Title?

SIANA RIZWAN	180042105
FARZANA TABASSUM	180042119
SABRINA ISLAM	180042122

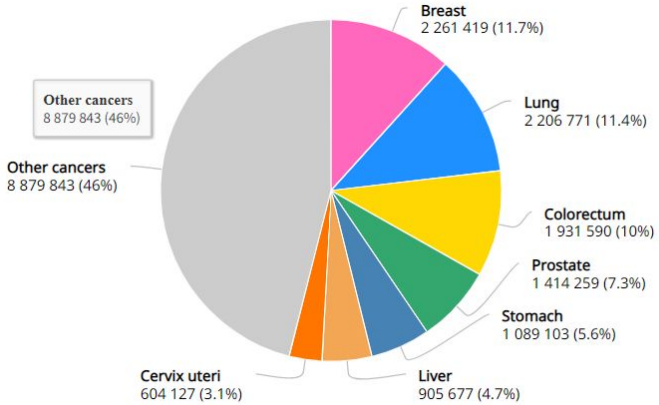
Under the supervision of:

Tareque Mohmud Chowdhury
Assistant Professor
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT), OIC

Contents

Introduction

- Cancer is a genetic disease which involves abnormal growth and proliferation of cells in the body.
- Being a leading cause of death worldwide, cancer accounts for nearly 10 million deaths in 2020.
- By 2040, the number of new cancer cases per year is expected to rise to 29.5 million and the number of cancer-related deaths to 16.4 million.



Total : 19 292 789

Motivation

- Cancers having fewer specific and sensitive biomarkers makes it difficult for traditional diagnosis methods to detect early.
- Traditional diagnosis methods (Biopsy and physical examinations, X-rays, CT scans, MRIs, etc.) are time-consuming and expensive.
- Traditional methods may struggle to handle the growing volume of medical data, particularly in the era of big data and precision medicine.

Literature Review

1. Cancer Type Prediction and Classification Based on RNA-sequencing Data *by Hsu, Y.H.; Si, D. In 2018 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*
2. Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection *by Lopez-Rincon A, Martinez-Archundia M, Martinez-Ruiz GU, Schoenhuth A, Tonda A. in BMC Bioinform. 2019*
3. PanClassif: Improving pan cancer classification of single cell RNA-seq gene expression data using machine learning *by Mahin, K.F.; Robiuddin, M.; Islam, M.; Ashraf et. el. in Genomics 2022.*
4. A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction *by Khalsan, M.; Machado, L.R.; Al-Shamery, E.S et. el. IEEE Access 2022.*
5. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data *by Li Y, Kang K, Krahn JM, Croutwater N et. el. in BMC Genomics 2018.*
6. Explainable Machine Learning to Identify Patient-specific Biomarkers for Lung Cancer *by M. Sobhan and A. M. Mondal.*
7. Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data *by García-Díaz P, Sánchez-Berriel I, Martínez-Rojas JA, et al. in Genomics 2020.*
8. A stacking ensemble deep learning approach to cancer type classification based on TCGA data *by Mohammed, M.; Mwambi, H.; Mboya, I.B.; Elbashir, M.K.; Omolo, B in Sci. Rep. 2021.*
9. Deep Learning to Discover Genomic Signatures for Racial Disparity in Lung Cancer *by M. Sobhan, A. Al Mamun, R. B. Tanvir, M. J. Alfonso in Proc. - 2020 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2020.*

Literature Review

Cancer Type Prediction and Classification Based on RNA-sequencing Data

– *Hsu, Yi-Hsin, and Dong Si. IEEE, 2018* *[2]*

Performance:

Testing Variables	Accuracy Score	Training Time	Ave. Precision	Ave. Recall	Ave. F1
DT	0.86014	23m 42s 121ms	0.86	0.86	0.86
kNN	0.89212	30s 751ms	0.90	0.89	0.89
Linear SVM	0.94988	~4hr	0.95	0.95	0.95
Poly SVM	0.76754	52m 52s 518ms	0.86	0.77	0.77
ANN	0.94797	18m 43s 312ms	0.95	0.95	0.95

Dataset:

TCGA Pan-Can: 33 types of cancer

Limitations:

1. Feature selection is not considered.
2. Mostly traditional machine learning models were used.

Literature Review

PanClassif: Improving pan cancer classification of single cell RNA-seq gene expression data using machine learning. – *Mahin, Kazi Ferdous, et al. Genomics 114.2 (2022) [3]*

Performance:

	Binary classification			**Multi-class** classification		
	105	204	571	105	204	571
KNN	1	1	1	1	1	0.99
RF	1	1	1	1	0.99	0.99
ANN	0.97	0.99	0.99	0.98	0.98	0.98

Dataset:

- TCGA Pan-Can: 22 types of cancer
- GEO: Breast cancer & Skin melanoma cancer

Limitations:

1. 22 types of cancer classification.
2. No mentions of any particular feature.
3. Patient specific feature selection is not considered.

Literature Review

Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection. – *Lopez-Rincon, Alejandro, et al. BMC bioinformatics 20.1 (2019) [4]*

Performance:

Classifier	TCGA	TT vs NT	GEO	TCGA (Subtype)	GEO (Subtype)	Global
Gradient Boosting	0.9359	0.9846	0.6697	0.9725	0.8909	0.8907
Random Forest	0.9324	0.9839	0.8085	0.9725	0.8634	0.9121
Logistic Regression	0.9237	0.9799	0.9351	0.9647	0.8476	0.9302
Passive Aggressive	0.8831	0.9606	0.8678	0.9556	0.8197	0.8974
SGD	0.9035	0.9767	0.9393	0.9490	0.8145	0.9166
SVC	0.9154	0.9791	0.7724	0.9451	0.8355	0.8895
Ridge	0.8305	0.9470	0.8867	0.9503	0.8300	0.8889
Bagging	0.9110	0.9812	0.7682	0.9555	0.9070	0.9046

Dataset:

- **TCGA Pan-Can**: 28 types of cancer
- **GEO**: 14 datasets of 5 different platforms to validate

Limitations:

1. 28 types of cancer classification.
2. Common set of features generalized for all types of cancers in the dataset.
3. Patient specific feature selection is not considered.

Our Hypothesis

Not all genes are responsible for all types of cancer.
Each cancer type is characterized by specific genetic alterations and molecular signatures.

By identifying the cancer-specific gene sets or biomarkers, researchers and clinicians can develop more accurate and targeted diagnostic tests and guide the development of targeted therapies.

Research Aims & Objectives

Aim 1

Propose a pipeline to classify 33 types of cancer with high accuracy.

Objectives

- Determining the performance on raw dataset.
- Evaluating the effects of data Normalization and Feature selection techniques on the performance.
- Ensembling to improve the performance.

Research Aims & Objectives

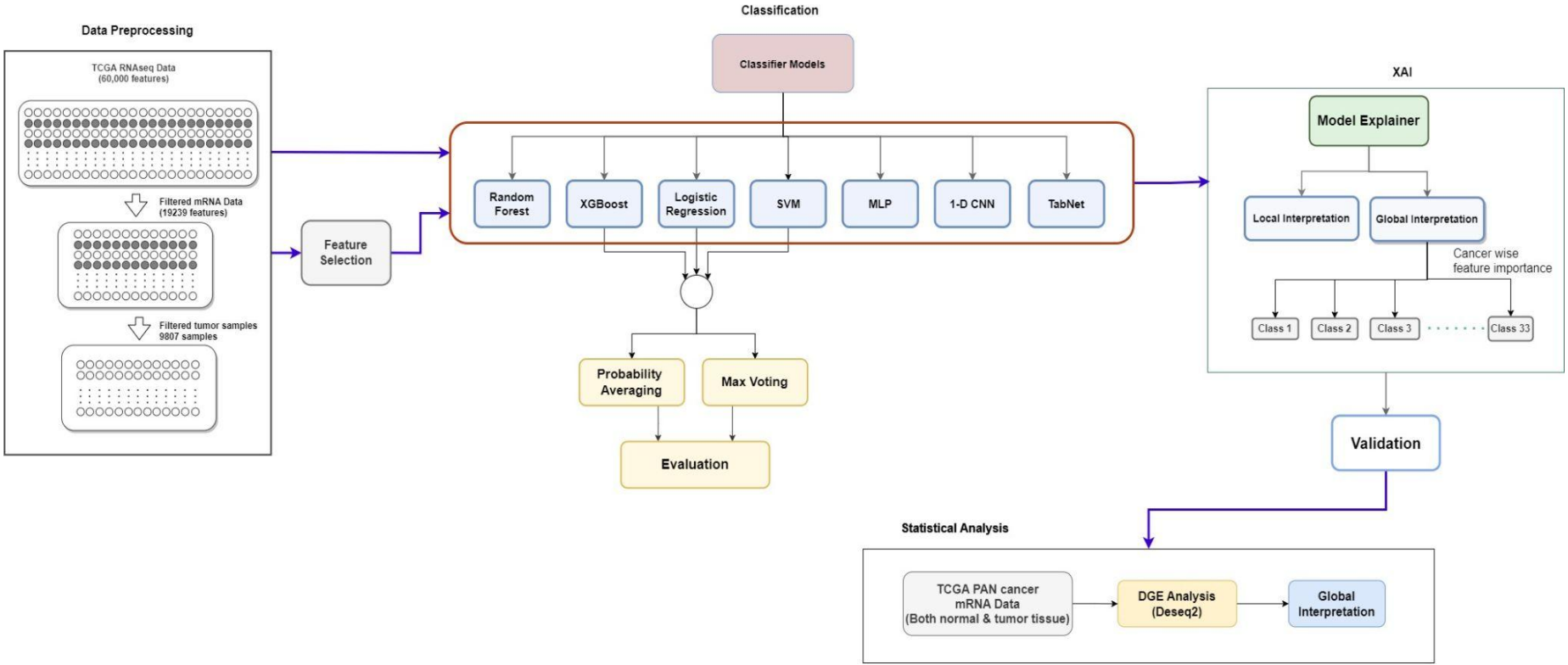
Aim 2

Identification of Cancer-Specific Important Gene Set.

Objectives

- ➢ Determining the feature contribution for each sample by applying Explainable machine learning models.
- ➢ Extracting the list of globally significant genes for each cancer types.
- ➢ Extracting patient-specific gene sets.
- ➢ Validating the gene sets by statistical analysis.

Proposed Pipeline



Dataset

TCGA Pan-Cancer Dataset

- Gene expression RNA-seq data (From UCSC Xena browser)
- Sample of 33 types of cancer
- Sample size: 10,535 and 19238 Features
- We have converted the gene ensemble ID to gene symbols.

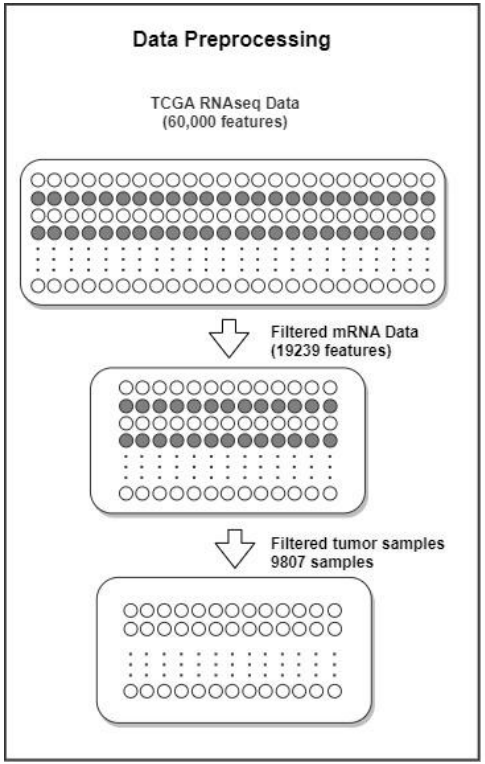
60,499 identifiers X 10535 samples All Identifiers All Samples

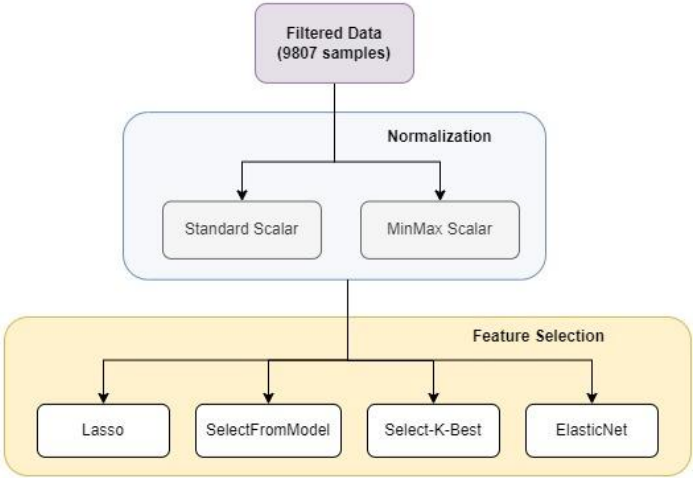
	TCGA-19-1787-01	TCGA-S9-A7J2-01	TCGA-G3-A3CH-11	TCGA-EK-A2RE-01	TCGA-44-6778-01	TCGA-F4-6854-01
ENSG00000000003.14	5.076	4.679	5.495	4.362	3.55	6.429
ENSG00000000005.5	2.431	-2.466	-3.626	-9.966	-9.966	-1.47
ENSG00000000419.12	4.766	4.005	4.141	5.512	4.822	6.365
ENSG00000000457.13	0.7664	1.647	1.345	1.736	2.39	1.975
ENSG00000000460.16	2.444	0.8246	-0.9132	2.516	2.144	2.406

	Unnamed: 0	RAB4B	TIGAR	RNF44	DNAH3	RPL23A	ARL8B	CALB2	DACH1	FMO2	...	ARHGAP21
0	TCGA-19-1787-01	4.8324	3.0411	3.7794	-5.5735	9.6898	5.1102	7.4223	-2.4659	1.0573	...	4.8929
1	TCGA-S9-A7J2-01	4.1962	1.6093	4.6888	-9.9658	9.0745	5.1285	2.1574	0.4761	-1.0262	...	5.6308
2	TCGA-G3-A3CH-11	3.3952	-0.0574	1.6695	-9.9658	8.2107	3.3407	-9.9658	-1.3183	-1.2481	...	1.9490
3	TCGA-EK-A2RE-01	3.9099	3.2722	3.1062	-2.1779	9.5378	4.7929	4.1962	-5.5735	-4.2934	...	2.8760
4	TCGA-44-6778-01	4.9031	3.1507	4.5862	-3.3076	9.3566	5.0313	-1.5105	1.7273	4.5142	...	3.7825
...	...	...	...	...	...	...	...	...	...	...	...	...
10530	TCGA-VQ-AA6F-01	4.8294	2.6255	4.9566	-0.0130	9.8004	4.0960	-0.7834	-0.6193	-0.8339	...	4.5681
10531	TCGA-BR-8588-01	3.7464	3.2251	3.9682	-1.2481	9.8467	4.9069	-3.0469	1.7617	2.0742	...	3.7432
10532	TCGA-24-2254-01	4.4810	1.6140	5.0700	-0.7588	9.4778	5.3234	3.0550	2.2663	1.1250	...	3.5584
10533	TCGA-DD-A115-01	3.7006	1.8282	2.9356	-4.2934	8.8128	3.9384	-5.5735	-3.6259	-0.3752	...	3.0859
10534	TCGA-FV-A3I0-11	2.8720	-0.7834	1.4756	-9.9658	8.0752	3.3535	-5.0116	-2.5479	-1.8314	...	1.9415

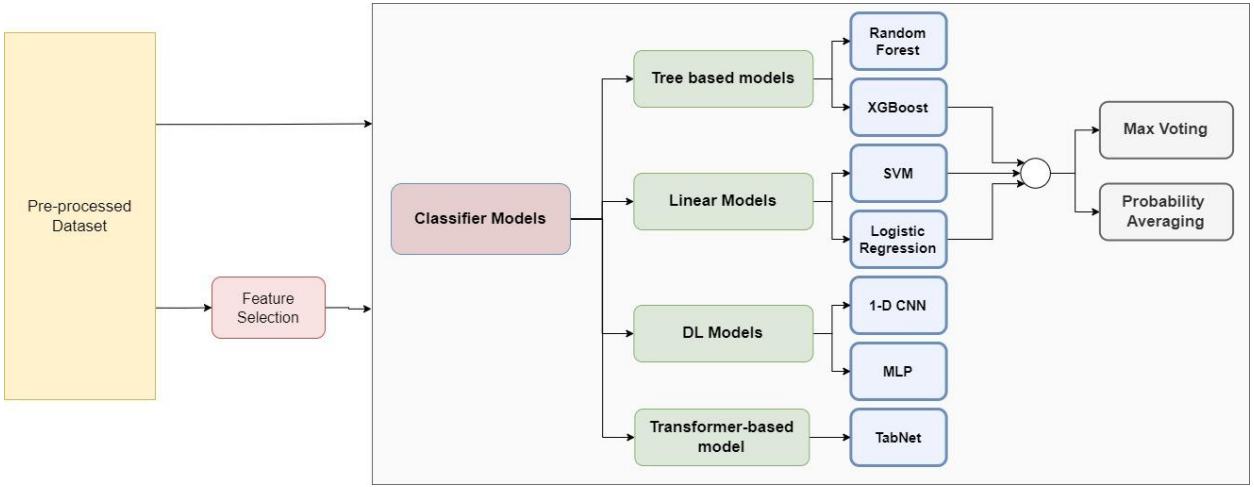
10535 rows × 19239 columns

Data Pre-processing



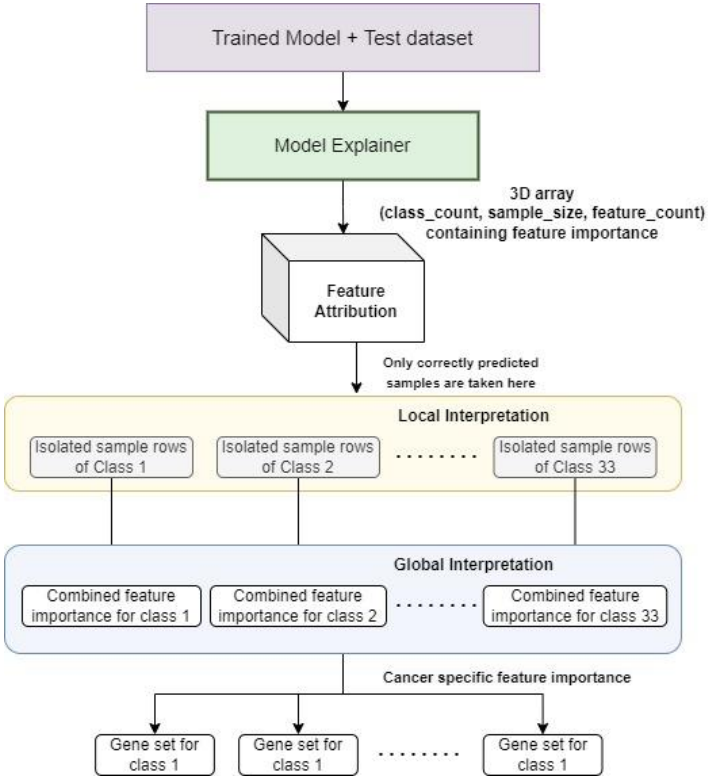


Classification



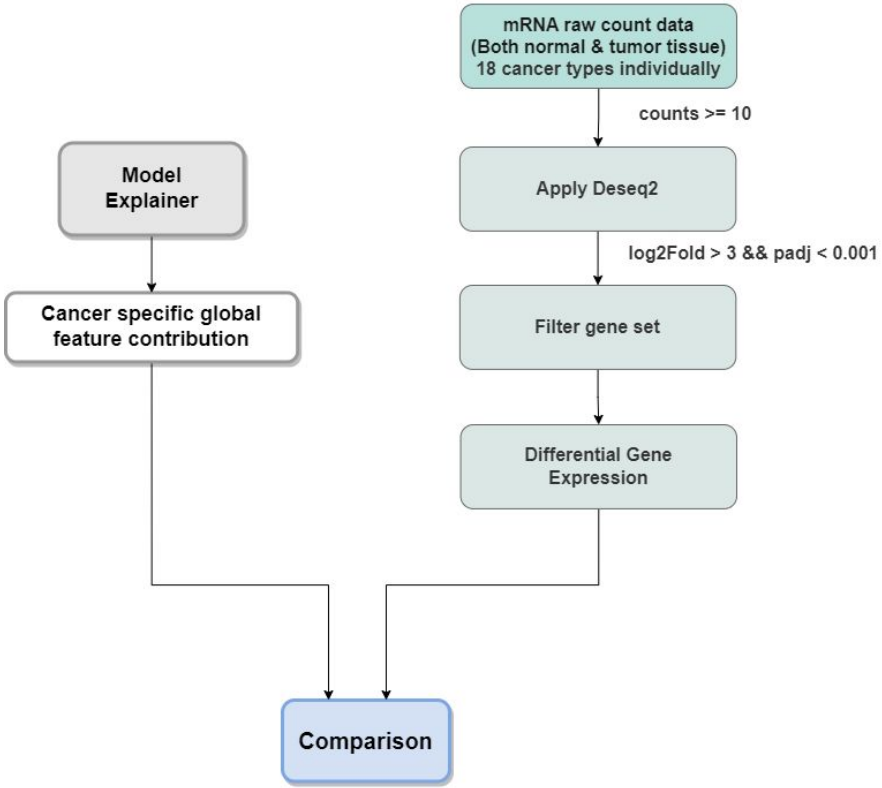
- We have applied this Classification architecture on the preprocessed dataset with 19k features.
- Also applied this Classification architecture on the datasets after applying different feature selection methods.

Explainability Analysis



- We have calculated the feature contribution for all models with 19k features.
- Then we have calculated the feature contribution for the models trained with feature selection (500 features).

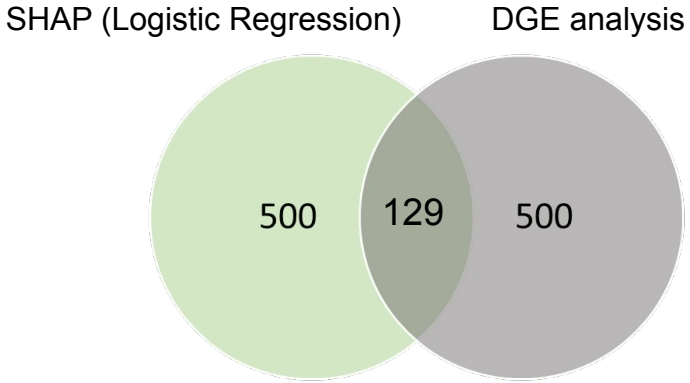
Comparison



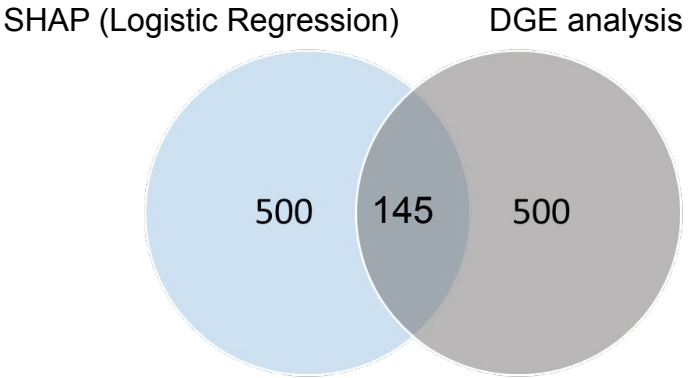
- Performed statistical DGE analysis using DESeq2 on 18 cancer types individually to get statistically significant gene set for each cancer types.
- Using the cancer specific feature importance from SHAP, identified the common set of genes for both top 500 features from 19k and the SFM 500 features.

Validation

For UCEC cancer, top 500 gene features from both statistical analysis and models trained on 19k features



For UCEC cancer, top 500 gene features from statistical analysis and all 500 features from models trained using SelectFromModel approach



Patient Specific Validation

Calculated patient specific feature contribution for 33 types of cancers

		0	1	2	3	4	5	6	7	8	9	...
0	0	TBX20	GUCA1A	PRR9	SPRR2A	HMX3	MYH14	LY6G6D	TFAP2B	PRSS33	TJP3	...
	1	TBX20	GUCA1A	PRR9	SPRR2A	HMX3	MYH14	LY6G6D	TFAP2B	PRSS33	TJP3	...
	2	TBX20	GUCA1A	PRR9	SPRR2A	HMX3	MYH14	LY6G6D	TFAP2B	PRSS33	TJP3	...
	3	TBX20	GUCA1A	OTOS	PRR9	SPRR2A	HMX3	MYH14	LY6G6D	TFAP2B	PRSS33	...
	4	TBX20	GUCA1A	OTOS	PRR9	SPRR2A	HMX3	MYH14	LY6G6D	TFAP2B	PRSS33	...
...	...	...	...	...	...	...	...	...	...	...	...	...
32	73	ACTG2	MSX1	LYPD8	GOLT1A	FCN2	XKRX	GGTLC3	EFNA2	FGF3	GUCY2C	...
	74	USH1C	PABPC3	PSG2	HOXB6	ERBB4	ITIH2	CXorf49	ACTG2	OCLN	DSC3	...
	75	CRABP1	NPY4R	TRIM40	CSF3	MOGAT3	SLC9A4	PABPC3	PYURF	AMY1B	H2BC15	...
	76	ITIH2	IAPP	NOTUM	BAAT	SOX14	NPY4R	MOK	FYB2	VIT	DMRT2	...
	77	HBB	WNT7A	FAT2	RAP1GAP	ANKRD1	NPPB	AIPL1	UBE2U	MMP15	WNK2	...

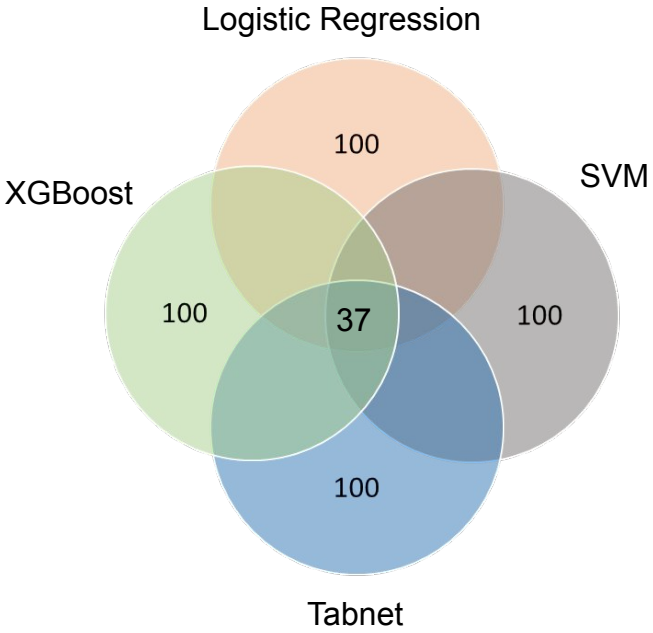


Fig: Common gene count for cancer BLCA

Conclusion

- We aim to propose an efficient pipeline for classifying 33 types of cancer based on gene expression data
- We extracted globally significant genes for each cancer type and locally significant genes for each patient
- The extracted genesets were validated using statistical tools to ensure appropriateness

Future Work

- Implementing SSGSEA to identify patient specific gene set for precision medicine.
- Patient specific pathway analysis.

References

[1]. Hsu, Yi-Hsin, and Dong Si. "Cancer type prediction and classification based on rna-sequencing data." 2018 *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018.

[2]. Mahin, Kazi Ferdous, et al. "PanClassif: Improving pan cancer classification of single cell RNA-seq gene expression data using machine learning." *Genomics* 114.2 (2022): 110264.

[3]. Lopez-Rincon, Alejandro, et al. "Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection." *BMC bioinformatics* 20.1 (2019): 1-17.

[4]. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.

[5]. Prokhorenkova, Liudmila, et al. "CatBoost: unbiased boosting with categorical features." *Advances in neural information processing systems* 31 (2018).

[6]. Jasim, Mahmood, et al. "A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction." *IEEE Access* (2022).

Thank You!