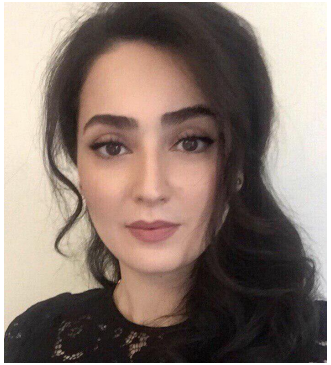
Cancer Classification through a Hybrid Machine Learning Approach

Elmira Amiri Souri*, Sophia Tsoka
*elmira.amiri@kcl.ac.uk

About me



Elmira Amiri is currently a PhD student of Bioinformatics in the Department of Informatics at King's College London. She got her M.Sc. degree in Information Technology and two B.Sc. degrees in Business Management and Information Technology. Her PhD research focuses on developing machine learning and analytical methods to extract biological 'meaning and pattern' from health-related data. The title of her PhD dissertation is "Machine learning based genes signature selection for cancer prediction, survival analysis and drug discovery''. She has recently completed developing a machine learning pipeline on genomics data that serves as a diagnostics factor of breast cancer metastasis risk predictor for classifying tumours and improving target treatment. In addition to cancer genomics research, she works on embedding-based machine learning models to predict drug-target interactions and perform drug discovery.

Motivation

- Identify cancer subtypes → simplifying the problem
- Increasing the depth of our knowledge in driver genes and oncogenic pathways
- Understanding of progression and formation of cancer
- Genes never act alone, methods that analyse a single gene are less efficient
- Extracting a new class of gene signatures that act as **mediators** in forming oncogenic pathways.

Introduction

- Cancer is a major cause of death (in 2014, about 14.1 million new cases a year)
- Early detection of cancer and understanding the biology of it can improve prognosis, however
  - **Heterogeneity** of tissues and genetics of patients
  - **High dimensionality** of the microarray data
  - **Combined effects** of genes lead to a variety of resultant phenotypes
- Using machine learning and deep learning methods
  - **Analyze** the data, **extract information**, and **understand cancer** better

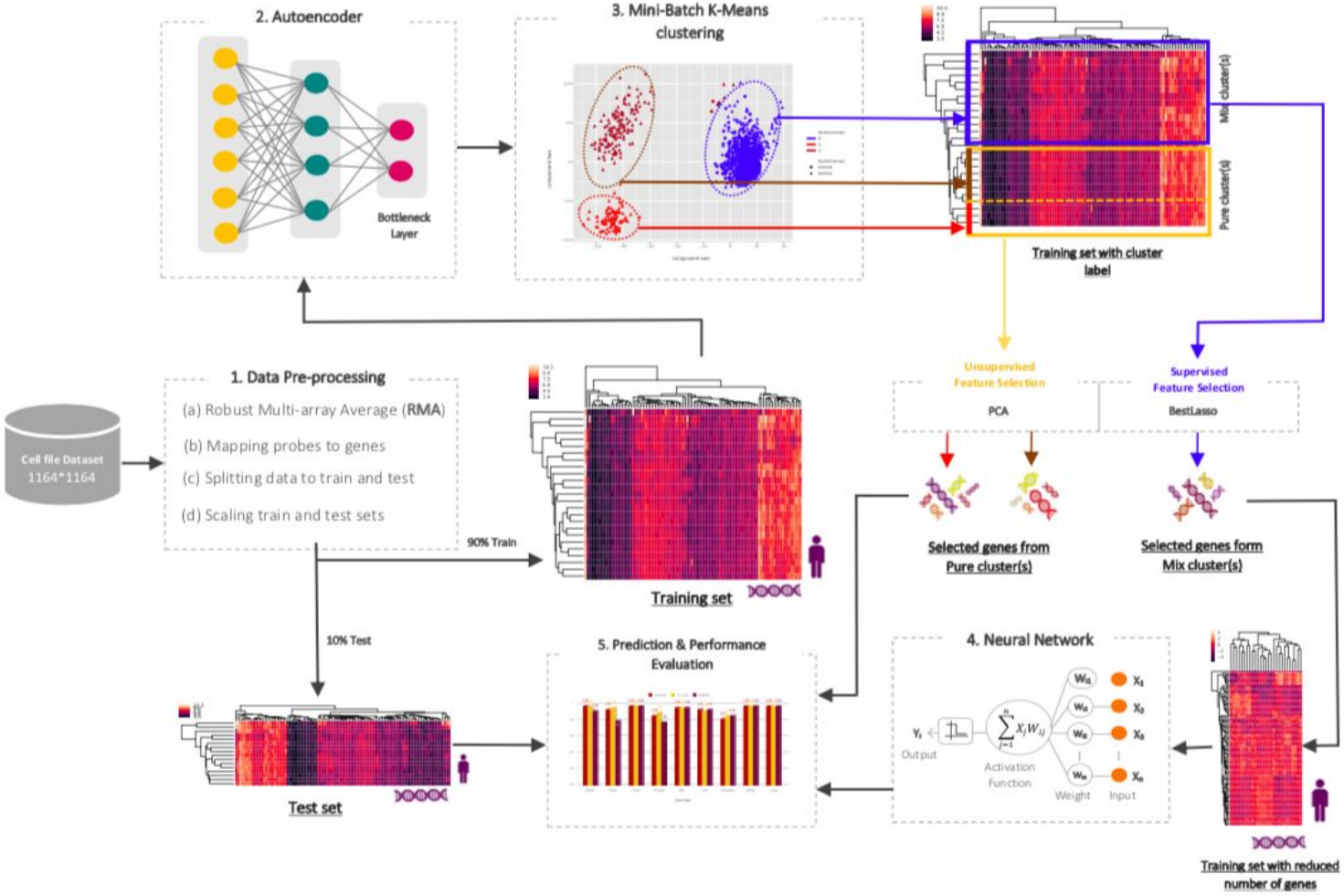
Contribution

- **Cancer subgrouping based on compressed set of genes :**
  - Divide samples into two main group of cancer
- **Gene biomarkers selection for each subgroups**
- **Extract a new group of genes associated with cancer**
  - Mediator genes not differentially expressed
- **Pathway enrichment analysis and creating network of cancers**
  - Important pathways in cancer
  - Hub gene(s) and connection of cancers

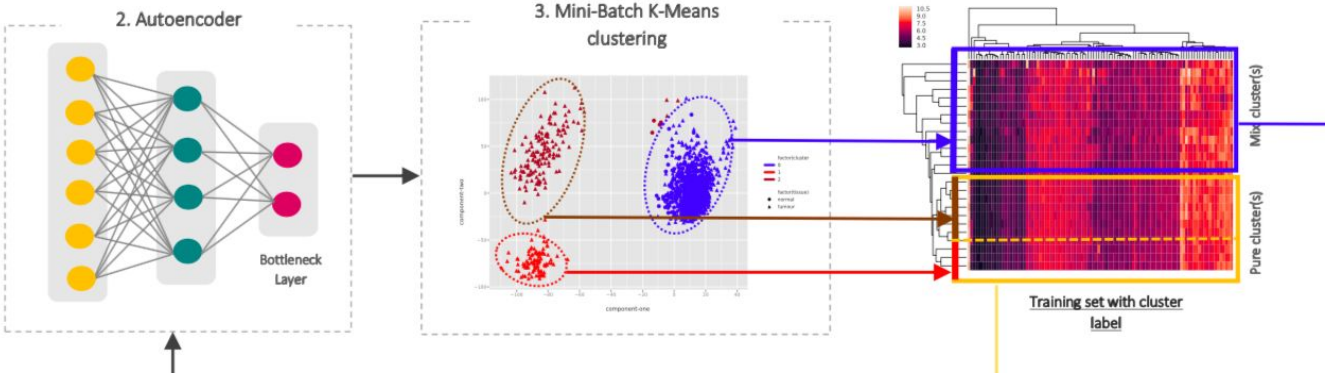
Dataset

- Gene expression data from Gene Expression Omnibus (GEO)
- Number of genes: 22485

Cancer	# of samples	# of tumor samples	# of normal samples
breast	2113	1984	129
Ovary	954	839	115
Colon	1765	1557	208
Prostate	389	299	90
Skin	621	357	264
Liver	588	279	309
Pancreatic	259	178	81
Kidney	1031	589	442
Lung	1087	875	212
Total	**8807**	**6957**	**1850**


2. Autoencoder
Bottleneck Layer
3. Mini-Batch K-Means clustering
Mix cluster(s)
Pure cluster(s)
Training set with cluster label
Unsupervised Feature Selection
PCA
Supervised Feature Selection
BestLasso
Selected genes from Pure cluster(s)
Selected genes form Mix cluster(s)
Cell file Dataset 1164*1164
1. Data Pre-processing
(a) Robust Multi-array Average (RMA)
(b) Mapping probes to genes
(c) Splitting data to train and test
(d) Scaling train and test sets
90% Train
Training set
10% Test
Test set
5. Prediction & Performance Evaluation
4. Neural Network
Output
Activation Function
Weight
Input
Training set with reduced number of genes

Feature Compression and Clustering

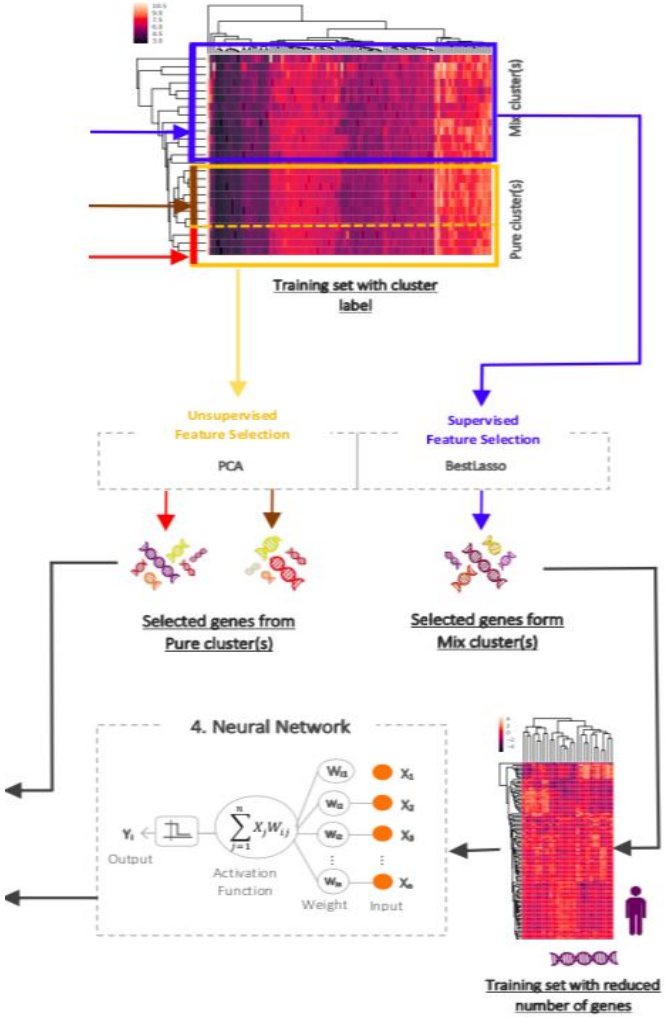


- To find the best subgroups → decrease the dimensionality
- Propose the use of autoencoder to create new representation of data
- Finding subgroups of cancers by Mini-batch K-means
- Determining the optimal number of clusters → Silhouette index
- Divide samples into two main groups of cancer: pure and mixed

$$C_i = \begin{cases} pure & \text{if } n_i^t/n_i^n < \alpha \\ & \text{or } n_i^n/n_i^t < \alpha, \\ mixed & \text{otherwise.} \end{cases}$$

Genes Selection and Classification

- Selecting only differentially expressed genes results:
  - Focus on **single genes** separately→ interaction of genes
  - **Ignore mediator** genes
- Machine learning feature selection → Biomarkers
  - **Pure cluster**: unsupervised (PCA)
  - **Mix cluster**: supervised (BestLasso)
- To evaluate importance of selected genes
  - The classification → only on **mixed clusters**
  - For samples in **pure clusters** → label is **set by the cluster**
  - All the results **are summed** to calculate evaluation metrics



Results

- **Cancer subgroup identification**

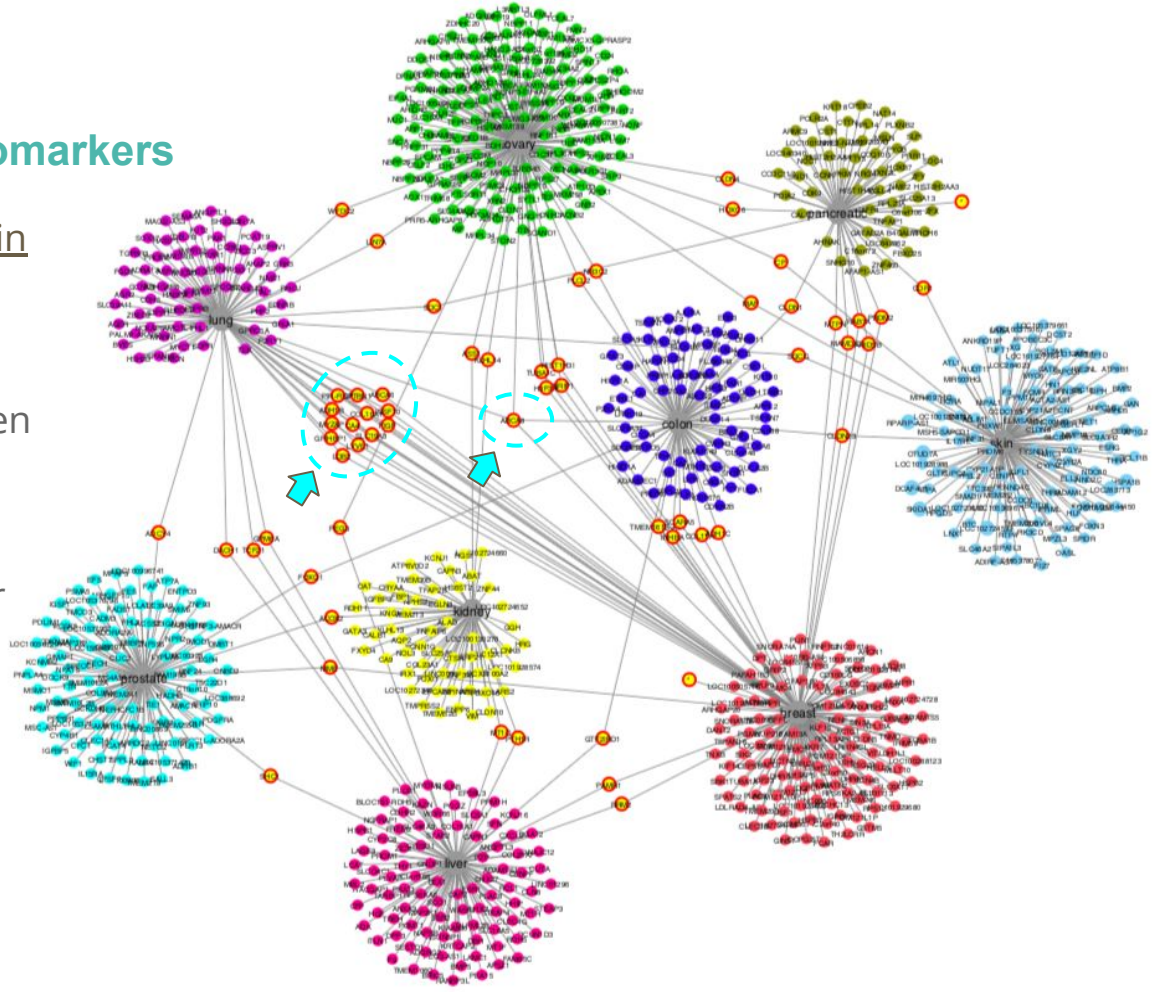
Cancer \ Number of Clusters	2	3	4	5	6
Breast	0.80	**0.82**	0.60	0.55	0.49
Ovary	0.68	**0.73**	0.46	0.47	0.36
Colon	**0.73**	0.55	0.42	0.41	0.38
Prostate	0.59	**0.70**	0.55	0.60	0.54
Skin	0.53	0.56	**0.61**	0.58	0.51
Liver	0.41	**0.52**	0.50	0.39	0.39
Pancreatic	0.55	**0.62**	0.58	0.48	0.42
Kidney	**0.69**	0.50	0.48	0.44	0.41
Lung	**0.51**	0.44	0.34	0.33	0.35

- **Cancer subgroups identification**

Cancer	**Cluster Types (# of Gene Signature)**			
Breast	Pure (28)	Pure (36)	Mixed(70)	
Ovary	Pure (41)	Mixed (90)	Mixed(19)	
Colon	Pure (normal)	Mixed (82)		
Prostate	Mixed (29)	Mixed(53)	Mixed(25)	
Skin	Pure (normal)	Pure (49)	Pure(38)	Mixed(28)
Liver	Mixed (34)	Mixed (27)	Mixed (45)	
Pancreatic	Pure (27)	Mixed (26)	Mixed (9)	
Kidney	Pure (normal)	Mixed (59)		
Lung	Pure (30)	Mixed (56)		

Results

- **Network of cancer based on biomarkers**
  - ABCA8 as a hub gene and it cited in the literature as related to cancer several times
  - Common gene signatures between breast and lung cancers
  - Lung is proven to be the most probable organ to metastasize for breast



Results

- **Prediction metrics**

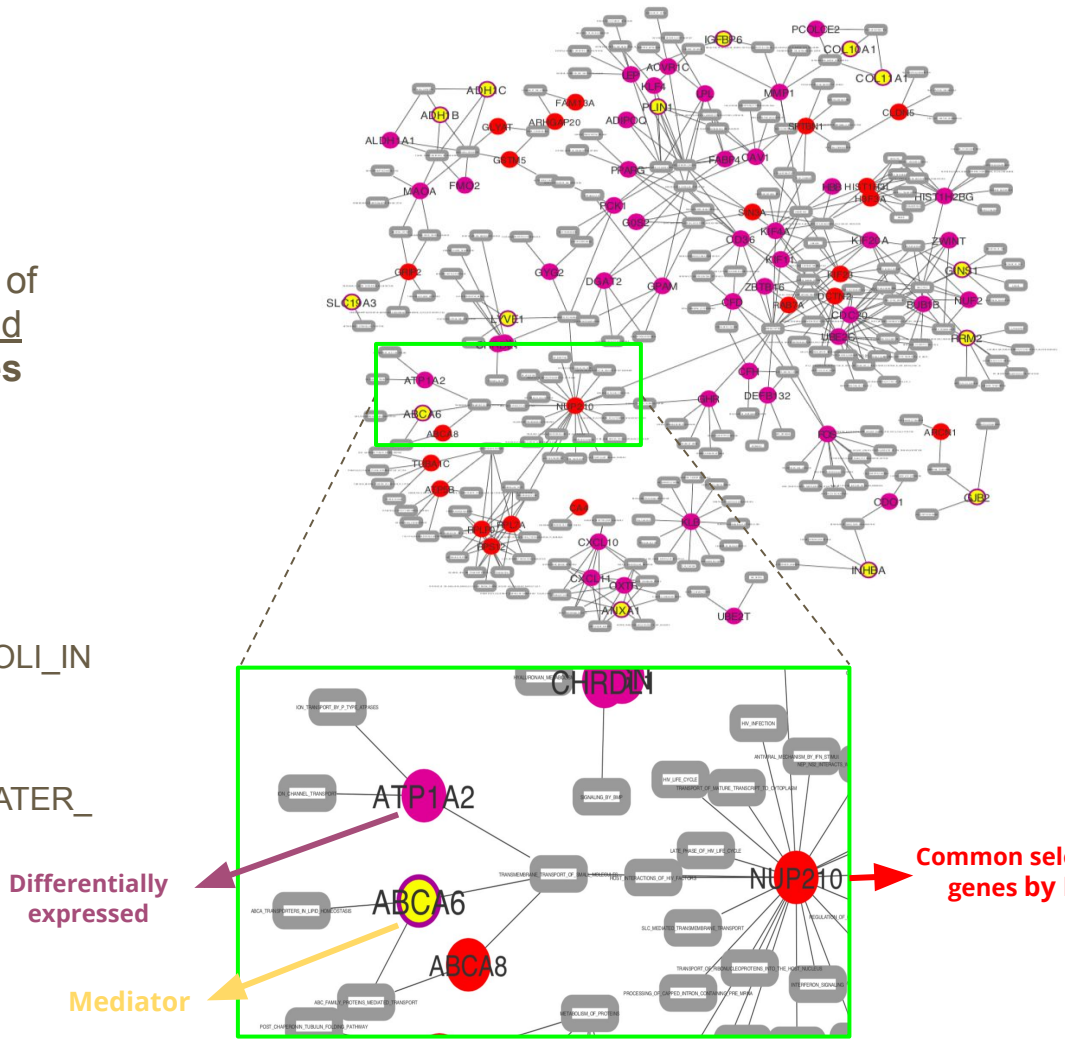
Cancer	Acurracy	$f1_score$	AUROC
Breast	0.99	0.99	0.96
Ovary	0.97	0.98	0.90
Colon	0.99	0.99	0.99
Prostate	0.93	0.98	0.89
Skin	0.98	0.98	0.98
Liver	0.97	0.96	0.97
Pancreatic	0.91	0.93	0.93
Kidney	0.99	0.99	0.99
Lung	0.99	0.99	0.99

- **Pathway Enrichment Analysis (KEGG pathway)**
  - PURINE_METABOLISM
  - PATHWAYS_IN_CANCER
  - LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION
  - PYRIMIDINE_METABOLISM
  - MAPK_SIGNALING_PATHWAY
  - FOCAL_ADHESION
  - NEUROACTIVE_LIGAND_RECEPTOR _INTERACTION

NOTE: The number of selected gene signatures in each pathway is determined and **normalised by the total number of genes** in the pathway

Results

- **Mediators genes (breast)**
  - From all the gene signatures, some of them are differentially expressed and some are not called **mediator genes**
  - Important pathways**:**
    - breast (mixed): RENIN_ANGIOTENSIN_SYSTEM
    - breast (pure): PATHOGENIC_ESCHERICHIA_COLI_INFECTION
    - breast (pure): VASOPRESSIN_REGULATED_WATER_RREABSORPTION



Conclusion

- Hybrid deep **learning-based model** able to
  - Find cancer subgroups → autoencoders and a clustering technique
  - Select a subset of biomarkers that predict cancer with high accuracy (more than 20.000 to less than 100 genes)
  - Find mediator genes

- The cancer tissues can be separated into **two main subgroups**
  - Pure: being very different than normal samples
  - Mixed: group more similar to normal tissues

- Created a **network of cancers** from the selected genes
  - To discover the relationships of the cancers
  - Found out that lung and breast cancer are more related
  - ABCA8 as hub gene

- Finding **important pathways** in KEGG
  - showed that they are known to have a role in cancer formation and progression

- **Publication**
  - Collaboration with Cancer Antibody Discovery and Immunotherapy group:
    Immune mediator expression signatures are associated with improved outcome in ovarian carcinoma, Journal Paper (Accepted in OncoImmunology)

Future Work

- Extend this work by
  - Predict the grades and stage of cancer and analyze them in each subgroups
  - Comparing cancer subgroups with known cancer subtypes
  - Adding relevant clinical data and target therapy for each subtype
  - Predicting survival time