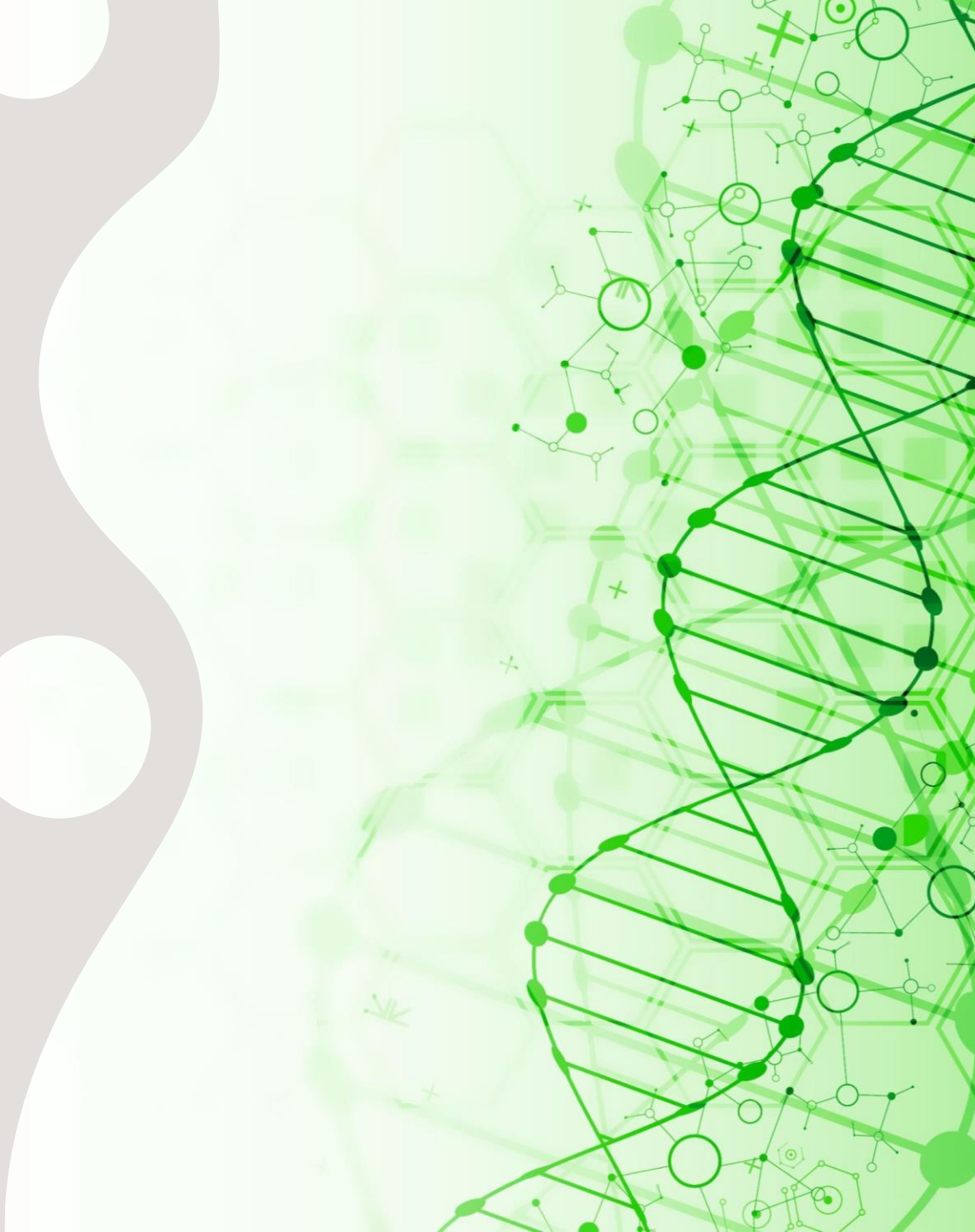
Survival Analysis: Harnessing Genomic, Imaging, Drug, and Clinical Data

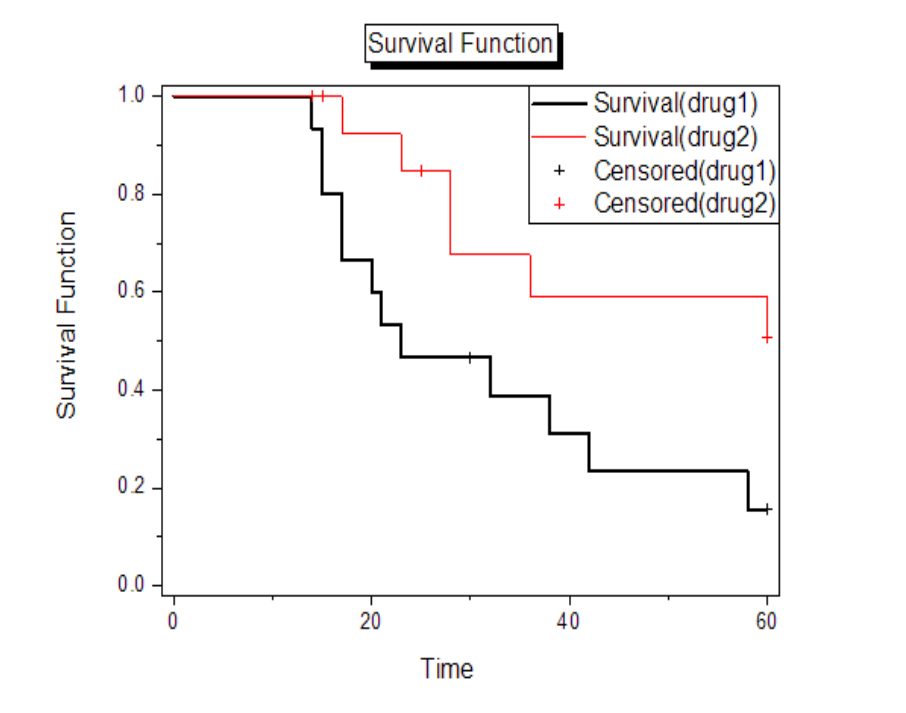
Amruta Mahendra Parulekar

20d070009

EE 492- BTP II



What is Survival Analysis?



Kaplan-Meier curve

Time-to-event data:

- Survival analysis deals with time-to-event data, where the "event" could be death, disease recurrence, or other endpoints.

Censored data:

- Not all patients experience the event of interest within the study period. Some may be lost to follow-up or still alive at the end. Their data is censored, meaning the exact time of the event isn't observed, but it's known the event hasn't occurred by the study's end.

Survival function:

- The survival function, denoted as $S(t)$, gives the probability of surviving beyond time t. It's calculated as the proportion of patients surviving beyond t.

Hazard function:

- The hazard function, denoted as $h(t)$, describes the instantaneous event occurrence rate at time t, given the patient has survived up to that time.

Motivation

Genetic testing is expensive: which genes should be tested for?

What treatments and drugs should be given for a better chance of survival?

What is the chance of survival of **a** patient given their genetic **makeup**, drugs given and other clinical data.

What are the clinical features of the at-risk population, so that these cases can be regularly screened?

What insights does imaging data add to survival prediction?

The TCGA Dataset

3314 patients each given multiple drugs. 8909 instances.

Genomic data	Values of 748 different genes
Drug data	Multiple drugs with different spellings, converted to 12-bit encoding vectors corresponding to 12 chemotherapy categories. If multiple drugs were given to a patient, the encodings were added up.
Image data	Whole slide images (Histopathology) available for all cancers. CT scans and MRI data was available for some cancers.
Clinical data	Clinical data, such as age, gender, histological type, chemotherapy type, cancer type and organ was obtained and relevant features were chosen by a clinician. There were 25 cancer types across 20 organs.
PFI and OS	PFI or progression free interval is the time for which the drug curbs the progression of tumor. OS represents death of patient. These are censored variables.

Baseline

Cox regression (Coz proportional hazards model) : CI without clinical data 0.65

- The core assumption of Cox regression is that hazards between individuals are proportional over time.

covariates $z \in \mathbb{R}^p$:

$$h(t; z) = h_0(t) \exp(z^\mathsf{T} \beta), \quad \beta \in \mathbb{R}^p.$$

- Cox regression enables inclusion of covariates, such as demographic or clinical factors, affecting the hazard rate.
- Cox regression maximizes a partial likelihood, comparing hazards for censored and uncensored data, adjusting for covariates.
- The Cox regression output provides hazard ratios (HRs) where HR > 1 signifies increased risk, HR < 1 decreased risk; e.g., an HR of 1.5 implies a 50% increase in risk per unit change in the variable

$$\frac{h(t; z)}{h(t; z')} = \exp((z - z')^\mathsf{T} \beta)$$

Loss function and Metric

We aim to create a neural network that simulates cox regression. We do this using Partial likelihood maximisation instead of loss minimization

Partial Likelihood

Partial likelihood penalizes when the order of dead patients is incorrect, or in case of censoring, it penalizes wrong order if the live patient is known to live more than dead patient

D=Dead set, R=Risk set (Tl>t)

The partial likelihood is

$$L(\beta) = \prod_{l \in D} \frac{\exp(z_l^T \beta)}{\sum_{k \in R(t_l)} \exp(z_k^T \beta)}$$

Cox suggested to estimate β by maximizing $L(\beta)$.

- does not depend on h_0
- does not depend on actual death times - only their order
- censored observations only appear in risk set (as for Kaplan-Meier)

Concordance index

The C-index measures the model's ability to rank survival times, ranging from 0.5 for random chance to 1 for perfect prediction, useful for assessing discriminatory power but not calibration.

Method 1: Graph neural networks

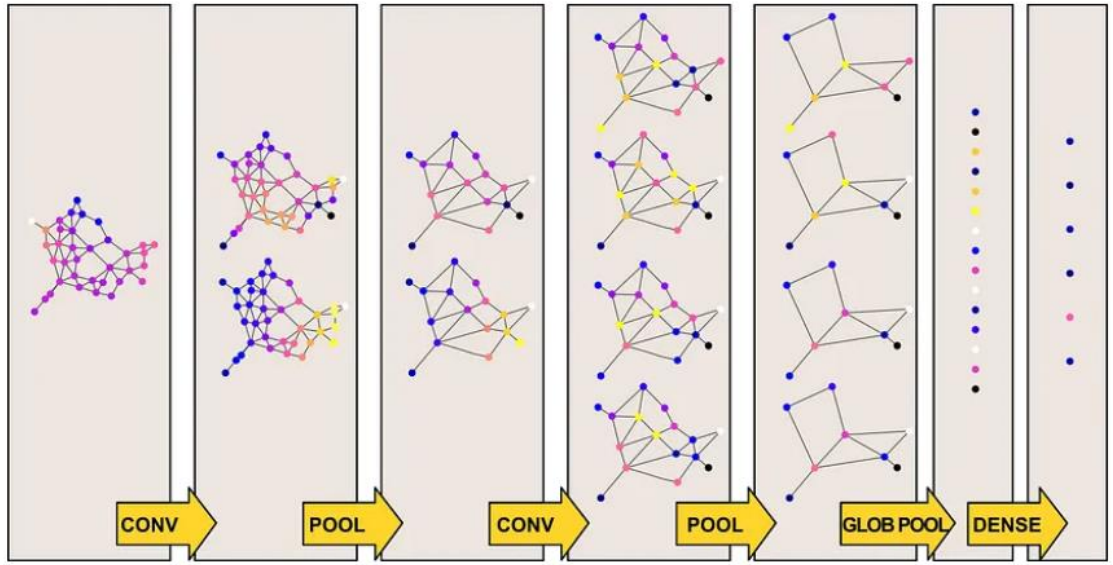
Protein- protein interaction data was used to create graph edges between the 748 genes.

Graph convolutions were used to train the model.**(weights shared)**

In each iteration, each **node is an aggregation of the embeddings of neighboring nodes**

Eventually, each node has information of distant nodes.

Model was discarded because overfitting was occurring in one epoch due to high model complexity and less amount of data



Method 2: Neural Ranking model

Data

- Train, Val, Test data size : 2563, 285, 712 patients
- Taken 5000 pairs of patients from the sets for training

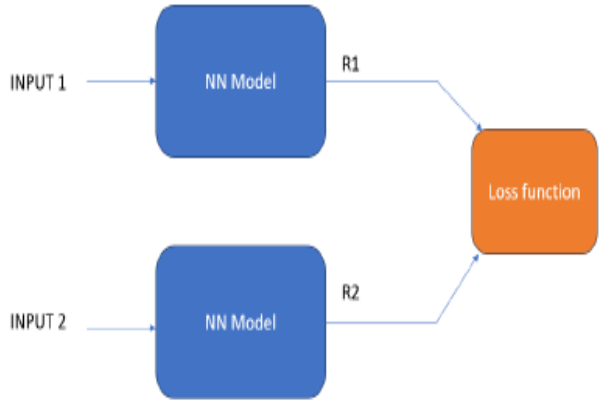
Model architecture

- Simple feed forward neural network.
- Two copies with shared weights

Loss function

- A pair of patients taken and survival is predicted by each copy
- If ordering wrong, **hinge** loss function penalizes

C-index 0.69 on genomic and drug data



Loss Function : $-log(\frac{R_1}{R_1+R_2})\delta(t_2 > t_1) - log(\frac{R_2}{R_1+R_2})\delta(t_1 > t_2)$

Method 3: Utilizing Mouse Data

Cell line data

- Mouse data, same cancer types, classified according to cell lines which are basically cancer cell cultures.
- 723 overlapping genes with human data
- 31769 instances
- Output – drug response (IC50 – drug quantity needed to inhibit a biological process by half)

Data preprocessing

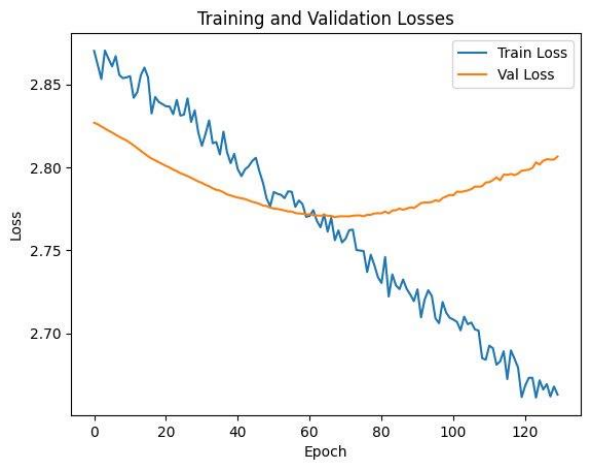
- Test train split on cell line data was performed cell line wise to prevent data leakage.
- Model performance was bad for survival prediction (CI =0.49), so instead of death, progression free interval was predicted for humans, which had a higher correlation with drug response

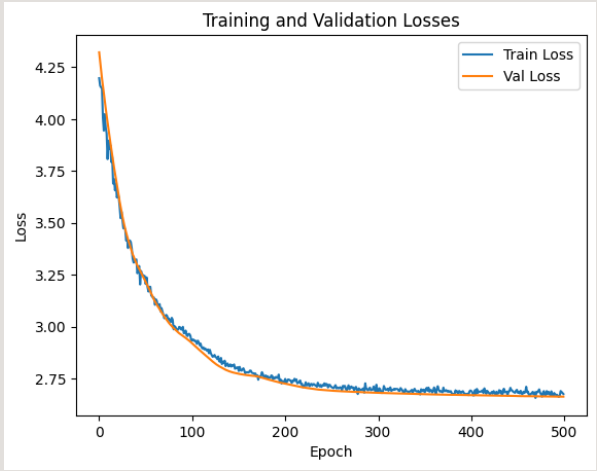
Model architecture

- 723 genes were passed through feed forward layers with ReLU to reduce dimensions to 12
- 12 bit drug encoding was concatenated with 12 length genomic vector and passed further through a feed forward network to a single output neuron

Method 3: Utilizing Mouse Data

Problem: Overfitting if model trained on only TCGA





Solution	Description	Concordance Index for PFI
Zero shot testing	Train on mouse data, test on human data	0.5346
Transfer learning	Train on mouse data, freeze all layers except last two, fine-tune on human data, test on human data	0.6412 (0.7225 CI on mouse data for drug response prior to finetuning)
Domain adaptation	Discriminator added which differentiates between mouse data and human data to try to align their distributions	0.5745

Method 4: Incorporating Clinical Data

Relevance of clinical parameters for survival was tested by training with one clinical parameter at a time along with genomic data

Clinical feature + Genomic data	Concordance Index for survival
Organ	0.6112
Race	0.4863
Stage	0.5577

After discussion with clinician , organ, stage and type were the most clinically relevant. Organ is highly correlated with type.
It was found that some clinical parameters benefitted even more than genomic and drug data for the simple neural cox model.

Type of data	Concordance Index for survival
Genomic+Drug	0.63
Clinical+Genomic	0.69
Clinical+Genomic+Drug	0.70

Method 5: Gene subset selection

Not based on cancer type

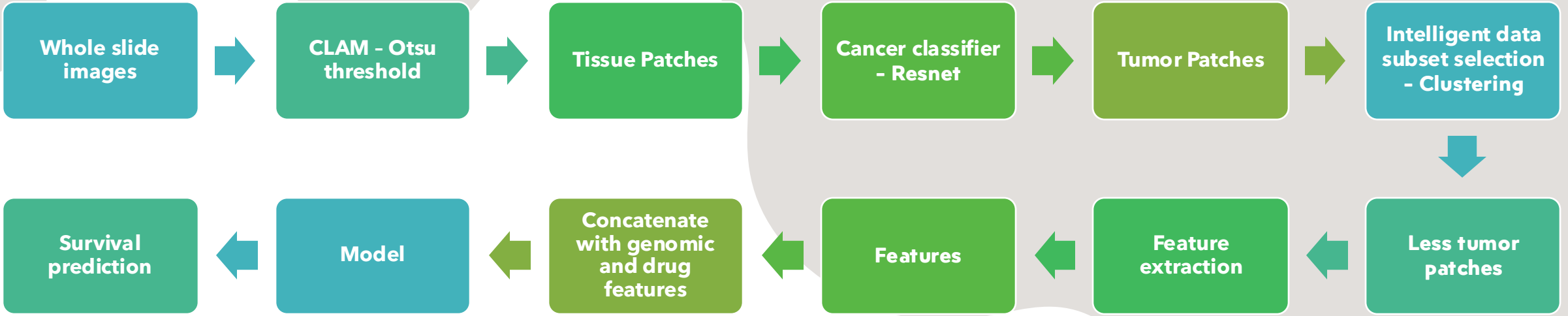
- Initially, 50 genes were chosen, based on cell-line drug response, without considering cancer type
- Several techniques like L1 regularization, PCA, etc were tried.
- Since only 50 gene columns, model architecture was simplified to a normal FFNN
- Best results (CI on PFI):
  - Zero-shot testing – 0.4998, Transfer learning – 0.6638

Based on cancer type

- The first layer was 1 to 1 connections of 723 genes followed by a FFNN.
- L1 regularization was used to zero out weights for the first layer of all but 20 genes.
- This was done separately for every cancer type, to get a gene list corresponding to cancer type. The list was given to a clinician for verification.
- While testing, only the type corresponding genes were passed.

	On all genes	20 selected genes	Instances
Adrenal	0.611111	0.659722	223
Bladder	0.518336	0.553366	401
Brain	0.723641	0.770465	665
Breast	0.546534	0.634615	986
Cervix	0.486667	0.5	251
Colon	0.390402	0.553826	447
Esophagus	0.57971	0.57971	184
Gall Bladder	0.666667	0.777778	30
Head & Neck	0.49465	0.546541	508
Kidney	0.683008	0.690293	878
Liver	0.450428	0.426561	357
Lung	0.574332	0.511932	999
Ovary	0.579964	0.5	420
Pancreas	0.631148	0.545082	146
Prostate	0.285714	0.5	476
Rectum	0.283333	0.791667	147
Skin	0.506511	0.530983	443
Uterus	0.375752	0.382766	367
stomach	0.545151	0.621237	459
thyroid	0.672566	0.884956	522

Method 6: Incorporating Medical Images



- The parts of the pipeline were created and functioning but required a lot of computation - discarded
- To avoid classification and subset selection, which requires a lot of computation, tissue patches were randomly sampled and sent to a TransMIL model.
- TransMIL does multi-instance learning assuming that atleast one positive patch present
- With 546 WSI of LUAD dataset, each having 1000 patches, Concordance Index for only image data is 0.5984

Challenges faced

Less instances of genomic data

- Overfitting on larger and complex models, especially due to more features, model complexity requirements increase and these complex models cant be completely trained by the lesser instances.

Large size of image data

- WSIs are large, giving millions of patches per patient of which only a few hold relevant information. It is hard to narrow this down accurately. They also need a lot of computational power.

More alive patients than dead

- Patients who don't die get censored and this is similar to incomplete information as it cannot be compared to patients who die after this timestamp.

Lack of annotated image data

- The WSIs are not annotated, leading to creation of a large number of patches of which only cancerous ones hold important information. A classifier was trained on open-source data but cannot comment on its efficiency.

Conclusions

- Graph convolutional networks did not work for this problem due to overfitting
- Linear cox performed better than neural cox base by CI 0.02
- Neural ranking model gave an improvement of CI 0.06
- Transfer learning with mouse data (PFI not OS) gave an improvement over direct training and prevented overfitting.
- Incorporation of clinical data gave an improvement of 0.07
- Gene subset selection helped a lot in preventing overfitting.
- Cancer specific genes gave the highest boost to CI
- Using WSIs alone had benefits, but not better than genomic+drug+clinical

Future work

- We hypothesize that for survival prediction, combining cancer specific gene selection, clinical data addition, drug encodings and image data will give optimal benefits and will test this out.
- Combining neural ranking with clinical data can also potentially give good results
- For PFI prediction, transfer learning with mouse data is the best option.

References

[1] Shao, Zhuchen & Bian, Hao & Chen, Yang & Wang, Yifeng & Zhang, Jian & Ji, Xiangyang & Zhang, Yongbing. (2021). TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classication.

[2] Fotso, Stephane. (2018). Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework.

[3] He, D., Liu, Q., Wu, Y. et al. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. Nat Mach Intell 4, 879–892 (2022). https://doi.org/10.1038/s42256-022-00541-0

[4] Kipf, Thomas & Welling, Max. (2016). Semi-Supervised Classification with Graph Convolutional Networks.

[5] Ilse, Maximilian & Tomczak, Jakub & Welling, Max. (2018). Attention-based Deep Multiple Instance Learning.

Thank you!