# 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.

(Manaswita Saikia)

Date:

Reg. no. TZ155525 of 2015

Place: Tezpur University, Napaam, Tezpur

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. Dhruba 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

# LIST OF FIGURES

# 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 Componenets |
| GO_MF | Gene Ontology Molecular Funxtions |
| 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 | Princpal 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 |