

# Declaration

I, Manaswita Saikia, hereby declare that the thesis entitled "**Biomarker Identification for Critical Diseases using Machine Learning Techniques**" submitted to the Department of Computer Science and Engineering under the School of Engineering, Tezpur University, in partial fulfillment for the award of the degree of Doctor of Philosophy in Computer Science and Engineering is a bona-fide work carried out by me. The results presented in this thesis have not been submitted in part or in full, to any other University or Institute for the award of any degree or diploma.

Date:

Place: Tezpur University, Napaam, Tezpur

(Manaswita Saikia)

Reg. no. TZ155525 of 2015

Enrollment No. CSP16105



# Tezpur University

## Certificate

This is to certify that the thesis entitled "**Biomarker Identification for Critical Diseases using Machine Learning Techniques**" submitted to the Department of Computer Science and Engineering under the School of Engineering, Tezpur University, in partial fulfillment for the award of the degree of Doctor of Philosophy in Computer Science and Engineering is a record of research work carried out by **Ms. Manaswita Saikia** under our supervision and guidance.

All helps received by her from various sources have been duly acknowledged. No part of this thesis has been submitted elsewhere for award of any other degree.

Signature of Supervisor

Signature of Co-supervisor

(Prof. Dhruva K. Bhattacharyya)

(Prof. Jugal K. Kalita)

Professor

Professor

Computer Science and Engineering

Department of Computer Science

School of Engineering

College of Engineering and Applied Science

Tezpur University

University of Colorado

Napaam, Tezpur

Colorado Springs, Colorado, USA

# Acknowledgement

The tenure of my research has been an intense learning experience which does not merely extend over research skills, but also inculcates a number of social skills at the personal level. It gives me immense pleasure to take this opportunity to express my deep sense of gratitude to my esteemed supervisor Prof. Dhruba K. Bhattacharyya. He gave freedom to pursue my ideas and work at my own pace, and was always available to discuss various problems on the way. His constant support, trust, valuable feedback, encouragement, innumerable advice and guidance have provided a good basis for completion of my research work. It is my privilege to thank my co-supervisor Prof. Jugal K. Kalita, for his invaluable feedback, suggestions and guidance in shaping my research papers and PhD.the thesis.

I would like to acknowledge Dr. Bhabesh Nath, and Dr. Rosy Sharmah, members of my doctoral research committee for their valuable suggestions and feedback throughout the period of the work. I also convey my heartiest thanks to all members of the faculty, Department of Computer Science and Engineering for their constructive suggestions and encouragements in this journey.

I am grateful to the authorities of Tezpur University and the Department of Computer Science and Engineering for providing me with the facilities during the pursuit. The support will always be remembered.

The blessings, untiring moral support and constant encouraging words of my parents, family and friends boosted me enough to carry out my research work up to this level. They are the pillars behind this accomplishment. This note of acknowledgment can never be complete without a mention to members of my extended family at Tezpur University especially my dear friends, Trishna Barman, Parthajit Borah, Upasana Sarmah, Dr. Nilakshi Devi and Dr. Hussain A. Chowdhury, for their help throughout the work.

Finally, I would like to thank all those who have directly or indirectly helped me in different capacities to complete my work.

# LIST OF TABLES

2.1	Summary of the microarray datasets, GSE20347 and GSE23400, and the bulk RNA-Seq dataset, GSE130078 for ESCC . . . . .	35
2.2	Summary of the scRNA-Seq dataset, GSE160269 for ESCC . . . . .	37
3.1	Eight chosen biclustering methods: A Comparison . . . . .	57
3.2	Summary of the biclusters detected by BicGenesis in all three datasets. . . . .	70
3.3	Subset of Normal and Disease Biclusters . . . . .	74
3.4	Preservation analysis of modules in the microarray datasets, GSE20347 and GS23400, and the bulk RNA-Seq dataset, GSE130078 . . . . .	77
3.5	Top 20 hub-genes for each extracted MoI in the two microarray and one bulk RNA-Seq datasets . . . . .	79
3.6	Percentages of genes in each MoT that are annotated to the Gene Ontology (GO) databases and KEGG pathways. . . . .	82
3.7	Summary of BCGs detected by BicGenesis in the microarray dataset, GS20347, that are annotated to top 3 GO terms in the three GO databases. . . . .	86
3.8	Summary of BCGs detected by BicGenesis in the microarray dataset, GSE23400, that are annotated to top 3 GO terms in the three GO databases. . . . .	88
3.9	Summary of BCGs detected by BicGenesis in the bulk RNA-Seq dataset, GS130078, that are annotated to top 3 GO terms in the three GO databases . . . . .	89
3.10	Summary of BCGs detected by BicGenesis in the two microarray and one bulk RNA-Seq datasets that have been annotated to the top 5 KEGG enriched pathways . . . . .	90
3.11	Summary of potential biomarkers identified by BicGenesis . . . . .	96
3.12	Summary of potential ESCC biomarkers identified by BicGenesis using the biomarker criteria . . . . .	100
4.1	DE methods for Microarray and bulk RNA-Seq data. . . . .	112
4.2	Summary of detected DEGs by the three RNA-Seq methods and the three microarray methods for three datasets . . . . .	120
4.3	Preservation analysis of modules detected by our Integrative DEA method in the two microarray and one RNA-Seq datasets . . . . .	126
4.4	Top 20 hub-genes for each extracted MoI in the two microarray and one RNA-Seq datasets . . . . .	127

4.5	Percentages of genes in each MoI that are annotated to the Gene Ontology (GO) databases and KEGG pathways. . . . .	129
4.6	DEGs that are annotated to most enriched GO term in all three GO databases (BP, CC and MF) as well as the most enriched pathway . . . .	130
4.7	Summary of BCGs detected by Integrative DEA in the microarray dataset, GS20347, that are annotated to top 3 GO terms in the three GO databases.	133
4.8	Summary of BCGs detected by Integrative DEA in the microarray dataset, GSE23400, that are annotated to top 3 GO terms in the three GO databases.	134
4.9	Summary of BCGs detected by Integrative DEA in the bulk RNA-Seq dataset, GS130078, that are annotated to top 3 GO terms in the three GO databases . . . . .	135
4.10	Summary of BCGs detected by our method, Integrative DEA, in the three datasets . . . . .	136
4.11	Summary of potential biomarkers identified by our framework, Integrative DEA . . . . .	142
4.12	Summary of potential ESCC biomarkers identified by Integrative DEA using the biomarker criteria . . . . .	145
4.13	Comparison of our method, Integrative DEA with two recent works that employ DEA on ESCC datasets . . . . .	147
5.1	Centrality Measures for hub-gene finding employed in CBDCEM . . . .	157
5.2	Comparison of the seven centrality measures employed by CBDCEM . .	159
5.3	Symbols used in proposed Hub-gene finding algorithm . . . . .	163
5.4	Preservation Analysis of Modules in the two microarray datasets, GSE20347 and GS23400, and the bulk RNA-Seq dataset, GSE130078 . . . . .	175
5.5	Top 20 hub genes for each extracted module of interest in all three datasets using our hub-gene finding algorithm. . . . .	176
5.6	Percentage of genes in each module of two microarray and one bulk RNA-Seq datasets that are annotated in the GO databases and KEGG pathways. . . . .	178
5.7	Summary of hub-genes detected by CBDCEM in GSE20347 , and GSE23400 annotated to the top 20 KEGG enriched pathways . . . . .	181
5.8	Summary of hub-genes detected by CBDCEM in GS130078 annotated to the top 20 KEGG enriched pathways . . . . .	182
5.9	Summary of hub-genes detected by CBDCEM in GS20347, and GSE23400 annotated to top GO terms in the three GO databases . . . . .	183
5.10	Summary of hub-genes detected by CBDCEM in bulk RNA-Seq dataset, GS130078 annotated to top GO terms in the three GO databases . . . .	185
5.11	Summary of potential biomarkers identified by CBDCEM. . . . .	188

5.12	Summary of potential ESCC biomarkers identified by CBDCEM using the biomarker criteria . . . . .	190
5.13	Summary of potential biomarkers detected by CBDCEM and four other hub-gene finding methods . . . . .	192
5.14	Summary of performance of CBDCEM vs. four other methods in terms of proportion of modules . . . . .	198
6.1	Centrality Measures for hub-gene finding employed in scDiffCoAM . . . . .	205
6.2	Comparison of the four of the seven measures employed by scDiffCoAM. . . . .	207
6.3	Preservation Analysis ( $Z_{\text{summary}}$ ) of CD45+ modules in CD45- dataset and vice versa . . . . .	222
6.4	Top 20 hub genes for each extracted MoIs in CD45+ and CD45- datasets using our hub-gene finding algorithm . . . . .	227
6.5	Percentage of genes in each MoI that are annotated in the GO databases and KEGG pathways . . . . .	230
6.6	Summary of hub-genes detected by scDiffCoAM that have been annotated to the Top 20 KEGG enriched pathways in the CD45+ cell types . . . . .	231
6.7	Summary of hub-genes detected by scDiffCoAM that have been annotated to the Top 20 KEGG enriched pathways in CD45- cell types . . . . .	232
6.8	Summary of hub-genes detected by scDiffCoAM that have been annotated to the top enriched GO terms in the three GO databases for CD45+ cell types . . . . .	233
6.9	Summary of hub-genes detected by scDiffCoAM that have been annotated to the top enriched GO terms in the three GO databases for CD45- cell types . . . . .	235
6.10	Summary of potential biomarkers candidates identified by scDiffCoAM. Here, All 3 under GO databases imply all three databases, BP, CC, and MF. . . . .	247
6.11	Summary of potential ESCC biomarkers identified by scDiffCoAM using the biomarker criteria . . . . .	251
6.12	Summary of potential biomarkers detected by scDiffCoAM and three other hub-gene finding methods . . . . .	254
6.13	Ranking all potential biomarkers for ESCC identified by all four proposed frameworks . . . . .	261

# LIST OF FIGURES

1.1	Steps involved in the Central Dogma of molecular biology. . . . .	2
2.1	Pre-processing pipeline employed by our proposed frameworks for the three types of gene expression data . . . . .	37
3.1	Biclustering Approaches: A Taxonomy . . . . .	43
3.2	Proposed Biclustering Analysis Framework . . . . .	61
3.3	Pipeline for DCA . . . . .	64
3.4	Hierarchical trees for normal and disease in the microarray datasets GSE20347 and GSE23400, and bulk RNA-Seq dataset GSE130078 for ESCC . . . .	71
3.5	Soft thresholds for normal and disease in the microarray datasets GSE20347 and GSE23400, and bulk RNA-Seq dataset GSE130078 for ESCC . . . .	72
3.6	Dendrograms for normal and disease in the microarray datasets GSE20347 and GSE23400, and bulk RNA-Seq dataset GSE130078 for ESCC . . . .	75
3.7	<i>Zsummary</i> plots for normal and disease in the microarray datasets GSE20347 and GSE23400, and bulk RNA-Seq dataset GSE130078 for ESCC . . . .	76
3.8	GRN for normal modules a) <i>skyblue</i> and b) <i>white</i> in GSE20347, disease modules c) <i>yellowgreen</i> , d) <i>white</i> e) <i>salmon4</i> , and f) <i>purple</i> in GSE20347	84
3.9	GRN for normal modules a) <i>brown2</i> in GSE23400 and b) <i>lightcyan</i> in GSE130078 . . . . .	85
3.10	GRN for disease module <i>orange</i> in GSE130078 . . . . .	85
4.1	Proposed Integrative Differential Expression Analysis Framework . . . .	115
4.2	Pipeline for DCA . . . . .	117
4.3	Dendrograms for normal and disease in the microarray datasets, GSE20347 and GSE23400, and bulk RNA-Seq dataset, GSE130078 for ESCC . . . .	123
4.4	<i>Zsummary</i> plots for normal and disease in the microarray datasets GSE20347 and GSE23400, and bulk RNA-Seq dataset GSE130078 for ESCC . . . .	125
4.5	GRN for normal module a) <i>pink</i> and disease modules b) <i>greenyellow</i> in GSE20347, disease modules c) <i>darkgreen</i> , d) <i>lightsteelblue1</i> e) <i>black</i> in GSE20347. GRN for disease module f) <i>magenta</i> in GSE23400. . . . .	131
4.6	GRN for normal modules a) <i>purple</i> and b) <i>greenyellow</i> in GSE20347, and disease modules c) <i>blue</i> in GSE23400. GRN for disease modules d) <i>lightyellow</i> e) <i>violet</i> , and f) <i>steelblue</i> in GSE130078. . . . .	132

5.1	Proposed Centrality Based DCA Framework, CBDCEM . . . . .	161
5.3	Dendrograms for normal and disease in the microarray datasets GSE20347 and GSE23400, and bulk RNA-Seq dataset GSE130078 for ESCC . . . .	171
5.4	Heiarchical trees for module detection for normal and disease in the microarray datasets GSE20347 and GSE23400, and bulk RNA-Seq dataset GSE130078 for ESCC . . . . .	172
5.5	$Z_{summary}$ plots for normal and disease in the microarray datasets GSE20347 and GSE23400, and bulk RNA-Seq dataset GSE130078 for ESCC . . . .	173
5.6	GRN for normal module a) <i>paleturquoise</i> , disease modules b) <i>darkturquoise</i> and c) <i>orange</i> in GSE20347. GRN for disease module d) <i>grey60</i> in GSE23400 . . . . .	179
5.7	GRN for disease module a) <i>lightcyan</i> , d) <i>tan</i> e) <i>green</i> in GSE23400, and disease module f) <i>salmon</i> in GSE130078. . . . .	180
6.1	Steps involved in WGCNA analysis for high dimensional data using hd-WGCNA . . . . .	208
6.2	Proposed framework for DCA on scRNA-Seq Dataset, scDiffCoAM . . . .	210
6.3	Violin Plots for CD45+ and CD45- . . . . .	216
6.4	Elbow Plots for CD45+ and CD45- . . . . .	217
6.5	Soft Thresholds for three CD45+ and three CD45- cell types . . . . .	219
6.6	Dendrograms for three CD45+ cell types and one CD45- cell type . . . .	220
6.7	Dendrograms for two CD45- cell types . . . . .	221
6.8	$Z_{summary}$ plot for Tcell (CD45+) in Epithelial (CD45-) and Endothelial (CD45-) and vice versa. . . . .	223
6.9	$Z_{summary}$ plot for Tcell (CD45+) in Fibroblast (CD45-) and Bcell (CD45+) in Epithelial (CD45-) and Endothelial (CD45-), and vice versa. . . . .	224
6.10	$Z_{summary}$ plot for Bcell (CD45+) in Fibroblast (CD45-) and Myeloid (CD45+) in Epithelial (CD45-) and Endothelial (CD45-), and vice versa. . . . .	225
6.11	$Z_{summary}$ plot for Myeloid (CD45+) in Fibroblast (CD45-) and vice versa. . . . .	226
6.12	GRN for two modules in CD45+ cell type, Tcell. . . . .	237
6.13	GRN for one, two, and two modules in CD45+ cell types, Tcell, Bcell, and Myeloid and one module CD45- cell types, Epithelial . . . . .	238
6.14	GRN for three and one modules in CD45- cell types, Endothelial and Fibroblast . . . . .	239
6.15	GRN for module <i>turquoise</i> in CD45- cell type, Fibroblast . . . . .	239
6.16	Summary of performances of scDiffCoAM vs. four other methods . . . .	258



# Glossary

BC	Bcell
BCG	Biomarker Candidate Gene
CEN	Co-expression Network
CPM	Counts Per Million
DAVID	Database for Annotation, Visualisation, and Integrated Discovery
DCA	Differential Co-expression analysis
DCE	Differential Co-Expression
DCG	Differentially Co-expressed Gene
DEA	Differential Expression Analysis
DEG	Differentially Expressed Gene
DNA	Deoxyribonucleic Acid
EBAM	Empirical Bayes analysis of Microarrays
EN	Endothelial
EP	Epithelial
ESCC	Esophageal Squamous Cell Carcinoma
FDR	False Discovery Rate
FI	Fibroblast
GO	Gene Ontology
GO_BP	Gene Ontology Biological Processes
GO_CC	Gene Ontology Cellular Components
GO_MF	Gene Ontology Molecular Functions
GRN	Gene Regulatory Network
GSEA	Gene Set Enrichment Analysis
HNSCC	Head and Neck Squamous Cell Carcinoma
KEGG	Kyoto Encyclopedia of Genes and Genomes
LaSCC	Laryngeal Squamous Cell Carcinoma
lFDR	Local False Discovery Rate
lgEGo	List of Enriched GO Terms
lgEP	List of Enriched Pathways
LSCC	Lung Squamous Cell Carcinoma

MoI	Module of Interest
MY	Myeloid
MSR	Mean Squared Residue
OSCC	Oral Squamous Cell Carcinoma
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PEGoT	Percentage of Enriched GO Terms
PEP	Percentage of Enriched Pathways
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
SAM	Significance Analysis of Microarrays
SCC	Squamous Cell Carcinoma
scRNA-Seq	Single cell RNA-Sequencing
TC	Tcell
TED	Top Enriched DEG
TF	Transcription Factor
TG	Target Gene
TOM	Topological Overlap Matrix
TSCC	Tongue Squamous Cell Carcinoma
WGCNA	Weighted Gene Co-expression Network Analysis