that contribute to the discrimination and classification of different phenotypes. Furthermore, the prediction of cancer diagnosis and prognosis based on ML-based schemes that incorporate both feature selection and classification algorithms have yielded promising results for accelerating decision making in clinical practice and research. What is more, the benefits that ML and data science have brought to cancer research and management allowed to (i) better forecast population trends, (ii) deliver higher preventive patient care, and (iii) tailor treatments according to each patient profile.

Towards this direction, the present thesis contributes to the computational modeling of cancer diagnosis, prognosis and treatment aiming at developing robust and exploitable methodologies that further accelerate the selection of personalized options towards patients' management compromising the quality of life. The modeling of time series gene expression data has been studied thoroughly in terms of DBNs development and application to microarray datasets. Going beyond the state-of-the-art, this thesis deals with knowledge from the pathway level (i.e. regulatory molecules and transcription factors) in conjunction with longitudinal data to better predict the patient phenotype and further classify both cancerous and non-cancerous samples.

The first proposed research study utilizes transcriptomic data from different time points during the follow-up period of the patients aiming at predicting OSCC recurrence through the development of the respective DBN model. The obtained results indicate that modeling gene expression data in terms of DEGs among the two groups of patients can provide better knowledge regarding the prediction of a disease recurrence. The identification of DEGs among patients allowed us to define specific gene patterns that are present to certain phenotypes. Moreover, their fusion with knowledge from the pathway level allowed the extraction of more accurate DBN models, i.e. gene interactions among the different patient groups.

In the next proposed study, time series gene expression data are exploited this time to identify DEGs with their potential MRs for cancer classification. Both promoter and pathway analysis are conducted based on the informative genes for extracting statistically significant regulatory molecules that may contribute to the transcription mechanisms of tumorigenesis. A DBN-based approach is thereby proposed which can model time series gene expression data for classification purposes alongside knowledge from the transcription factors. We achieved an overall higher classification performance by discriminating more than 90% of the test set samples from the first gene expression microarray dataset. Likewise, satisfactory results were achieved when the other two datasets were employed in our methodology, which demonstrates that DBN-based models can accurately classify patient samples. The estimated interactions (i.e. conditional probabilities) though, among genes and their potential regulatory factors could be further examined. Therefore, the application of the current knowledge (meaningful and generalized cancer classification models) to the healthcare systems and cancer treatment domains could leverage predictive modeling and clinician's interpretability.

In the final section of the proposed computational approaches to cancer prediction, we studied thoroughly the potential exploitation of clinicogenomic profiles of patients for predicting lymphoma development during SS progression. Going one step beyond the state-of-the-art we designed and proposed an ML-based methodology and demonstrated that ensemble methods with fusion strategies could obtain better classification results when dealing with predictive modeling. The proposed methodology integrates knowledge from both the genetic profiles and their clinical findings of the patients and further predicts with a good overall performance lymphoma development. The presented ML pipeline assembles several steps that can be cross validated together while setting different parameters. Moreover, it can be applied to heterogenous datasets across many diseases. Finally, our study anticipates that international efforts among groups alongside the conduction of SS prospective studies could provide an additional clinical impact to the disease management and the patients' daily activity through the application of predictive models towards accelerating cancer diagnosis and prognosis.

8.2 Perspectives

Based on the proposed methodologies and the corresponding results of this thesis we can further advance our computational approaches as regards to the integration of heterogenous data sources and the implementation of integrative ML-based approaches for developing multi-modal fusion models towards cancer diagnosis, prognosis. Thus, better treatment options could be adopted.

Concerning the modeling of time series gene expression data and their integration to DBN models for knowledge extraction our methodology can be extended to other types of cancer and diseases while integrating knowledge from various biomedical databases including protein-protein interactions and other omics datasets. Thus, we could further assess and estimate the regulatory interactions that may be altered during the disease onset.

Moreover, the combination of data from different data sources when dealing with biological networks of complex diseases such as cancer could improve drastically our knowledge for estimating gene networks which constitutes the gene regulation mechanisms within cells for functioning.

In addition, the integration of pathway knowledge for modeling cancer progression could be advanced by elucidating the exact transcription factors (TFs) that are included within certain biological processes of cancer cells, identifying the binding sites for known TFs and discovering TF binding motifs in genomic regions. Using genome wide data in the lab, such as genetic interactions, protein-protein interactions and protein domain similarity network data would empower our understanding about the formation of gene regulatory interaction networks and how they change over time based on mutations.

Lymphoma development presents an unmet clinical need in the research field of SS, and several studies have been published for deciphering the onset and progression of cancer occurrence. On this basis, several clinical variables and genetic variants have been proven to increase lymphoma risk development. Moreover, predicting lymphoma in terms of ML-based methodologies could reveal the adverse risk factors during lymphoma diagnosis, prognosis and treatment. We plan to expand our study on lymphoma development prediction by means of other well established ML-based pipelines, such as MKL and DL. Apparently, Genome Wide Association studies (GWAs) could provide observational studies of genome-wide genetic variants which can be easily incorporated in our proposed methodology; thus, enhancing the identification of new population-based risk genetic variants in SS. Towards this direction, the exploitation of large and heterogeneous SS datasets in the future from multicenter studies could contribute to the development of more accurate predictive models through ML techniques. Furthermore, the rise of omics data and their exploitation in the biomedical sciences could permit the identification of key factors involved in lymphomagenesis and the detection of high-risk patients at early stages.

Another brief selection of research directions that we plan to consider for further advancing and accelerating the application of ML in precision oncology and increase predictive accuracy correspond to the development of integrative models for cancer prognosis prediction. On this basis, the integration of heterogeneous data sources would allow to design and deploy predictive models for improving cancer management.

Specifically, a multi-modal fusion strategy that exploits sparse ensembles based on machine learning principles could be designed and deployed for accelerating treatment optimization and de-escalation in cancer. A hybrid approach may be adopted for improving cancer patients' therapy response, cancer diagnostics and treatment, and facilitating patient care. In order to increase the performance of our future strategy, the predictions of several models will be combined in an intelligent way providing a fusion at the decision level. Therefore, to enhance computer-aided diagnosis integrative methods such as MKL and multi-modal deep learning will be adopted. MKL learning tools will be implemented for building predictive models based on multi-view omics data as well as clinical findings. In addition, a deep neural network-based multi-modal structure can be followed to integrate the multi-view omics data and capture their high-level associations for predicting cancer patients that do respond to treatment. A separate subnetwork will be selected for each data view and the output of individual subnetworks in higher layers will be then integrated.

The present thesis deals with the development and application of novel analysis methods and algorithms from the field of ML for modeling cancer diagnosis, prognosis and treatment. High-throughput data, such as gene expression microarray data are modeled in terms of DBNs for cancer classification and prediction. The integration of knowledge from the pathway level, such as TFBSs and other regulatory molecules, revealed the superiority of combining heterogeneous data sources towards accelerating the progression in precision oncology. In addition, cancer risk prediction based on ensemble ML-based pipelines has been studied to highlight the potential usefulness of combining clinical findings along with the genetic susceptibility profiles of patients for personalized treatment selection and disease prevention.

REFERENCES

[1] G. M. Cooper and R. E. Hausman, *The cell: Molecular approach*: Medicinska naklada, 2004.

[2] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell,* vol. 144, pp. 646-674, 2011.

[3] R. Weinberg and D. Hanahan, "The hallmarks of cancer," *Cell,* vol. 100, pp. 57-70, 2000.

[4] H. A. Loomans-Kropp and A. Umar, "Cancer prevention and screening: the next step in the era of precision medicine," *NPJ precision oncology,* vol. 3, pp. 1-8, 2019.

[5] S. McGuire, "World cancer report 2014. Geneva, Switzerland: World Health Organization, international agency for research on cancer, WHO Press, 2015," ed: Oxford University Press, 2016.

[6] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians,* vol. 68, pp. 394-424, 2018.

[7] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros*, et al.*, "Global cancer observatory: cancer today," *Lyon, France: International Agency for Research on Cancer,* 2018.

[8] J. Ferlay, M. Colombet, I. Soerjomataram, C. Mathers, D. Parkin, M. Piñeros*, et al.*, "Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods," *International journal of cancer,* vol. 144, pp. 1941-1953, 2019.

[9] K. Jordan, M. Aapro, S. Kaasa, C. Ripamonti, F. Scotté, F. Strasser*, et al.*, "European Society for Medical Oncology (ESMO) position paper on supportive and palliative care," *Annals of Oncology,* vol. 29, pp. 36-43, 2018.

[10] C. D. Runowicz, C. R. Leach, N. L. Henry, K. S. Henry, H. T. Mackey, R. L. Cowens-Alvarado*, et al.*, "American cancer society/American society of clinical oncology breast cancer survivorship care guideline," *CA: a cancer journal for clinicians,* vol. 66, pp. 43-73, 2016.

[11] W. C. Hahn, C. M. Counter, A. S. Lundberg, R. L. Beijersbergen, M. W. Brooks, and R. A. Weinberg, "Creation of human tumour cells with defined genetic elements," *Nature,* vol. 400, pp. 464-468, 1999.

[12] N. Rivlin, R. Brosh, M. Oren, and V. Rotter, "Mutations in the p53 tumor suppressor gene: important milestones at the various steps of tumorigenesis," *Genes & cancer,* vol. 2, pp. 466-474, 2011.

[13] G. Schneider, M. Schmidt-Supprian, R. Rad, and D. Saur, "Tissue-specific tumorigenesis: context matters," *Nature Reviews Cancer,* vol. 17, p. 239, 2017.

[14] K. Chiba, F. K. Lorbeer, A. H. Shain, D. T. McSwiggen, E. Schruf, A. Oh, *et al.*, "Mutations in the promoter of the telomerase gene TERT contribute to tumorigenesis by a two-step mechanism," *Science,* vol. 357, pp. 1416-1420, 2017.

[15] S. Kumar, S. Jensen, D. Peeney, A. Chowdhury, B. Wei, and W. G. Stetler-Stevenson, "Loss of function mutation in TIMP2 gene accelerates tumorigenesis and mortality in murine model of lung cancer through EGFR signaling," ed: AACR, 2017.

[16] Y. Cho, S. Gorina, P. D. Jeffrey, and N. P. Pavletich, "Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations," *Science,* vol. 265, pp. 346-355, 1994.

[17] M. R. Junttila and G. I. Evan, "p53—a Jack of all trades but master of none," *Nature reviews cancer,* vol. 9, pp. 821-829, 2009.

[18] J. M. Adams and S. Cory, "The Bcl-2 apoptotic switch in cancer development and therapy," *Oncogene,* vol. 26, pp. 1324-1337, 2007.

[19] D. Pan, "The hippo signaling pathway in development and cancer," *Developmental cell,* vol. 19, pp. 491-505, 2010.

[20] C. D. Stefanski and J. R. Prosperi, "Role of Wnt/β-catenin pathway in cancer signaling," in *Predictive Biomarkers in Oncology,* ed: Springer, 2019, pp. 289-295.

[21] M. L. Metzker, "Sequencing technologies—the next generation," *Nature reviews genetics,* vol. 11, pp. 31-46, 2010.

[22] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, *et al.*, "Complementary DNA sequencing: expressed sequence tags and human genome project," *Science,* vol. 252, pp. 1651-1656, 1991.

[23] F. S. Collins, M. Morgan, and A. Patrinos, "The Human Genome Project: lessons from large-scale biology," *Science,* vol. 300, pp. 286-290, 2003.

[24] V. Schneider and D. Church, "Genome reference consortium," *The NCBI Handbook,* p. 117, 2013.

[25] A. J. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock, "Repetitive elements may comprise over two-thirds of the human genome," *PLoS genetics,* vol. 7, 2011.

[26] C. D. Allis and T. Jenuwein, "The molecular hallmarks of epigenetic control," *Nature Reviews Genetics,* vol. 17, p. 487, 2016.

[27] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic acids research,* vol. 35, pp. D61-D65, 2007.

[28] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, *et al.*, "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation," *Nucleic acids research,* vol. 44, pp. D733-D745, 2016.

[29] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, *et al.*, "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome research,* vol. 22, pp. 1760-1774, 2012.

[30] A. Frankish, B. Uszczynska, G. R. Ritchie, J. M. Gonzalez, D. Pervouchine, R. Petryszak, *et al.*, "Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction," *BMC genomics,* vol. 16, p. S2, 2015.

[31] S. Povey, R. Lovering, E. Bruford, M. Wright, M. Lush, and H. Wain, "The HUGO gene nomenclature committee (HGNC)," *Human genetics,* vol. 109, pp. 678-680, 2001.

[32] T. Boveri, "Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris," *Journal of cell science,* vol. 121, pp. 1-84, 2008.

[33] A. G. Knudson, "Two genetic hits (more or less) to cancer," *Nature Reviews Cancer,* vol. 1, pp. 157-162, 2001.

[34] T. Boveri, "Zur frage der entstehung maligner tumoren (Jena: Gustav Fischer)," *The origin of malignant tumors,* 1914.

[35] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nature medicine,* vol. 10, pp. 789-799, 2004.

[36] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature,* vol. 458, pp. 719-724, 2009.

[37] E. Dolgin, "The most popular genes in the human genome," *Nature,* vol. 551, pp. 427-431, Nov 2017.

[38] D. Stehelin, H. E. Varmus, J. M. Bishop, and P. K. Vogt, "DNA related to the transforming gene (s) of avian sarcoma viruses is present in normal avian DNA," *Nature,* vol. 260, pp. 170-173, 1976.

[39] M. Gonzalez-Cao, P. Iduma, N. Karachaliou, M. Santarpia, J. Blanco, and R. Rosell, "Human endogenous retroviruses and cancer," *Cancer biology & medicine,* vol. 13, p. 483, 2016.

[40] A. E. Karnoub and R. A. Weinberg, "Ras oncogenes: split personalities," *Nature reviews Molecular cell biology,* vol. 9, pp. 517-531, 2008.

[41] D. Grandér, "How do mutated oncogenes and tumor suppressor genes cause cancer?," *Medical oncology,* vol. 15, pp. 20-26, 1998.

[42] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *science,* vol. 339, pp. 1546-1558, 2013.

[43] F. McCormick, "Signalling networks that cause cancer," *Trends in Genetics,* vol. 15, pp. M53-M56, 1999.

[44] R. A. Previs, R. L. Coleman, A. L. Harris, and A. K. Sood, "Molecular pathways: translational and therapeutic implications of the Notch signaling pathway in cancer," *Clinical Cancer Research,* vol. 21, pp. 955-961, 2015.

[45] N. E. Hynes and H. A. Lane, "ERBB receptors and cancer: the complexity of targeted inhibitors," *Nature Reviews Cancer,* vol. 5, pp. 341-354, 2005.

[46] N. Perrimon, C. Pitsouli, and B.-Z. Shilo, "Signaling mechanisms controlling cell fate and embryonic patterning," *Cold Spring Harbor perspectives in biology,* vol. 4, p. a005975, 2012.

[47] N. Turner and R. Grose, "Fibroblast growth factor signalling: from development to cancer," *Nature Reviews Cancer,* vol. 10, pp. 116-129, 2010.

[48] H. Ying, A. C. Kimmelman, C. A. Lyssiotis, S. Hua, G. C. Chu, E. Fletcher-Sananikone*, et al.*, "Oncogenic Kras maintains pancreatic tumors through regulation of anabolic glucose metabolism," *Cell,* vol. 149, pp. 656-670, 2012.

[49] M. Ljungman and D. P. Lane, "Transcription—guarding the genome by sensing DNA damage," *Nature Reviews Cancer,* vol. 4, pp. 727-737, 2004.

[50] R. Medema and L. Macůrek, "Checkpoint control and cancer," *Oncogene,* vol. 31, pp. 2601-2613, 2012.

[51] B.-B. S. Zhou and S. J. Elledge, "The DNA damage response: putting checkpoints in perspective," *Nature,* vol. 408, pp. 433-439, 2000.

[52] M. Song, B. Vogelstein, E. L. Giovannucci, W. C. Willett, and C. Tomasetti, "Cancer prevention: molecular and epidemiologic consensus," *Science,* vol. 361, pp. 1317-1318, 2018.

[53] V. Brower, "Tumor angiogenesis—new drugs on the block," *Nature biotechnology,* vol. 17, pp. 963-968, 1999.

[54] S. M. Lippman and J. V. Heymach, "The convergent development of molecular-targeted drugs for cancer treatment and prevention," *Clinical Cancer Research,* vol. 13, pp. 4035-4041, 2007.

[55] D. Schottenfeld and J. F. Fraumeni Jr, *Cancer epidemiology and prevention*: Oxford University Press, 2006.

[56] R. L. Siegel, K. D. Miller, and A. Jemal, "Descriptive Epidemiology," *The American Cancer Society's Principles of Oncology: Prevention to Survivorship,* p. 3, 2018.

[57] A. C. Society. (2020, 30/03/2020). *Global Cancer Facts & Figures 4th Edition*. Available: https://www.cancer.org/research/cancer-facts-statistics/global.html

[58] S. Anand and A. Sen, "Human Development Index: Methodology and Measurement," 1994.

[59] M. Plummer, C. de Martel, J. Vignat, J. Ferlay, F. Bray, and S. Franceschi, "Global burden of cancers attributable to infections in 2012: a synthetic analysis," *The Lancet Global Health,* vol. 4, pp. e609-e616, 2016.

[60] B. Jönsson, T. Hofmarcher, P. Lindgren, and N. Wilking, "The cost and burden of cancer in the European Union 1995–2014," *European Journal of Cancer,* vol. 66, pp. 162-170, 2016.

[61] E. B. Elkin and P. B. Bach, "Cancer's next frontier: addressing high and increasing costs," *Jama,* vol. 303, pp. 1086-1087, 2010.

[62] R. Rubin, "Profile: Institute for Health Metrics and Evaluation, WA, USA," *The Lancet,* vol. 389, p. 493, 2017.

[63] D. S. Chan, L. Abar, M. Cariolou, N. Nanu, D. C. Greenwood, E. V. Bandera, *et al.*, "World Cancer Research Fund International: Continuous Update Project—systematic

literature review and meta-analysis of observational cohort studies on physical activity, sedentary behavior, adiposity, and weight change and breast cancer risk," *Cancer Causes & Control*, pp. 1-18, 2019.

[64] S. F. Sener, "Disease without borders," *CA: a cancer journal for clinicians*, vol. 55, pp. 7-9, 2005.

[65] W. H. Organization, *National cancer control programmes: policies and managerial guidelines*: World Health Organization, 2002.

[66] K. Sharma, B. K. Mohanti, G. K. Rath, and S. Bhatnagar, "Pattern of palliative care, pain management and referral trends in patients receiving radiotherapy at a tertiary cancer center," *Indian journal of palliative care*, vol. 15, p. 148, 2009.

[67] F. M. Knaul, P. E. Farmer, E. L. Krakauer, L. De Lima, A. Bhadelia, X. J. Kwete, *et al.*, "Alleviating the access abyss in palliative care and pain relief—an imperative of universal health coverage: the Lancet Commission report," *The Lancet*, vol. 391, pp. 1391-1454, 2018.

[68] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, p. 333, 2016.

[69] W. P. Janzen, *High throughput screening: methods and protocols*: Springer Science & Business Media, 2001.

[70] H. E. Duckworth, R. C. Barber, and V. Venkatasubramanian, "Mass spectroscopy," *Mass Spectroscopy, by HE Duckworth, RC Barber, VS Venkatasubramanian, Cambridge, UK: Cambridge University Press, 1990*, 1990.

[71] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemporary oncology*, vol. 19, p. A68, 2015.

[72] G. England, "The 100,000 genomes project," *The*, vol. 100, pp. 0-2, 2016.

[73] S. J. Mooney, D. J. Westreich, and A. M. El-Sayed, "Epidemiology in the era of big data," *Epidemiology (Cambridge, Mass.)*, vol. 26, p. 390, 2015.

[74] A.-T. Maia, S.-J. Sammut, A. Jacinta-Fernandes, and S.-F. Chin, "Big data in cancer genomics," *Current Opinion in Systems Biology*, vol. 4, pp. 78-84, 2017.

[75] D. Cirillo and A. Valencia, "Big data analytics for personalized medicine," *Current opinion in biotechnology*, vol. 58, pp. 161-167, 2019.

[76] E. Capobianco, "Precision Oncology: The Promise of Big Data and the Legacy of Small Data," *Frontiers in ICT,* vol. 4, p. 22, 2017.

[77] T. Hulsen, S. S. Jamuar, A. Moody, J. H. Karnes, V. Orsolya, S. Hedensted, *et al.*, "From big data to precision medicine," *Frontiers in medicine,* vol. 6, p. 34, 2019.

[78] M. A. Haendel, C. G. Chute, and P. N. Robinson, "Classification, ontology, and precision medicine," *New England Journal of Medicine,* vol. 379, pp. 1452-1462, 2018.

[79] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England journal of medicine,* vol. 372, pp. 793-795, 2015.

[80] C. Vesteghem, R. F. Brøndum, M. Sønderkær, M. Sommer, A. Schmitz, J. S. Bødker, *et al.*, "Implementing the FAIR Data Principles in precision oncology: review of supporting initiatives," *Briefings in bioinformatics,* 2019.

[81] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific data,* vol. 3, 2016.

[82] E. M. Cahan, T. Hernandez-Boussard, S. Thadaney-Israni, and D. L. Rubin, "Putting the data before the algorithm in big data addressing personalized healthcare," *NPJ Digital Medicine,* vol. 2, pp. 1-6, 2019.

[83] S. Munevar, "Unlocking big data for better health," *Nature biotechnology,* vol. 35, pp. 684-686, 2017.

[84] G. Bertier, J. Carrot-Zhang, V. Ragoussis, and Y. Joly, "Integrating precision cancer medicine into healthcare—policy, practice, and research challenges," *Genome medicine,* vol. 8, p. 108, 2016.

[85] I. Budin-Ljøsne, J. Isaeva, B. M. Knoppers, A. M. Tassé, H.-y. Shen, M. I. McCarthy, *et al.*, "Data sharing in large research consortia: experiences and recommendations from ENGAGE," *European Journal of Human Genetics,* vol. 22, pp. 317-321, 2014.

[86] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, *et al.*, "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers," *BMC genomics,* vol. 13, p. 341, 2012.

[87] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype–phenotype interactions," *Nature Reviews Genetics,* vol. 16, pp. 85-97, 2015.

[88] H. Carter, M. Hofree, and T. Ideker, "Genotype to phenotype via network analysis," *Current opinion in genetics & development,* vol. 23, pp. 611-621, 2013.

[89] B. Ristevski and M. Chen, "Big data analytics in medicine and healthcare," *Journal of integrative bioinformatics,* vol. 15, 2018.

[90] A. R. Sonawane, S. T. Weiss, K. Glass, and A. Sharma, "Network medicine in the age of biomedical big data," *Frontiers in Genetics,* vol. 10, 2019.

[91] D. J. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent, "Expression profiling using cDNA microarrays," *Nature genetics,* vol. 21, pp. 10-14, 1999.

[92] E. R. Mardis, "DNA sequencing technologies: 2006–2016," *Nature protocols,* vol. 12, p. 213, 2017.

[93] G. Lightbody, V. Haberland, F. Browne, L. Taggart, H. Zheng, E. Parkes*, et al.*, "Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application," *Briefings in bioinformatics,* vol. 20, pp. 1795-1811, 2019.

[94] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes*, et al.*, "Nucleotide sequence of bacteriophage φX174 DNA," *nature,* vol. 265, pp. 687-695, 1977.

[95] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome biology,* vol. 18, p. 83, 2017.

[96] M. Kim and I. Tagkopoulos, "Data integration and predictive modeling methods for multi-omics datasets," *Molecular omics,* vol. 14, pp. 8-25, 2018.

[97] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science,* vol. 270, pp. 467-470, 1995.

[98] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes," *Proceedings of the National Academy of Sciences,* vol. 93, pp. 10614-10619, 1996.

[99] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee*, et al.*, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature biotechnology,* vol. 14, pp. 1675-1680, 1996.

[100] B. Wold and R. M. Myers, "Sequence census methods for functional genomics," *Nature methods,* vol. 5, pp. 19-21, 2008.

[101] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature reviews genetics,* vol. 10, pp. 57-63, 2009.

[102] T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W.-C. Ngau, P. Ledoux, *et al.*, "NCBI GEO: mining millions of expression profiles—database and tools," *Nucleic acids research,* vol. 33, pp. D562-D566, 2005.

[103] M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, *et al.*, "The NCBI dbGaP database of genotypes and phenotypes," *Nature genetics,* vol. 39, pp. 1181-1186, 2007.

[104] J. A. Vizcaíno, E. W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Rios, *et al.*, "ProteomeXchange provides globally coordinated proteomics data submission and dissemination," *Nature biotechnology,* vol. 32, pp. 223-226, 2014.

[105] R. A. Spicer and C. Steinbeck, "A lost opportunity for science: journals promote data sharing in metabolomics but do not enforce it," *Metabolomics,* vol. 14, p. 16, 2018.

[106] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, *et al.*, "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)," *Nucleic acids research,* vol. 45, pp. D896-D901, 2017.

[107] C. Manzoni, D. A. Kia, J. Vandrovcova, J. Hardy, N. W. Wood, P. A. Lewis, *et al.*, "Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences," *Briefings in bioinformatics,* vol. 19, pp. 286-302, 2018.

[108] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, *et al.*, "The sequence of the human genome," *science,* vol. 291, pp. 1304-1351, 2001.

[109] C. Gonzaga-Jauregui, J. R. Lupski, and R. A. Gibbs, "Human genome sequencing in health and disease," *Annual review of medicine,* vol. 63, pp. 35-61, 2012.

[110] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, *et al.*, "ArrayExpress—a public database of microarray experiments and gene expression profiles," *Nucleic acids research,* vol. 35, pp. D747-D750, 2007.

[111] R. Petryszak, M. Keays, Y. A. Tang, N. A. Fonseca, E. Barrera, T. Burdett, *et al.*, "Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants," *Nucleic acids research,* vol. 44, pp. D746-D752, 2016.

[112] L. Martens, "Proteomics databases and repositories," in *Bioinformatics for Comparative Proteomics,* ed: Springer, 2011, pp. 213-227.

[113] G. C. Koh, P. Porras, B. Aranda, H. Hermjakob, and S. E. Orchard, "Analyzing protein–protein interaction networks," *Journal of proteome research,* vol. 11, pp. 2014-2031, 2012.

[114] V. N. Kristensen, O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi, and A.-L. Børresen-Dale, "Principles and methods of integrative genomic analyses in cancer," *Nature Reviews Cancer,* vol. 14, pp. 299-313, 2014.

[115] P.-Y. Wu, C.-W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, "–Omic and electronic health record big data analytics for precision medicine," *IEEE Transactions on Biomedical Engineering,* vol. 64, pp. 263-273, 2016.

[116] E. J. Nestler, "Transgenerational epigenetic contributions to stress responses: fact or fiction?," *PLoS biology,* vol. 14, 2016.

[117] E. Org, B. W. Parks, J. W. J. Joo, B. Emert, W. Schwartzman, E. Y. Kang*, et al.*, "Genetic and environmental control of host-gut microbiota interactions," *Genome research,* vol. 25, pp. 1558-1569, 2015.

[118] E. Org, M. Mehrabian, and A. J. Lusis, "Unraveling the environmental and genetic interactions in atherosclerosis: central role of the gut microbiota," *Atherosclerosis,* vol. 241, pp. 387-399, 2015.

[119] B. Neumann, T. Walter, J.-K. Hériché, J. Bulkescher, H. Erfle, C. Conrad*, et al.*, "Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes," *Nature,* vol. 464, pp. 721-727, 2010.

[120] A. Iudin, P. K. Korir, J. Salavert-Torres, G. J. Kleywegt, and A. Patwardhan, "EMPIAR: a public archive for raw electron microscopy image data," *Nature methods,* vol. 13, p. 387, 2016.

[121] E. Williams, J. Moore, S. W. Li, G. Rustici, A. Tarkowska, A. Chessel*, et al.*, "Image Data Resource: a bioimage data integration and publication platform," *Nature methods,* vol. 14, p. 775, 2017.

[122] J. Wu, K. K. Tha, L. Xing, and R. Li, "Radiomics and radiogenomics for precision radiotherapy," *Journal of radiation research,* vol. 59, pp. i25-i31, 2018.

[123] R. L. Gullo, I. Daimiel, E. A. Morris, and K. Pinker, "Combining molecular and imaging metrics in cancer: radiogenomics," *Insights into Imaging,* vol. 11, p. 1, 2020.

[124] M. Zanfardino, K. Pane, P. Mirabelli, M. Salvatore, and M. Franzese, "TCGA-TCIA Impact on Radiogenomics Cancer Research: A Systematic Review," *International Journal of Molecular Sciences,* vol. 20, p. 6033, 2019.

[125] N. Genes, S. Violante, C. Cetrangol, L. Rogers, E. E. Schadt, and Y.-F. Y. Chan, "From smartphone to EHR: a case report on integrating patient-generated health data," *NPJ digital medicine,* vol. 1, pp. 1-6, 2018.

[126] M. C. Dunn and P. E. Bourne, "Building the biomedical data science workforce," *PLoS biology,* vol. 15, p. e2003082, 2017.

[127] S. G. Komen. (2020). *Big Data for Breast Cancer (BD4BC)*. Available: https://bd4bc.komen.org/

[128] N. J. Nilsson, *The quest for artificial intelligence*: Cambridge University Press, 2009.

[129] J. Haugeland, *Artificial intelligence: The very idea*: MIT press, 1989.

[130] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering,* vol. 2, pp. 719-731, 2018.

[131] M. Fenech, N. Strukelj, and O. Buston, "Ethical, social, and political challenges of artificial intelligence in health," *Future Advocacy & Wellcome Trust, London) https://wellcome. ac. uk/sites/default/files/ai-in-health-ethicalsocial-political-challenges. pdf (accessed 26 Jul 2019),* 2018.

[132] M. Fenech, "Maximising the Opportunities of Artificial Intelligence for People Living With Cancer," *Clinical Oncology,* vol. 32, pp. e80-e85, 2020.

[133] F. Azuaje, "Artificial intelligence for precision oncology: beyond patient stratification," *NPJ precision oncology,* vol. 3, pp. 1-5, 2019.

[134] Y. Li, F.-X. Wu, and A. Ngom, "A review on machine learning principles for multi-view biological data integration," *Briefings in bioinformatics,* vol. 19, pp. 325-340, 2018.

[135] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering,* vol. 22, pp. 1345-1359, 2009.

[136] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," *arXiv preprint arXiv:1711.06104,* 2017.

[137] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 9, p. e1312, 2019.

[138] K. Y. Michael, J. Ma, J. Fisher, J. F. Kreisberg, B. J. Raphael, and T. Ideker, "Visible machine learning for biomedicine," *Cell,* vol. 173, pp. 1562-1565, 2018.

[139] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, 2016, pp. 3504-3512.

[140] O. Lahav, N. Mastronarde, and M. van der Schaar, "What is interpretable? Using machine learning to design interpretable decision-support systems," *arXiv preprint arXiv:1811.10799,* 2018.

[141] A. M. Alaa and M. van der Schaar, "Forecasting individualized disease trajectories using interpretable deep learning," *arXiv preprint arXiv:1810.10489,* 2018.

[142] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*: Pearson Education India, 2016.

[143] C. M. Bishop, *Pattern recognition and machine learning*: springer, 2006.

[144] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," *Jama,* vol. 319, pp. 1317-1318, 2018.

[145] C. J. Lynch and C. Liston, "New machine-learning technologies for computer-aided diagnosis," *Nature medicine,* vol. 24, pp. 1304-1305, 2018.

[146] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics,* vol. 19, pp. 1236-1246, 2018.

[147] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in bioinformatics,* vol. 18, pp. 851-869, 2017.

[148] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Molecular systems biology,* vol. 12, 2016.

[149] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics and intelligent laboratory systems,* vol. 39, pp. 43-62, 1997.

[150] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of machine learning research,* vol. 11, pp. 3371-3408, 2010.

[151] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks,* vol. 8, pp. 98-113, 1997.

[152] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems,* 2012, pp. 1097-1105.

[153] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, pp. 1735-1780, 1997.

[154] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

[155] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature,* vol. 521, pp. 436-444, 2015.

[156] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," in *Proceedings of the international conference on machine learning*, 2013.

[157] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature,* vol. 542, pp. 115-118, 2017.

[158] P. Gupta and A. K. Malhi, "Using deep learning to enhance head and neck cancer diagnosis and classification," in *2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA)*, 2018, pp. 1-6.

[159] Y. Chen, Y. Li, R. Narayan, A. Subramanian, and X. Xie, "Gene expression inference with deep learning," *Bioinformatics,* vol. 32, pp. 1832-1839, 2016.

[160] D. Baptista, P. G. Ferreira, and M. Rocha, "Deep learning for drug response prediction in cancer," *Briefings in Bioinformatics*.

[161] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*: MIT press, 2018.

[162] K. H. Shain, D. Hart, A. Siqueira Silva, R. Alugubelli, G. De Avila, P. R. Sudalagunta, *et al.*, "Reinforcement Learning to Optimize the Treatment of Multiple Myeloma," ed: American Society of Hematology Washington, DC, 2019.

[163] Z. Liu, C. Yao, H. Yu, and T. Wu, "Deep reinforcement learning with its application for lung cancer detection in medical internet of things," *Future Generation Computer Systems,* vol. 97, pp. 1-9, 2019.

[164] W. Zhang, J. Chien, J. Yong, and R. Kuang, "Network-based machine learning and graph theory algorithms for precision oncology," *NPJ precision oncology,* vol. 1, pp. 1-15, 2017.

[165] A. L. Beam, A. K. Manrai, and M. Ghassemi, "Challenges to the Reproducibility of Machine Learning Models in Health Care," *JAMA,* 2020.

[166] K. Y. Ngiam and W. Khor, "Big data and machine learning algorithms for health-care delivery," *The Lancet Oncology,* vol. 20, pp. e262-e273, 2019.

[167] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov, "Applications of deep learning in biomedicine," *Molecular pharmaceutics,* vol. 13, pp. 1445-1454, 2016.

[168] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*: MIT press, 2009.

[169] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature,* vol. 521, pp. 452-459, 2015.

[170] F. M. Delgado and F. Gómez-Vela, "Computational methods for Gene Regulatory Networks reconstruction and analysis: A review," *Artificial intelligence in medicine,* vol. 95, pp. 133-145, 2019.

[171] S. Guo, Q. Jiang, L. Chen, and D. Guo, "Gene regulatory network inference using PLS-based methods," *BMC bioinformatics,* vol. 17, p. 545, 2016.

[172] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of computational biology,* vol. 7, pp. 601-620, 2000.

[173] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano, "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks," *Journal of bioinformatics and computational biology,* vol. 2, pp. 77-98, 2004.

[174] A. Livshits, A. Git, G. Fuks, C. Caldas, and E. Domany, "Pathway-based personalized analysis of breast cancer expression data," *Molecular oncology,* vol. 9, pp. 1471-1483, 2015.

[175] C. A. Penfold, A. Shifaz, P. E. Brown, A. Nicholson, and D. L. Wild, "CSI: a nonparametric Bayesian approach to network inference from multiple perturbed time series gene expression data," *Statistical applications in genetics and molecular biology,* vol. 14, pp. 307-310, 2015.

[176] A. Darwiche, *Modeling and reasoning with Bayesian networks*: Cambridge university press, 2009.

[177] W. Yang, K. Yoshigoe, X. Qin, J. S. Liu, J. Y. Yang, A. Niemierko, *et al.*, "Identification of genes and pathways involved in kidney renal clear cell carcinoma," *BMC bioinformatics,* vol. 15, p. S2, 2014.

[178] K. Murphy, "The bayes net toolbox for matlab," *Computing science and statistics,* vol. 33, pp. 1024-1034, 2001.

[179] K. P. Murphy and S. Russell, "Dynamic bayesian networks: representation, inference and learning," 2002.

[180] S. Isci, H. Dogan, C. Ozturk, and H. H. Otu, "Bayesian network prior: network analysis of biological data using external knowledge," *Bioinformatics,* vol. 30, pp. 860-867, 2014.

[181] Y. Ni, P. Müller, L. Wei, and Y. Ji, "Bayesian graphical models for computational network biology," *BMC bioinformatics,* vol. 19, p. 63, 2018.

[182] D. Heckerman, C. Meek, and G. Cooper, "A Bayesian approach to causal discovery," *Computation, causation, and discovery,* vol. 19, pp. 141-166, 1999.

[183] A. V. Werhli and D. Husmeier, "Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge," *Statistical applications in genetics and molecular biology,* vol. 6, 2007.

[184] K. Murphy and S. Mian, "Modelling gene expression data using dynamic Bayesian networks," Technical report, Computer Science Division, University of California …1999.

[185] S. Y. Kim, S. Imoto, and S. Miyano, "Inferring gene networks from time series microarray data using dynamic Bayesian networks," *Briefings in bioinformatics,* vol. 4, pp. 228-235, 2003.

[186] B. Baur and S. Bozdag, "A canonical correlation analysis-based dynamic bayesian network prior to infer gene regulatory networks from multiple types of biological data," *Journal of Computational Biology,* vol. 22, pp. 289-299, 2015.

[187] R. Agrahari, A. Foroushani, T. R. Docking, L. Chang, G. Duns, M. Hudoba*, et al.*, "Applications of Bayesian network models in predicting types of hematological malignancies," *Scientific reports,* vol. 8, pp. 1-12, 2018.

[188] M. B. Sesen, A. E. Nicholson, R. Banares-Alcantara, T. Kadir, and M. Brady, "Bayesian networks for clinical decision support in lung cancer care," *PloS one,* vol. 8, 2013.

[189] D. Koch, R. S. Eisinger, and A. Gebharter, "A causal Bayesian network model of disease progression mechanisms in chronic myeloid leukemia," *Journal of theoretical biology,* vol. 433, pp. 94-105, 2017.

[190] K. P. Exarchos, Y. Goletsis, and D. I. Fotiadis, "A multiscale and multiparametric approach for modeling the progression of oral cancer," *BMC medical informatics and decision making,* vol. 12, p. 136, 2012.

[191] K. Glass, J. Quackenbush, D. Spentzos, B. Haibe-Kains, and G.-C. Yuan, "A network model for angiogenesis in ovarian cancer," *BMC bioinformatics,* vol. 16, p. 115, 2015.

[192] J. Zheng, I. Chaturvedi, and J. C. Rajapakse, "Integration of epigenetic data in bayesian network modeling of gene regulatory network," in *IAPR International Conference on Pattern Recognition in Bioinformatics*, 2011, pp. 87-96.

[193] R. Gendelman, H. Xing, O. K. Mirzoeva, P. Sarde, C. Curtis, H. S. Feiler*, et al.*, "Bayesian network inference modeling identifies TRIB1 as a novel regulator of cell-cycle progression and survival in cancer cells," *Cancer research,* vol. 77, pp. 1575-1585, 2017.

[194] H. Xing, P. D. McDonagh, J. Bienkowska, T. Cashorali, K. Runge, R. E. Miller*, et al.*, "Causal modeling using network ensemble simulations of genetic and gene expression data predicts genes involved in rheumatoid arthritis," *PLoS computational biology,* vol. 7, 2011.

[195] A. Stojadinovic, A. Bilchik, D. Smith, J. S. Eberhardt, E. B. Ward, A. Nissan*, et al.*, "Clinical decision support and individualized prediction of survival in colon cancer: Bayesian belief network model," *Annals of surgical oncology,* vol. 20, pp. 161-174, 2013.

[196] J. A. Forsberg, J. Eberhardt, P. J. Boland, R. Wedin, and J. H. Healey, "Estimating survival in patients with operable skeletal metastases: an application of a bayesian belief network," *PloS one,* vol. 6, 2011.

[197] D. Zhao and C. Weng, "Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction," *Journal of biomedical informatics,* vol. 44, pp. 859-868, 2011.

[198] Y. Kim, S. Han, S. Choi, and D. Hwang, "Inference of dynamic networks using time-course data," *Briefings in bioinformatics,* vol. 15, pp. 212-228, 2014.

[199] H. Li, N. Wang, P. Gong, E. J. Perkins, and C. Zhang, "Learning the structure of gene regulatory networks from time series gene expression data," *BMC genomics,* vol. 12, p. S13, 2011.

[200] P. Petousis, S. X. Han, D. Aberle, and A. A. Bui, "Prediction of lung cancer incidence on the low-dose computed tomography arm of the National Lung Screening Trial: A dynamic Bayesian network," *Artificial intelligence in medicine,* vol. 72, pp. 42-55, 2016.

[201] R. M. Austin and A. Onisko, "Increased cervical cancer risk associated with extended screening intervals after negative human papillomavirus test results: Bayesian risk estimates using the Pittsburgh Cervical Cancer Screening Model," *Journal of the American Society of Cytopathology,* vol. 5, pp. 9-14, 2016.

[202] Y. Li, K. Kang, J. M. Krahn, N. Croutwater, K. Lee, D. M. Umbach, *et al.*, "A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data," *BMC genomics,* vol. 18, p. 508, 2017.

[203] T. Ching, X. Zhu, and L. X. Garmire, "Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data," *PLoS computational biology,* vol. 14, p. e1006076, 2018.

[204] R.-J. Kuo, M.-H. Huang, W.-C. Cheng, C.-C. Lin, and Y.-H. Wu, "Application of a two-stage fuzzy neural network to a prostate cancer prognosis system," *Artificial intelligence in medicine,* vol. 63, pp. 119-133, 2015.

[205] K. Park, A. Ali, D. Kim, Y. An, M. Kim, and H. Shin, "Robust predictive model for evaluating breast cancer survivability," *Engineering Applications of Artificial Intelligence,* vol. 26, pp. 2194-2205, 2013.

[206] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin, *et al.*, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature communications,* vol. 7, p. 12474, 2016.

[207] T.-P. Lu, K.-T. Kuo, C.-H. Chen, M.-C. Chang, H.-P. Lin, Y.-H. Hu, *et al.*, "Developing a prognostic gene panel of epithelial ovarian cancer patients by a machine learning model," *Cancers,* vol. 11, p. 270, 2019.

[208] S. Zhang, Y. Xu, X. Hui, F. Yang, Y. Hu, J. Shao, *et al.*, "Improvement in prediction of prostate cancer prognosis with somatic mutational signatures," *Journal of Cancer,* vol. 8, p. 3261, 2017.

[209] Z. Qian, Y. Li, Y. Wang, L. Li, R. Li, K. Wang, *et al.*, "Differentiation of glioblastoma from solitary brain metastases using radiomic machine-learning classifiers," *Cancer letters,* vol. 451, pp. 128-135, 2019.

[210] W. Kim, K. S. Kim, J. E. Lee, D.-Y. Noh, S.-W. Kim, Y. S. Jung, *et al.*, "Development of novel breast cancer recurrence prediction model using support vector machine," *Journal of breast cancer,* vol. 15, pp. 230-238, 2012.

[211] C.-J. Tseng, C.-J. Lu, C.-C. Chang, and G.-D. Chen, "Application of machine learning to predict the recurrence-proneness for cervical cancer," *Neural Computing and Applications,* vol. 24, pp. 1311-1316, 2014.

[212] M. Waddell, D. Page, and J. Shaughnessy Jr, "Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma," in *Proceedings of the 5th international workshop on Bioinformatics*, 2005, pp. 21-28.

[213] J. Listgarten, S. Damaraju, B. Poulin, L. Cook, J. Dufour, A. Driga, *et al.*, "Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms," *Clinical cancer research,* vol. 10, pp. 2725-2737, 2004.

[214] H. Lu, H. Wang, and S. W. Yoon, "A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis," *Expert Systems with Applications,* vol. 116, pp. 340-350, 2019.

[215] P. Vasudevan and T. Murugesan, "Cancer subtype discovery using prognosis-enhanced neural network classifier in multigenomic data," *Technology in cancer research & treatment,* vol. 17, p. 1533033818790509, 2018.

[216] S. Sepehri, T. Upadhaya, M.-C. Desseroit, D. Visvikis, C. C. Le Rest, and M. Hatt, "Comparison of machine learning algorithms for building prognostic models in non-small cell lung cancer using clinical and radiomics features from 18F-FDG PET/CT images," *Journal of Nuclear Medicine,* vol. 59, pp. 328-328, 2018.

[217] L. G. Ahmad, A. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, and A. Razavi, "Using three machine learning techniques for predicting breast cancer recurrence," *J Health Med Inform,* vol. 4, p. 3, 2013.

[218] K. P. Exarchos, Y. Goletsis, and D. I. Fotiadis, "Multiparametric decision support system for the prediction of oral cancer reoccurrence," *IEEE Transactions on Information Technology in Biomedicine,* vol. 16, pp. 1127-1134, 2011.

[219] S.-W. Chang, S. Abdul-Kareem, A. F. Merican, and R. B. Zain, "Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods," *BMC bioinformatics,* vol. 14, p. 170, 2013.

[220] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 16, pp. 841-850, 2018.

[221] X. Li, S. Zhang, Q. Zhang, X. Wei, Y. Pan, J. Zhao, *et al.*, "Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study," *The Lancet Oncology,* vol. 20, pp. 193-201, 2019.

[222] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, *et al.*, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature medicine,* vol. 24, pp. 1559-1567, 2018.

[223] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," *Computer methods and programs in biomedicine,* vol. 153, pp. 1-9, 2018.

[224] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology,* vol. 292, pp. 60-66, 2019.

[225] A. Akselrod-Ballin, M. Chorev, Y. Shoshan, A. Spiro, A. Hazan, R. Melamed*, et al.*, "Predicting breast cancer by applying deep learning to linked health records and mammograms," *Radiology,* vol. 292, pp. 331-342, 2019.

[226] A. Rodríguez-Ruiz, E. Krupinski, J.-J. Mordang, K. Schilling, S. H. Heywang-Köbrunner, I. Sechopoulos*, et al.*, "Detection of breast cancer with mammography: effect of an artificial intelligence support system," *Radiology,* vol. 290, pp. 305-314, 2019.

[227] E. Wulczyn, D. F. Steiner, Z. Xu, A. Sadhwani, H. Wang, I. Flament*, et al.*, "Deep learning-based survival prediction for multiple cancer types using histopathology images," *arXiv preprint arXiv:1912.07354,* 2019.

[228] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis*, et al.*, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS medicine,* vol. 16, 2019.

[229] Y. Wang, D. Wang, X. Ye, Y. Wang, Y. Yin, and Y. Jin, "A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction," *Information Sciences,* vol. 474, pp. 106-124, 2019.

[230] N. C. Institute. (30/03/2020). *The Surveillance, Epidemiology, and End Results (SEER) Program*. Available: https://seer.cancer.gov/

[231] R. J. Kate and R. Nadig, "Stage-specific predictive models for breast cancer survivability," *International journal of medical informatics,* vol. 97, pp. 304-311, 2017.

[232] L. Zhu, W. Luo, M. Su, H. Wei, J. Wei, X. Zhang*, et al.*, "Comparison between artificial neural network and Cox regression model in predicting the survival rate of gastric cancer patients," *Biomedical reports,* vol. 1, pp. 757-760, 2013.

[233] L. Papp, N. Pötsch, M. Grahovac, V. Schmidbauer, A. Woehrer, M. Preusser*, et al.*, "Glioma survival prediction with combined analysis of in vivo 11C-MET PET features, ex vivo features, and patient features by supervised machine learning," *Journal of Nuclear Medicine,* vol. 59, pp. 892-899, 2018.

[234] Z. Hasnain, J. Mason, K. Gill, G. Miranda, I. S. Gill, P. Kuhn, *et al.*, "Machine learning models for predicting post-cystectomy recurrence and survival in bladder cancer patients," *PloS one,* vol. 14, 2019.

[235] C. M. Lynch, B. Abdollahi, J. D. Fuqua, R. Alexandra, J. A. Bartholomai, R. N. Balgemann, *et al.*, "Prediction of lung cancer patient survival via supervised machine learning classification techniques," *International journal of medical informatics,* vol. 108, pp. 1-8, 2017.

[236] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, 2000, pp. 1-15.

[237] C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, and A. C. Culhane, "Dimension reduction techniques for the integrative analysis of multi-omics data," *Briefings in bioinformatics,* vol. 17, pp. 628-641, 2016.

[238] G. Tini, L. Marchetti, C. Priami, and M.-P. Scott-Boyer, "Multi-omics integration—a comparison of unsupervised clustering methodologies," *Briefings in bioinformatics,* vol. 20, pp. 1269-1279, 2019.

[239] A. C. Yeh, H. Li, Y. Zhu, J. Zhang, G. Khramtsova, K. Drukker, *et al.*, "Radiogenomics of breast cancer using dynamic contrast enhanced MRI and gene expression profiling," *Cancer Imaging,* vol. 19, p. 48, 2019.

[240] J. Klughammer, B. Kiesel, T. Roetzer, N. Fortelny, A. Nemc, K.-H. Nenning, *et al.*, "The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space," *Nature medicine,* vol. 24, pp. 1611-1624, 2018.

[241] P. Luo, Y. Li, L.-P. Tian, and F.-X. Wu, "Enhancing the prediction of disease–gene associations with multimodal deep learning," *Bioinformatics,* vol. 35, pp. 3735-3742, 2019.

[242] W. H. Land Jr, L. Wong, D. McKee, T. Masters, F. Anderson, and S. Sarvaiya, "Data fusion of several support-vector-machine breast-cancer diagnostic paradigms using a GRNN oracle," in *Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications 2004*, 2004, pp. 423-430.

[243] D. H. Wolpert, "Stacked generalization," *Neural networks,* vol. 5, pp. 241-259, 1992.

[244] Y. Wang, D. Wang, N. Geng, Y. Wang, Y. Yin, and Y. Jin, "Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection," *Applied Soft Computing*, vol. 77, pp. 188-204, 2019.

[245] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *Journal of machine learning research,* vol. 12, pp. 2211-2268, 2011.

[246] M. Tao, T. Song, W. Du, S. Han, C. Zuo, Y. Li*, et al.*, "Classifying breast cancer subtypes using multiple kernel learning based on omics data," *Genes,* vol. 10, p. 200, 2019.

[247] A. Rahimi and M. Gönen, "Discriminating early-and late-stage cancers using multiple kernel learning on gene sets," *Bioinformatics,* vol. 34, pp. i412-i421, 2018.

[248] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research,* vol. 9, pp. 2491-2521, 2008.

[249] K. Kourou, C. Papaloukas, and D. I. Fotiadis, "Integration of pathway knowledge and dynamic Bayesian networks for the prediction of oral cancer recurrence," *IEEE journal of biomedical and health informatics,* vol. 21, pp. 320-327, 2016.

[250] K. Kourou, G. Rigas, C. Papaloukas, M. Mitsis, and D. I. Fotiadis, "Cancer classification from time series microarray data through regulatory Dynamic Bayesian Networks," *Computers in Biology and Medicine,* vol. 116, p. 103577, 2020.

[251] K. D. Kourou, V. C. Pezoulas, E. I. Georga, T. Exarchos, C. Papaloukas, M. Voulgarelis*, et al.*, "Predicting Lymphoma Development by Exploiting Genetic Variants and Clinical Findings in a Machine Learning-Based Methodology With Ensemble Classifiers in a Cohort of Sjögren's Syndrome Patients," *IEEE Open Journal of Engineering in Medicine and Biology,* vol. 1, pp. 49-56, 2020.

[252] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," *PLoS computational biology,* vol. 8, 2012.

[253] K. Kourou, G. Rigas, K. P. Exarchos, C. Papaloukas, and D. I. Fotiadis, "Prediction of oral cancer recurrence using dynamic Bayesian networks," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 5275-5278.

[254] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences,* vol. 98, pp. 5116-5121, 2001.

[255] G. Wu, E. Dawson, A. Duong, R. Haw, and L. Stein, "ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis," *F1000Research,* vol. 3, 2014.

[256] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics,* vol. 27, pp. 431-432, 2011.

[257] K. Kourou, C. Papaloukas, and D. I. Fotiadis, "Gene-based pathway enrichment analysis of oral squamous cell carcinoma patients," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI),* 2016, pp. 360-363.

[258] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat*, et al.*, "The reactome pathway knowledgebase," *Nucleic acids research,* vol. 48, pp. D498-D503, 2020.

[259] T. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan*, et al.*, "Human protein reference database—2009 update," *Nucleic acids research,* vol. 37, pp. D767-D772, 2009.

[260] C. Ratineau, M. W. Petry, H. Mutoh, and A. B. Leiter, "Cyclin D1 represses the basic helix-loop-helix transcription factor, BETA2/NeuroD," *Journal of Biological Chemistry,* vol. 277, pp. 8847-8853, 2002.

[261] Y. A. Cho, J. S. Hong, E. J. Choe, H. J. Yoon, S. D. Hong, J. I. Lee*, et al.*, "The role of p300 in the tumor progression of oral squamous cell carcinoma," *Journal of Oral Pathology & Medicine,* vol. 44, pp. 185-192, 2015.

[262] P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez*, et al.*, "Pathway and network analysis of cancer genomes," *Nature methods,* vol. 12, p. 615, 2015.

[263] M. A. De Bastiani, B. Pfaffenseller, and F. Klamt, "Master regulators connectivity map: a transcription factors-centered approach to drug repositioning," *Frontiers in Pharmacology,* vol. 9, p. 697, 2018.

[264] H.-S. Chiu, S. Somvanshi, E. Patel, T.-W. Chen, V. P. Singh, B. Zorman*, et al.*, "Pan-cancer analysis of lncRNA regulation supports their targeting of cancer genes in each tumor context," *Cell reports,* vol. 23, pp. 297-312. e12, 2018.

[265] S. Davis and P. S. Meltzer, "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor," *Bioinformatics,* vol. 23, pp. 1846-1847, 2007.

[266] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi*, et al.*, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic acids research,* vol. 43, pp. e47-e47, 2015.

[267] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics,* pp. 1165-1188, 2001.

[268] F. Kolpakov, V. Poroikov, G. Selivanova, and A. Kel, "GeneXplain—identification of causal biomarkers and drug targets in personalized cancer pathways," *Journal of biomolecular techniques: JBT,* vol. 22, p. S16, 2011.

[269] V. Matys, E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, *et al.*, "TRANSFAC®: transcriptional regulation, from patterns to profiles," *Nucleic acids research,* vol. 31, pp. 374-378, 2003.

[270] J. Koschmann, A. Bhar, P. Stegmaier, A. E. Kel, and E. Wingender, "“Upstream analysis”: an integrated promoter-pathway analysis approach to causal interpretation of microarray data," *Microarrays,* vol. 4, pp. 270-286, 2015.

[271] F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, *et al.*, "Ensembl 2015," *Nucleic acids research,* vol. 43, pp. D662-D669, 2015.

[272] A. E. Kel, P. Stegmaier, T. Valeev, J. Koschmann, V. Poroikov, O. V. Kel-Margoulis, *et al.*, "Multi-omics “upstream analysis” of regulatory genomic regions helps identifying targets against methotrexate resistance of colon cancer," *EuPA open proteomics,* vol. 13, pp. 1-13, 2016.

[273] N. Friedman and D. Koller, "Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks," *Machine learning,* vol. 50, pp. 95-125, 2003.

[274] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, *et al.*, "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic acids research,* p. gkw937, 2016.

[275] S. Haider, C. Yao, V. Sabine, M. Grzadkowski, V. Stimper, M. H. Starmans, *et al.*, "Network-based biomarkers enable cross-disease biomarker discovery," *bioRxiv,* p. 289934, 2018.

[276] F. E. Froeling, C. Feig, C. Chelala, R. Dobson, C. E. Mein, D. A. Tuveson, *et al.*, "Retinoic acid–induced pancreatic stellate cell quiescence reduces paracrine Wnt–β-catenin signaling to slow tumor progression," *Gastroenterology,* vol. 141, pp. 1486-1497. e14, 2011.

[277] R. Kamps, R. D. Brandão, B. J. Bosch, A. D. Paulussen, S. Xanthoulea, M. J. Blok, *et al.*, "Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification," *International journal of molecular sciences,* vol. 18, p. 308, 2017.

[278] C. P. Mavragani and H. M. Moutsopoulos, "Sjögren syndrome," *Cmaj,* vol. 186, pp. E579-E586, 2014.

[279] A. Nezos, E. Gkioka, M. Koutsilieris, M. Voulgarelis, A. G. Tzioufas, and C. P. Mavragani, "TNFAIP3 F127C Coding Variation in Greek Primary Sjogren’s Syndrome Patients," *Journal of immunology research,* vol. 2018, 2018.

[280] S. Fragkioudaki, C. P. Mavragani, and H. M. Moutsopoulos, "Predicting the risk for lymphoma development in Sjogren syndrome: an easy tool for clinical use," *Medicine,* vol. 95, 2016.

[281] C. Baldini, P. Pepe, N. Luciano, F. Ferro, R. Talarico, S. Grossi*, et al.*, "A clinical prediction rule for lymphoma development in primary Sjögren's syndrome," *The Journal of rheumatology,* vol. 39, pp. 804-808, 2012.

[282] V. C. Pezoulas, K. D. Kourou, F. Kalatzis, T. P. Exarchos, A. Venetsanopoulou, E. Zampeli*, et al.*, "Medical data quality assessment: On the development of an automated framework for medical data curation," *Computers in biology and medicine,* vol. 107, pp. 270-283, 2019.

[283] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel*, et al.*, "Scikit-learn: Machine learning in Python," *Journal of machine learning research,* vol. 12, pp. 2825-2830, 2011.

[284] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in neural information processing systems*, 2013, pp. 431-439.

[285] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis,* vol. 38, pp. 367-378, 2002.

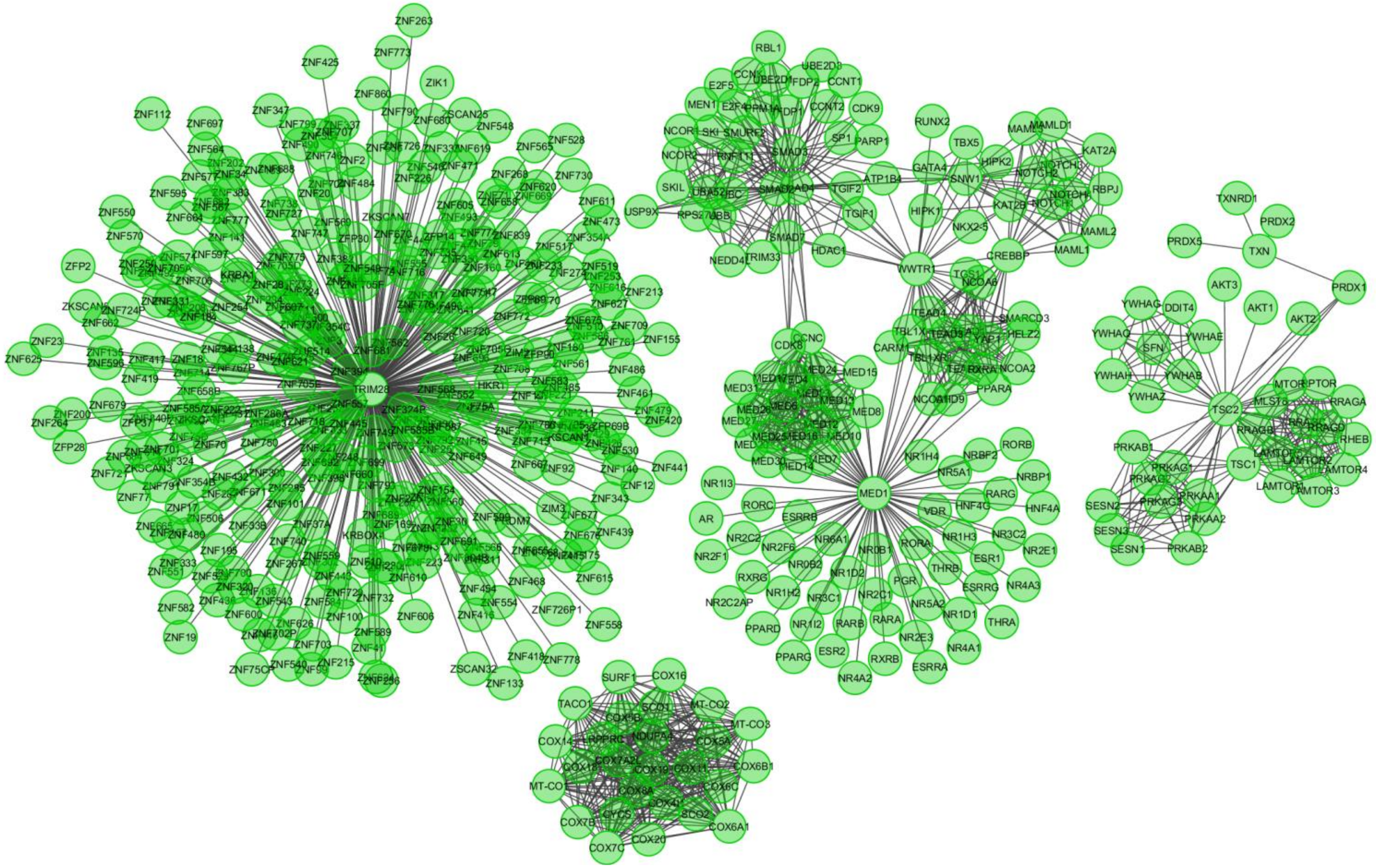
[286] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *The Journal of Machine Learning Research,* vol. 18, pp. 559-563, 2017.

[287] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," *University of California, Berkeley,* vol. 110, p. 24, 2004.

[288] A. Papageorgiou, D. C. Ziogas, C. P. Mavragani, E. Zintzaras, A. G. Tzioufas, H. M. Moutsopoulos*, et al.*, "Predicting the outcome of Sjogren's syndrome-associated non-Hodgkin's lymphoma patients," *PloS one,* vol. 10, 2015.

APPENDIX I

The functional interaction network of the GT pathway. Circles correspond to genes with their respective edges/connections.



APPENDIX II

The top-ranking master regulators obtained from the upstream analysis for the genes considered in each dataset.

GSE5462

Master molecule name	**Ranks sum**	**Master molecule name**	**Ranks sum**	**Master molecule name**	**Ranks sum**
NR1(h)	30	**LKB1-isoform1(h){ace K48}**	67	**PtdIns(3,5)P2**	99
AMPKbeta-2-isoform1(h)	31	**CLAN(h)**	68	**Bcl-x(h)**	99
PKAc(h):GSK3alpha(h)	32	**14-3-3eta(h)**	69	**Apaf-1:dATP**	100
PKAc(h):GSK3beta(h)	32	**granzymeB(h)**	69	**PtdIns(5)P**	102
AMPKalpha-2(h)	32	**ML-IAP(h)**	70	**proCaspase-3(h){pS150}**	103
AMPKbeta-2(h)	32	**calpain-4(h)**	70	**Septin4-isoform1(h)**	103
AMPKbeta-1(h)	32	**Cytochrome C:(Apaf-1)**	71	**PtdIns(3,4)P2**	105
PKCzeta{pT410}:PIP3	32	**c-FLIP-L(h)**	72	**PtdIns(3)P**	109
AMPKbeta-1(h)	33	**Pyk2-isoform1(h)**	74	**PtdIns(4)P**	111
LKB1(h)	34	**PHD3(h)**	74	**VEGF-A(h)**	124
PKACA(h)	38	**AKT(v.s.){pT308**	74		
thrombin(h)	41	**LCMT(h)**	74		
LKB1-isoform2(h)	46	**Hip-1(h)**	76		

thrombin(h)	49	**proCaspase-3(h)**	77
D1(h)	50	**SK1(h){p}**	77
CAMKKB(h)	51	**(Caspase-10)2**	77
Caspase-1(h)	52	**CaMKK-alpha-isoform1(h)**	78
DR4(h)	52	**cIAP-2(h)**	79
zVAD-fmk	53	**Ubc5B-isoform2(h)**	80
NR2B(h)	54	**FADD(h)**	81
CAMKKB-isoform2(h)	54	**Aven(h)**	81
glycine	56	**HGF(h)**	82
EGF:ErbB1[261]:ErbB2[261]	57	**GPRK5(h)**	86
CaMKK-alpha(h)	58	**TNFR1(h)**	88
14-3-3tau(h)	59	**Caspase-9-p35(h)**	92
p70S6K1-Alpha1(h){pT252}{pT412}	59	**Smac-isoform2(h)**	92
sphingosine	60	**Apaf-1L(h)**	93
LKB1-isoform1(h)	60	**dATP**	94
tPA(h)	64	**Caspase-9-p10(h)**	94

GSE14426

Master molecule name	**Ranks sum**	**Master molecule name**	**Ranks sum**	**Master molecule name**	**Ranks sum**
cyclinB1(h)	78	**Chfr-**	138	**Cdc23(h)**	162

		isoform2(h)			
ErbB2-isoform2(h)	88	**cyclinB:Cdk1{pT161}**	140	**cyclosome(h)**	166
MEG2(h)	102	**cyclosome{p}n:Cdc20{p}**	140	**TFIIH-CAK(h)**	169
acpp(h)	103	**cyclosome{p}n:Fzr:MAD2B**	140	**PARP(h)**	172
C/EBPalpha-isoform1(h)	104	**cyclinB1(h):Cdk1(h)**	142	**LOK(h){p}**	179
pkmyt1-isoform1(h)	106	**Fzr1(h)**	143	**NIPA(h)**	180
ErbB2-isoform1(h)	108	**MyoD(h)**	145	**Chk1(h)**	181
Gwl-isoform1(h)	109	**Cdc25A1(h)**	145	**CDP(h)**	181
Chk1(h){pS280}	109	**Cdc25A2(h)**	145	**ErbB2(h)[261]**	182
cyclosome(h):Cdc20(h)	110	**Cdk1(h){pY15}**	146	**Wee1(h):14-3-3beta(h)**	183
RASSF1-A(h)	110	**E2-C(h)**	146	**ErbB4(h)[261]**	184
cyclinB2(h)	110	**ANAPC2(h)**	148	**USP44(h)**	184
Cdc25C-isoform1(h)	111	**Cdk1(h){p}**	148	**Chk2(h)**	185
ErbB2(h)	114	**cyclinB:Cdk1{pY15}**	150	**LOK(h)**	185
CTAK1(h)	116	**cyclinA:Cdk2{pY15}**	151	**ErbB3(h)[261]**	186
Cdc25B-isoform3(h)	117	**cyclinA:Cdk1{pY15}**	152	**Cdc14A2(h)**	187
PKACA(h){pT198}	119	**Pin1(h)**	155	**ErbB1(h)[261]**	188
CaMKII{p}	120	**Skp1(h):NIPA**	156	**alpha5-integrin**	190
cyclinB1-isoform1(h)	123	**Roc1(h)**	156	**C/EBPalpha(h)**	193

CBL-3L(h)	125	**p57Kip2(h)**	157	**plk4-isoform1(h)**	193
Cdk1(h){pT14}{pY15}	125	**Cdc25C(h)**	158	**plk2(h)**	193
UBP41-isoform1(h)	128	**Cdc25C(h){p}**	159	**plk4-isoform2(h)**	193
APC11(h)	131	**cyclosome(h):Cdc20(h){ub}n**	159	**Cdc25B(h)**	195
E2-C-isoform1(h)	131	**cyclinB2(h)**	159	**tPA(h)**	197
UBP41(h)	132	**APC1(h):**	159	**plk3(h)**	199
cyclosome:Cdc20{ub}n	133	**Cdc25B(h){p}**	159	**NEK11-isoform1(h)**	200
(MAD2(h))4	136	**cyclosome(h)**	160	**cyclosome(h)**	201
DNA-PKcs-isoform1(h)	136	**Cdc23(h)**	160	**Wee1(h)**	201
NR1(h)	137	**Cdc25B(h){p}**	161	**ErbB1-p170(h)**	203
pkmyt1(h)	138	**Cdc25C(h){p}**	161	**PKAc(h):GSK3alpha(h)**	204

Master molecule name	**Ranks sum**	**Master molecule name**	**Ranks sum**	**Master molecule name**	**Ranks sum**
PKAc(h)	204	**Chk2-isoform1(h)**	223	**Mps1-isoform1(h)**	259
Cdk1-isoform1(h):cyclinB1	204	**p38beta(h){p}**	224	**FOXP3(h)**	260
Wee1-isoform1(h)	204	**Cdk1-isoform1(h)**	227	**DP97(h)**	260
NEK11(h)	204	**HIP14-xbb1(h)**	227	**Lyn(h){pY508}**	260
Cdc14A(h)	206	**APC1(h)**	228	**E2F-1(h)**	261
APC10(h)	209	**CLAN(h)**	229	**Chk2(h){pT68}**	261
Caspase-1(h)	210	**Cdc16(h)**	230	**Aurora-A(h){pT288}**	261

HAUSP(h)	210	**APC5(h)**	231	**Chk1(h){pS317}{pS345}**	261
Cdc25A(h){pS115}{pS320}	210	**APC7(h)**	231	**ATF-3(h)**	262
SLK-isoform2(h)	210	**IFNalpha8(h)**	231	**RPN-II(h)**	262
cyclinB(h){p}:Cdk1(h)	212	**IFNalpha6(h)**	231	**NEK11-isoform2(h)**	262
ErbB1(h){ub}n	213	**IFNalpha5(h)**	232	**Lyn(h){pY397}{pY508}**	262
cyclinB(h):Cdk1(h)	214	**Cdc23(h)**	233	**c-Cbl(h)**	264
LOK(h)	214	**ANAPC16(h)**	233	**AKT(h){p}**	264
plk1(h)	215	**Chk1-isoform1(h)**	234	**Lyn(h){pY397}**	264
ErbB4(h){ub}n	215	**IFNalpha4(h)**	234	**beta-TrCP1(h)**	267
Cdc27(h)	216	**LynA(h)**	235	**rnf11(h)**	268
plk2(h)	216	**cyclinA(h):Cdk1(h)**	236	**TRF2(h)**	270
plk4(h)	216	**LynB(h)**	237	**pot1(h)**	272
Chk2-xbb12(h)	216	**huntingtin(h)**	237	**E-cadherin(h)**	276
Chk1-isoform2(h)	216	**IFNalpha2(h)**	237	**FOXO6(h)**	276
Chfr(h)	217	**PARP(h)**	239	**securin(h)**	277
plk1(h)	218	**Fzr1-isoform2(h)**	247	**MUC4(h)**	279
DNA-PKcs(h)	218	**Raf-1(h){p}**	248	**CamKII(h)**	280
SLK(h)	218	**Raf-1(h){p}**	248	**acpp-isoform2(h)**	280
Cdk1(h)	220	**proCaspase-2(h)**	250	**GSK3alpha(h)**	282
plk3(h)	220	**proCaspase-6(h)**	253	**BARD1(h):brca1(h)**	283
huntingtin(h)	220	**Pin1(h)**	256	**DDB1(h)**	285
Cdc20(h)	222	**DNA-PKcs(h){p}**	257	**brca1(h):BARD1**	285

Lyn(h)	223	**beta-TrCP2(h)**	258	**(h)** **Caspase-8-p18(h)**	286

Master molecule name	**Ranks sum**	**Master molecule name**	**Ranks sum**	**Master molecule name**	**Ranks sum**
Caspase-8-p10(h)	286	**E1{ub(1)}**	300	**cyclinA1soform2**	327
Aurora-A{pT288}	287	**E1(h){ub(1)}**	302	**pim1**	329
BARD1(h):brca1-isoform1(h)	287	**E1:UbcH7**	302	**pim1-isoform**	329
PP2Cgamma	290	**E1:Ubc5A**	302	**APC4(h)**	339
Cdc25A(h){p}	290	**Cas(h)**	303	**p31-comet(h)**	346
IKK-alpha	291	**E1{ub(1)}**	304	**AMPKalpha-1-isoform1(h)**	356
B55A(h)	291	**p53-isoform1(h)**	306	**AMPKalpha-1-isoform2(h)**	356
Tome-1(h)	292	**p53-isoform4(h)**	306	**Daxx-isoform1(h)**	359
Cdc25A(h){p}	292	**p53-isoform2(h)**	306	**cyclinA(h):Cdk2**	384
HEF1(h)	295	**p27Kip1(h)**	309	**Cks1(h)**	390
E1:Ubc7	297	**cyclinB1(h)**	318	**p107(h)**	408
HIP2(h)	298	**mmp2(h)**	319	**SCF-Skp2(h)**	420

GSE37182

Master molecule name	**Ranks sum**	**Master molecule name**	**Ranks sum**	**Master molecule name**	**Ranks sum**
RSK2(h)	15	**Septin4-isoform1(h)**	103	**Cytochrome C**	137
XIAP(h)	49	**p70S6K1(h)**	106	**c-FLIP-S(h)**	138

XIAP(h)	52	**RSK1-isoform2(h)**	109	**Bcl-x(h)**	139
RSK3(h)	52	**beta1-integrin(h)**	110	**CD19(h)**	139
granzymeB(h)	52	**Fyn(h)[261]**	111	**cyclinE(h):Cdk2 (h)**	140
RSK3-isoform	55	**ERK5(h){p}**	112	**ERK(h){p}**	142
granzymeB(h)	69	**CARD4-isoform1(h)**	113	**Caspase-9(h)**	142
Cytochrome C(h)	72	**(angiotensin II)**	115	**RSK2(h)**	145
Caspase-3-p12(h)	76	**(proCaspase-9(h))2**	117	**ATR(h)**	146
Caspase-3-p17(h)	76	**TFF1(h)**	117	**BAP1(h)**	147
cIAP-2(h)	76	**MSK1-isoform1(h)**	117	**CUL4A(h)**	148
Caspase-10a(h)	77	**zVAD-fmk**	118	**p38alpha-CSBP1(h)**	148
Caspase-8a(h)	79	**cIAP-2(h)**	118	**p38alpha-CSBP2(h)**	148
usp13-isoform1(h)	79	**proCaspase-3(h){pS150}**	118	**p38alpha-Mxi2(h)**	148
usp13(h)	79	**E2F-7(h)**	118	**Apaf-1:dATP**	152
Src(h)[261]	80	**E2F-8(h)**	118	**apollon(h)**	153
Apaf-1(h)	86	**PRL3(h)**	120	**MSK1(h)**	155
Cytochrome C(h)	86	**Cytochrome C(h)**	122	**Cytochrome C(h)**	156
proCaspase-3(h)	93	**Fzr1(h)**	126	**XAF1(h)**	162
Smac-alpha(h)	94	**rrp1b(h)**	127	**ATM(h)**	164
Smac(h)	97	**Omi(h)**	128	**Caspase-3(h)**	165
Cytochrome C	98	**Smac-isoform2(h)**	128	**RSK2(h)**	165

Smac-delta(h)	98	**NADE(h)**	131	**PP2A(h)**	166
Aven(h)	99	**G{re}**	132	**Caspase-3-p20(h)**	166
Apaf-1L(h)	99	**Caspase-9-p10(h)**	133	**Caspase-3-p19(h)**	166
AKT(v.s.){pT	99	**TFIIH-CAK(h)**	134	**p38alpha(h)**	166
(Caspase-10)2	100	**Caspase-9-p35(h)**	135	**Csk(h)**	167
p38beta1(h)	101	**CARD4(h)**	137	**PP2Calpha1(h)**	167
ERK4(h)	101	**proCaspase-3(h):Hsp10**	137	**Wip1(h)**	167
dATP	103	**ML-IAP(h)**	137	**plk3(h)**	167

Master molecule name	**Ranks sum**	**Master molecule name**	**Ranks sum**	**Master molecule name**	**Ranks sum**
RSK1(h)	168	**DDB1(h)**	206	**Fbw5(h)**	258
CamKII(h)	168	**DR4(h)**	206	**FBXO9(h)**	258
Caspase-6(h)	169	**CaMKK**	207	**ZUBR1(h)**	258
IFNalpha1(h)	170	**c-FLIP-L(h)**	211	**MIB1(h)**	258
plk2(h)	170	**OTB1(h)**	221	**WDR48(h):USP1(h)**	258
MAK(h)	170	**FAF1(h)**	222	**alpha4-integrin(h)**	260
DNA-PKcs(h)	171	**USP28**	222	**Ubc13**	262
Murr1(h)	171	**USP28**	222	**ATM(h){pS1981**	263
plk4(h)	172	**IFNalpha8(h)**	224	**mgat5(h)**	263
(Caspase-8(h))2	173	**IFNalpha6(h)**	225	**LRR1(h)**	273
p38beta(h)	175	**BAP1(h)**	225	**Bcl-2(h)**	276
TIEG-1(h)	176	**IFNalpha5(h)**	226	**TLR9(h){ub}n**	287
apollon(h)	176	**Trx1(h)[80]2**	227	**Pyk2**	293
plk3(h)	178	**PTEN(h)**	235	**ML-IAP1(h)**	295
RSK1(h){pS221}	179	**IFNalpha2a(h)**	237	**hBre1(h)**	295

Hip-1(h)	180	**rnf5(h)**	238	**CDCA7L(h)**	295
IFNalpha4(h)	180	**Rad-18(h)**	239	**RARRES1(h)**	299
TOSO(h)	180	**FANCL**	240	**ML-IAP2(h)**	302
IFNalpha2(h)	183	**Fbx11(h): Cul-1(h)**	241	**IFNbeta(h)**	314
IMP-1(h)	185	**KPC1**	242	**ATM(h)**	328
USP28(h)	190	**ing4(h)**	243	**ATM(h)**	328
proCaspase-10(h)	193	**THBS2(h)**	248		
Fzr1-isoform2(h)	193	**RNF168(h)**	252		
p38beta(h){p}	193	**vcam1(h)**	253		
OGT(h)	195	**EULIR(h)**	253		
BAP1(h):ASXL1	195	**march8**	254		
DNA-PKcs	197	**brca1-isoform2(h)**	255		
PTEN(h)	198	**rnf5(h)**	256		
nanog(h)	202	**cyclinE2(h)**	257		
RelA-p65(h)	203	**RNF8(h)**	257		

APPENDIX III

CLASSIFICATION RESULTS FOR THE MICROARRAY DATASET OF GSE14426 PANCREATIC CANCER STUDY

Differentially Expressed Genes (Limma analysis)		**Master Regulators (GeneXplain Upstream Analysis)**						**Differentially Expressed Genes with Master Regulators**					
27 DEGs		10 MRS		15 MRS		20 MRS		27 DEGs & 10 MRs		27 DEGs & 15 MRs		27 DEGs & 20 MRs	
Acc	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***
65,00%	70,00%	50,00%	43,33%	50,00%	51,67%	55,00%	51,67%	71.66%	78.33%	71,68%	85,56%	73.33%	82.22%

CLASSIFICATION RESULTS FOR THE MICROARRAY DATASET OF GSE37182 COLON CANCER STUDY

Differentially Expressed Genes (Limma analysis)		**Master Regulators (GeneXplain Upstream Analysis)**						**Differentially Expressed Genes with Master Regulators**					
7 DEGs		10 MRS		15 MRS		20 MRS		7 DEGs & 10 MRs		7 DEGs & 15 MRs		7 DEGs & 20 MRs	
Acc	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***
97,14%	99,49%	99,28%	98,54%	97,14%	96,53%	97,85%	98,47%	95,71%	98,78%	97,14%	99,59%	98,57%	98,57%

CLASSIFICATION RESULTS FOR THE MICROARRAY DATASET OF GSE5462 BREAST CANCER STUDY

Differentially Expressed Genes (Limma analysis)		**Master Regulators (GeneXplain Upstream Analysis)**						**Differentially Expressed Genes with Master Regulators**					
19 DEGs		10 MRS		15 MRS		20 MRS		19 DEGs & 15 MRs		19 DEGs & 10 MRs		19 DEGs & 20 MRs	
Acc	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***
65,76%	45,84%	62,69%	45,26%	62,69%	51,53%	61,53%	49,91%	68,84%	50,85%	68.46%	49.55%	70,77%	56,29%

AUTHOR’S PUBLICATIONS

PUBLICATIONS RELATED TO THIS THESIS

Journal Articles

[1] **Kourou, K**., Pezoulas, V. C., Georga, E. I., Exarchos, T., Papaloukas, C., Voulgarelis, M., Goules A., Nezos A., Tzioufas A. G., Moutsopoulos H. M., Mavragani, C., & Fotiadis, D. I. (2020). Predicting lymphoma development by exploiting genetic variants and clinical findings in a machine learning-based methodology with ensemble classifiers in a cohort of Sjögren's Syndrome patients. IEEE Open Journal of Engineering in Medicine and Biology.

[2] **Kourou, K**., Rigas, G., Papaloukas, C., Mitsis, M., & Fotiadis, D. I. (2019). Cancer classification from time series microarray data through regulatory Dynamic Bayesian Networks. Computers in Biology and Medicine, 116, 103577.

[3] **Kourou, K**., Papaloukas, C., & Fotiadis, D. I. (2017). Integration of pathway knowledge and dynamic Bayesian networks for the prediction of oral cancer recurrence. IEEE journal of biomedical and health informatics, 21(2), 320-327.

[4] **Kourou, K**., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

[5] **Kourou, K**., & Fotiadis, D. I. (2015). Computational modelling in cancer: methods and applications. Biomed Data J, 1(1).

Conference Papers

[1] **Kourou, K**., Kondylakis, H., Koumakis, L., Manikis, G.C., Marias, K., Tsiknakis, M., Simos, P.G., Karademas, E. and Fotiadis, D.I., 2019, September. Computational Models for Predicting Resilience Levels of Women with Breast Cancer. In Mediterranean Conference on Medical and Biological Engineering and Computing (pp. 518-525). Springer, Cham.

[2] **Kourou, K**., Papaloukas, C., Mitsis, M., & Fotiadis, D. I. (2018). Assessing the Predictive Value of Regulatory Molecules for Patient Outcome in Pancreatic Cancer: A Computational Approach. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1307-1310). IEEE.

[3] **Kourou, K**., Papaloukas, C., & Fotiadis, D. I. (2018). Modeling Biological Data Through Dynamic Bayesian Networks for Oral Squamous Cell Carcinoma Classification. In World Congress on Medical Physics and Biomedical Engineering 2018 (pp. 375-379). Springer, Singapore.

[4] **Kourou, K.**, Papaloukas, C., & Fotiadis, D. I. (2017). Identification of differentially expressed genes through a meta-analysis approach for oral cancer classification. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 3876-3879). IEEE.

[5] **Kourou, K**., Papaloukas, C., & Fotiadis, D. I. (2017). A computational pipeline for deciphering the molecular mechanisms of oral cancer progression. In 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) (pp. 209-212). IEEE.

[6] **Kourou, K.**, Rigas, G., Exarchos, K. P., Papaloukas, C., & Fotiadis, D. I. (2016). Prediction of oral cancer recurrence using dynamic Bayesian networks. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 5275-5278). IEEE.

[7] **Kourou, K.,** Papaloukas, C., & Fotiadis, D. I. (2016). Gene-based pathway enrichment analysis of oral squamous cell carcinoma patients. In 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 360-363). IEEE.

[8] **Kourou, K.**, Exarchos, K. P., Papaloukas, C., & Fotiadis, D. I. (2015). A Bayesian Network-based approach for discovering oral cancer candidate biomarkers. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 7663-7666). IEEE.

OTHER PUBLICATIONS

Journal Articles

[1] **Kourou, K**., Pezoulas, V. C., Georga, E. I., Exarchos, T., Tsanakas, P., Tsiknakis, M. & Fotiadis, D. I. (2018). Cohort Harmonization and Integrative Analysis from a Biomedical Engineering Perspective. IEEE reviews in biomedical engineering.

[2] **Kourou, K**., Rigas, G., Exarchos, K. P., Goletsis, Y., Exarchos, T. P., Jacobs, S. & Fotiadis, D. I. (2016). Prediction of time dependent survival in HF patients after VAD implantation using pre-and post-operative data. Computers in biology and medicine, 70, 99-105.

Conference Papers

[1] Manikis, G. C., **Kourou, K**., Poikonen-Saksela, P., Kondylakis, H., Karademas, E., Marias, K., Katehakis, D. G., Koumakis, L., Kourabali, A., Pat-Horenczyk, R., Tsiknakis, M., Simos, P., & Fotiadis, D. I. (2019, October). Computational modeling of psychological resilience trajectories during breast cancer treatment. In 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 423-427). IEEE.

[2] Exarchos, K. P., **Kourou, K**., Exarchos, T. P., Papaloukas, C., Karamouzis, M. V., & Fotiadis, D. I. (2015). Sequence patterns mediating functions of disordered proteins. In GeNeDis 2014 (pp. 49-59). Springer, Cham.

SHORT CV

Konstantina D. Kourou, M.Sc., is a bioinformatician, with experience in the analysis of high-throughput datasets and the development of computational methods for predicting disease status and outcomes. She is currently working on the development of machine learning predictive models of cancer for the translation of patient data to precision diagnosis and treatment.

She received the B.Sc. degree from the Department of Informatics, Ionian University, and holds an M.Sc. in Bioinformatics from the King's College London University. She is currently a Ph.D. candidate at the University of Ioannina specializing on the development of machine learning models and their applications in cancer prognosis and prediction. Since 2013, she has been a member of the Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina.

Google Scholar profile:

https://scholar.google.com/citations?user=papOzucAAAAJ&hl=en