

LB/TH/08/2023

DCS 03/48

OVERALL SURVIVAL PREDICTION OF GLIOMA PATIENTS USING GENOMICS

M. R. Navodini Wijethilake

208012F

LIBRARY
UNIVERSITY OF MORATUWA, SRI LANKA
MORATUWA

Thesis/Dissertation submitted in partial fulfillment of the requirements for the
degree Master of Science by Research

004*2021

004(043)

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

University of Moratuwa



TH5103

September 2021

TH5103

TH 5103

DECLARATION

I, Navodini Wijethilake, declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature ***UOM Verified Signature*** Date: 23/09/2021

The above candidate has carried out research for the Masters thesis Dissertation under my supervision.

Name of Supervisor: Dr. Dulani Meedeniya

Signature of the Supervisor: Date: 23/09/2021

The above candidate has carried out research for the Masters thesis/Dissertation under my supervision.

Name of Supervisor: Dr. Charith Chithraranjan

Signature of the Supervisor: ***UOM Verified Signature*** ²¹

ABSTRACT

Overall survival prediction is a vital task that will lead for better patient management in clinical practise. Existing approaches mainly focus on imaging based survival prediction, which is non invasive, and thus, easier to be implemented at the initial diagnosis stages. However, the advancements in the DNA/RNA technologies has given access to genomic and transcriptomic profiles of the gliomas, that directly reflect the molecular level alterations. Thus, in this work we mainly focus on using transcriptomic profiles for survival prediction, an area that has not been widely analysed yet for survival prediction. We utilize the gene expression and mutation profiles, while augmenting the recent Artificial Intelligence approaches, such as deep probabilistic programming and multi task learning for prognosis prediction. Thereby we do not just focus on the application, we also contribute with novel learning paradigms to improve the classification task performances. Nonetheless, we also focus on proposing a novel loss function, since architectural wise the state of art performance has been achieved for classification tasks.

In addition, we also investigate ability to employ radiomics, for subtype classification, that is also associated with survival. Since subtypes mainly rely on the genomic alterations, we found it useful to focus on imaging features ability to predict prognosis of glioma.

Keywords: Survival Prediction; Prognosis; Deep Learning; Multi-task Learning; Deep Probabilistic Programming

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisors Dr. Dulani Meedeniya and Dr. Charith Chithraranjan for the immense guidance provided to successfully finish this research. I'm extremely thankful for your patience, motivation, and immense knowledge given throughout the period. I firmly believe without your courageous support this would not be able to reach this stage.

I wish to thank Dr. Indika Perera for his valuable insights and guidance from the very beginning of this research. I would like to convey my gratitude to the entire staff of the Department of Computer Science and Engineering, both academic and non-academic for all their support given throughout the entire Masters course period. This research was supported by the University of Moratuwa Senate Research Grant. I would like to acknowledge the grant and other relevant parties who work hard to provide the required facilities to up bring the research facilities in Sri Lanka more specifically, by providing the financial support.

Nonetheless, I would like to thank my family for all the love and support.

Thank you!

LIST OF ABBREVIATIONS

GBM	Glioblastoma
DL	Deep Learning
ML	Machine Learning
IR	Importance Ranking
PCA	Principle Component Analysis
RFE	Recursive Feature Elimination
CC	Correlation Coefficient
RF	Random Forest
LR	Linear Regression
SVM	Support Vector Machine
XGB	eXtreme Gradient Boosting
ST	Survival Tree
ANN	Artificial Neural Network
LASSO	Least Absolute Shrinkage and Selection Operator
KM	Keplen Meier
TCGA	The Cancer Genome Atlas
CNN	Convolutional Neural Network
CGGA	Chinese Glioma Genome Atlas
OBTS	the Ohio Brain Tumor Study
GEO	Gene Expression Omnibus
CPTAC	Clinical Proteomic Tumor Analysis Consortium
TCIA	The Cancer Imaging Archive
LGG	Lower Grade Glioma
CoxPH	Cox Proportional Hazard model
CE	Cross-Entropy
FL	Focal Loss
LSCE	Label Smoothing Cross-Entropy
LSF	Label Smoothing Focal

normLSF Normalized Label Smoothing Focal
Acc. Accuracy
Prec. Precision
Sens. Sensitivity
MTL Multi Task Learning



TABLE OF CONTENTS

Declaration of the Candidate & Supervisor	i
Abstract	ii
Acknowledgement	iii
List of Abbreviations	iv
Table of Contents	vi
List of Figures	ix
List of Tables	xi
1 Introduction	2
1.1 Domain Overview	2
1.1.1 Classification based on underlying histology	2
1.1.2 Classification based on molecular pathogenesis	3
1.1.3 Mortality of Glioma patients	4
1.1.4 Overall Glioma survival and Glioma survival	5
1.2 Overview of the Problem	5
1.3 Motivation	6
1.4 Importance of this research	6
1.5 Objectives	7
1.6 Contribution	8
1.7 Publications	8
2 Literature Review	10
2.1 Data types used in Glioma survival analysis	10
2.1.1 Glioma Screening and Data Collection in clinical practice	10
2.1.2 Radiomics	11
2.1.3 Genomics	16
2.1.4 Other data types used in Survival Analysis	20
2.1.5 Public Glioma Cohorts	21
2.1.6 Preprocessing of Glioma survival related Data	23
2.2 Glioma Survival Analysis Approaches	25

2.2.1	Machine Learning based Survival Analysis	25
2.2.2	Deep Learning based Survival Analysis	35
2.2.3	Statistical Analysis tools for Survival Analysis	36
2.2.4	Other methods used in survival prediction	39
3	Methods	43
3.1	The Approaches included in this work	43
3.2	Prognosis prediction with Gene Expression profiles	44
3.2.1	Basic Machine Learning for Survival Prediction	44
3.2.2	Deep Learning for survival prediction	50
3.3	Prognosis prediction with Mutation profiles	54
3.3.1	Dataset	54
3.3.2	Methods	55
3.4	Prognosis with Radiogenomics	56
3.4.1	Dataset	56
3.4.2	Segmentation	59
3.4.3	Feature Extraction	59
3.4.4	Statistical Analysis	60
3.4.5	Subtype Predictive model	61
4	Results	62
4.1	Prognosis prediction with gene expression profiles	62
4.1.1	System Evaluation - ML approaches & Risk Score Model	62
4.1.2	System Evaluation - DL approach	68
4.2	Prognosis prediction with mutation profiles	75
4.3	Prognosis Analysis with Radiogenomics	76
4.3.1	Correlation between radiomics, genomics and overall survival	76
4.3.2	Imaging biomarkers associated with molecular subtypes	78
4.3.3	Subtype Prediction	79
5	Discussion	81
5.1	Comparison with existing studies	81
5.2	Limitations	86
5.3	Future Directions	87

6 Conclusions	89
References	91



LIST OF FIGURES

Figure 2.1	A general radiomics extraction pipeline.	11
Figure 2.2	Sample survival tree for survival group classification into short, medium, and long survival.	30
Figure 2.3	Artificial Neural Network consists of a single hidden layer.	34
Figure 2.4	A sample nomogram.	42
Figure 3.1	Graphical Abstract.	43
Figure 3.2	Proposed Bayesian Neural network architecture.	49
Figure 3.3	Survival Time Distribution (in days) into three class samples - short-term, medium-term, and long term for CGGA and TCGA dataset.	50
Figure 3.4	The 1D array containing the expressions levels of 13094 genes are transformed into a 2D array of 116 x 116.	51
Figure 3.5	MTL model inspired with Resnet18 architecture for survival class and grade prediction.	56
Figure 3.6	Step 1: Freeze the grade classification block and optimize survival prediction task.	57
Figure 3.7	Step 2: Freeze the survival classification block and the shared encoder, and optimize grade prediction task.	57
Figure 4.1	Heat map for the (a) CGGA cohort (b) validation TCGA cohort with proposed gene signature.	65
Figure 4.2	Overall survival distribution of high and low risk groups.	66
Figure 4.3	Kaplan-Meier curves obtained for the high risk and low risk groups. (a) CGGA (b) TCGA dataset.	67
Figure 4.4	mRNA expression value distribution of each selected genes for the high and low risk groups.	68
Figure 4.5	Comparative classification performance analysis for model trained with state-of-the-art cross entropy and proposed cost function on TCGA dataset.	71

Figure 4.6	Reliability graph for normLSF comparing with CE, FL and LSF. The curve for normLSF is much closer to the perfect reliability curve (dash line).	72
Figure 4.7	tSNE projections of penultimate layer features of ResNet18 model trained on Genomics data with various loss functions.	73
Figure 4.8	SHAP analysis	74
Figure 4.9	The most crucial clinical genes responsible for true and false prediction in red and blue, respectively of all three categories.	74
Figure 4.10	Performance comparison between using Label smoothing and Cross entropy loss with and without gradnorm.	76
Figure 4.11	The comparison of the fractal dimensions of whole tumor, tumor core and necrosis regions between 4 subtypes of Glioblastoma.	79

LIST OF TABLES

Table 2.1	Open source Tools available for Preprocessing and Segmentation of MRI	13
Table 2.2	Segmentation methods and imaging features extracted in glioma related studies.	17
Table 2.3	Frequently used Glioma cohorts for survival analysis of Glioma patients. TCGA: The Cancer Genome Atlas; CPTAC: Clinical Proteomic Tumor Analysis Consortium; BraTS: Brain Tumor Segmentation; CGGA: Chinese Glioma Genome Atlas; GEO: Gene Expression Omnibus, OBTS: the Ohio Brain Tumor Study	22
Table 2.4	Feature selection and analyzing techniques used in survival prediction of gliomas with radiomics. IR: Importance Ranking; PCA: Principle Component Analysis; RFE: Recursive Feature Elimination; CC: Correlation Coefficient; RF: Random Forest; LR: Linear Regression; SVM: Support Vector Machine; XGB: eXtreme Gradient Boosting; ST: Survival Tree; ANN: Artificial Neural Network	26
Table 2.5	Feature selection and analysing techniques used in survival prediction of gliomas with radiogenomics. RFE: Recursive Feature Elimination; LASSO: Least Absolute Shrinkage and Selection Operator; KM: Keplern Meier; LR: Linear Regression; SVM: Support Vector Machine	26
Table 2.6	Related work with DL approaches. TCGA: The Cancer Genome Atlas; CNN: Convolutional Neural Network	36
Table 2.7	Related work on prognostic risk score calculation. TCGA: The Cancer Genome Atlas; CGGA: Chinese Glioma Genome Atlas; GEO: Gene Expression Omnibus; CPTAC: Clinical Proteomic Tumor Analysis Consortium; TCIA: The Cancer Imaging Archive; LGG: Lower Grade Glioma, GBM: Glioblastoma, Cox PH: Cox Proportional Hazard model	41



Table 2.8	Related work on prognostic nomograms development. TCGA: The Cancer Genome Atlas; CPTAC: Clinical Proteomic Tumor Analysis Consortium; GEO: Gene Expression Omnibus, OBTS: the Ohio Brain Tumor Study, LGG: Lower Grade Glioma; GBM: Glioblastoma	42
Table 3.1	Dataset Description	45
Table 3.2	Selected 7 genes and their corresponding posterior probabilities	46
Table 3.3	Parameters of the subtype prediction learning models	61
Table 4.1	Comparison of Overall Survival Prediction with Machine Learning - 4 fold cross validation on CGGA cohort	62
Table 4.2	Comparison of Overall Survival Prediction with Machine Learning - testing on TCGA cohort	63
Table 4.3	Univariate Cox Regression analysis on the chosen 7 genes	64
Table 4.4	Performance study of state-of-the-art CNN models trained with cross-entropy and proposed loss for classifying of 4 cross-validated CGGA genomic data samples. CE, normLSF, Acc., Prec., Sens. represents Cross-Entropy, Normalized Label Smoothing Focal, Accuracy, Precision and Sensitivity respectively.	69
Table 4.5	Average classification metrics for validation-set obtained with different loss functions on a 4-fold CGGA dataset. The values manifested in bold are best per column. CE, FL, LSCE, LSF, normLSF, Acc., Prec., Sens. represents Cross-Entropy, Focal Loss, Label Smoothing Cross-Entropy, Label Smoothing Focal, Normalized Label Smoothing Focal, Accuracy, Precision and Sensitivity respectively.	70
Table 4.6	4-fold cross evaluation metrics of ResNet18 model trained on CGGA Genomics dataset using ours loss.	70

Table 4.7	Model miscalibration quantification for CE, FL, LSF and proposed normLSF loss. Evaluation metrics such as Expected Calibration Error (ECE), Static Calibration Error (SCE), Thresholded Adaptive Calibration Error (TACE), Brier Score (BS) and Uncertainty Calibration Error (UCE) are used to calculate the calibration error for each model.	72
Table 4.8	Prediction of subtype and grade as a multi task classification using mutation profiles on CGGA dataset. Step 1. freeze grade prediction branch(Label Smoothing) Step 2. freeze shared encoder – survival prediction branch (Label Smoothing) Step 3. both unfreeze - GradNorm	75
Table 4.9	Prediction of subtype and grade as a multi task classification using mutation profiles on TCGA dataset. Step 1. freeze grade prediction branch(Label Smoothing) Step 2. freeze shared encoder – survival prediction branch (Label Smoothing) Step 3. both unfreeze - GradNorm	75
Table 4.10	Kruskal Wallis statistics for radiomic features ($p < 0.05$) between 4 subtypes.	78
Table 4.11	Prediction of subtypes as a binary classification using Radiomics	80
Table 4.12	Prediction of Molecular subtypes as a 4 class classification with Genomics	80
Table 5.1	Comparison with existing best performing works.	86