COMP3212: Understanding Cancers

Luke McClure - 29573904
May 28, 2020

1 Introduction

For this investigation I mainly focused on using the GDC hub data for various cancers located at https://xenabrowser.net/datapages/. Datasets obtained from only one hub was important to ensure the same attributes were used throughout each dataset.

Four cancers were chosen, Esophageal Cancer (ESGA), Glioblastoma (GBM), Pancreatic Cancer (PAAD), and Sarcoma (SARC). These were chosen due to the diverse nature of where each affects. This range hopefully will result in a clear distinction between each throughout investigation.

2 Phenotypes

The phenotype data in these datasets have several challenges when attempting to analyse and learn on them. There are some fields that are present in some cancers which are not in others, within analysis for this project the largest subset of fields that is present within every cancer is used.

These datasets are also riddled with NaN inputs. To translate these missing inputs for learning any NaN value was replaced with the mean of that column, doing so means that the vector associated with that column will not have any weight as would be the case if any NaN value was replaced with a constant like 0.

A decision tree was constructed of the remaining fields from the phenotype datasets from each cancer, shown as Figure 1, with a trained accuracy of 99%. This tree reveals an interesting number of characteristics between the different cancers investigated. While the second level of this tree at day_of_dcc_upload seems to be a fluke in being able to predict Sarcoma, other features of this tree can give hints of the behaviors of these cancers.

Level 1 separates a majority of Glioblastoma cases under tissue_prospective_collection_indicator, this may be due to the location of this disease in the brain meaning that prospective collection is not normally done. This is reflected in the high number of NaN values, resulting in a higher mean that will be put on these values.
Level 3 indicates the feature lost_follow_up as a good indicator between Pancreatic cancer and Esophageal cancer alongside outliers. As losing follow up can be a result of withdrawal or moving away, this may be a result of the unfortunately higher mortality rates of pancreatic cancer (5yr: 91.8%, 10yr: 97.8%) compared to esophageal cancer (5yr: 81%, 10yr: 85%) not allowing an individual to potentially move away and therefore become lost to a study.
Level 4 only separates from Esophageal cancer and outliers from the previous classes and therefore is not as revealing a feature.

Due to the difficulties faced in feature extraction and uniform data phenotype analysis is not going to be pursued further.

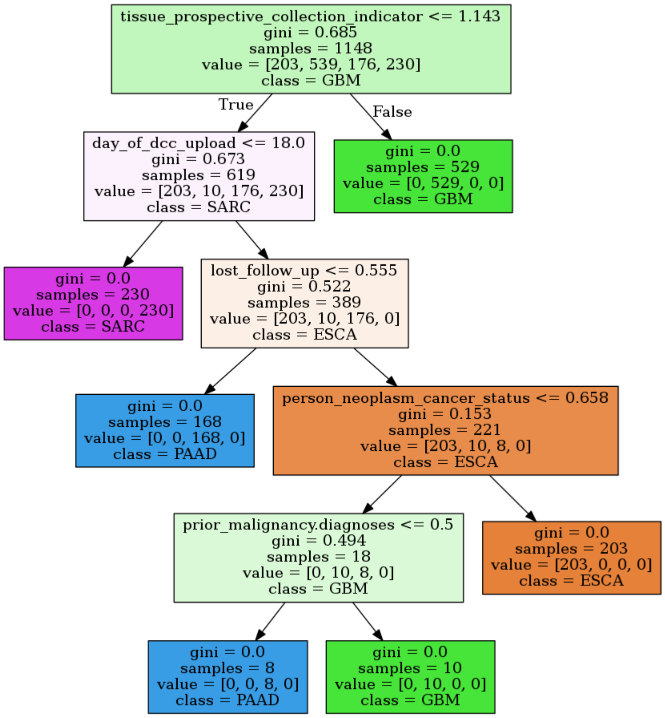

Figure 1: Phenotype Decision Tree

3 Genotypes

The numeric nature of genotypic data lends itself more suitably to many methods of machine learning. RNA sequencing gene expression count datasets were used for this, allowing different expression of proteins within the various types of cancers to be compared. This also results in the same domain for each dataset based off the genes found within the human genome.

Initially the same exploratory decision tree was generated, seen as Figure 2, was tested to 98% accuracy and shows how the different expression of proteins can give clues as to which cancer is being studied.

This tree largely sorts each class by the second level, with outlying points being sorted deeper, therefore only the first two levels will be focused on.
Level 1 of the decision tree is based around TTC22, with low counts leading to Sarcoma and Glioblastoma and high counts leading to Pancreatic and Esophageal cancer.

Within level 2 there are two branches, one to decide between Sarcoma and Glioblastoma, and one to decide between Pancreatic and Esophageal cancer. GFAP is used for the first branch, where low counts indicate Sarcoma and high counts indicated Glioblastoma. It is unsurprising that GFAP is used as this protein is expressed extensively in cells of the central nervous system, where Glioblastoma affects, so high counts of it indicate CNS cells. Meanwhile, ALO22718.1 is used to indicate between Pancreatic cancer with low counts, and Esophageal cancer with high counts. A study found high expression of this protein in the esophagus and comparatively none in the pancreas.

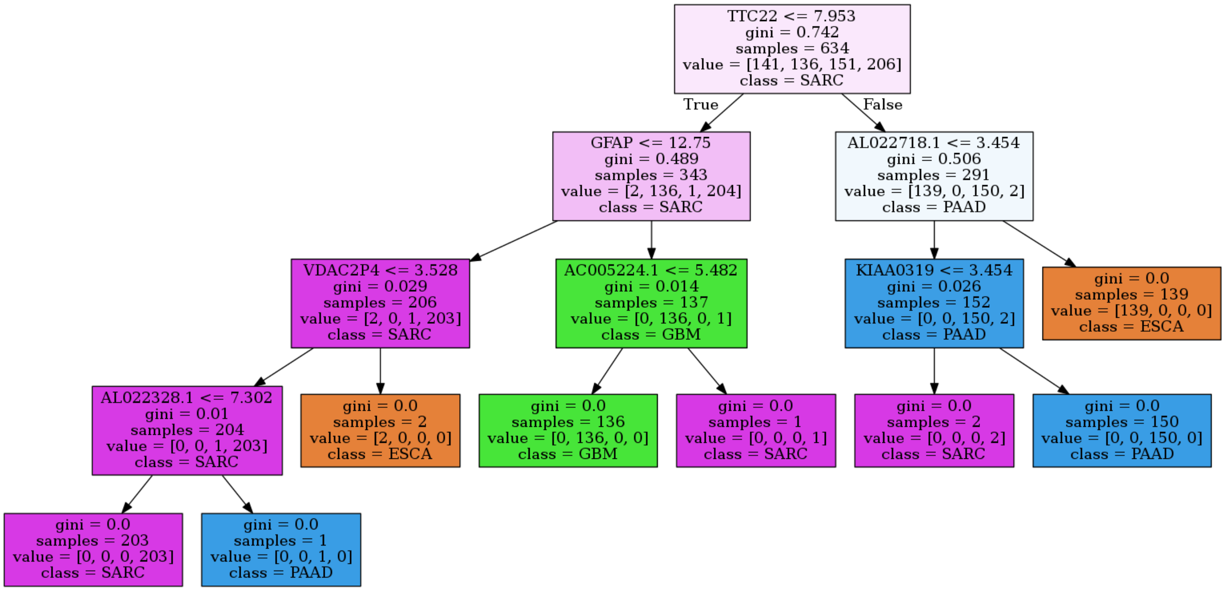

Figure 2: Genotype Decision Tree

Linear Discriminant Analysis was used to condense the 60,488 dimensional data features in the dataset as much as possible. This will both allow for visualisation of these datasets and grouping of these classes together.

The LDA performed in Figure 3 shows a 2D and 3D representation of this data, where between both it is clear how separated the classes are. While this is helpful for visualisation, the grouping does not provide a further scope for analysis unlike other methods.

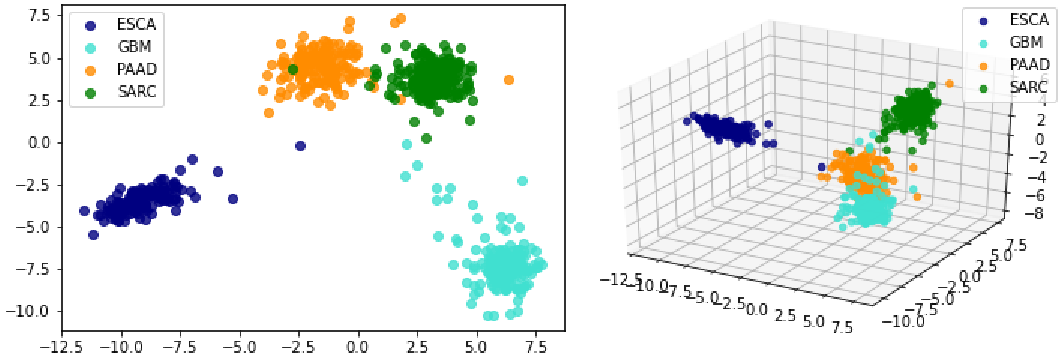

Figure 3: LDA Results

Principal Component Analysis is able to project high dimensional data into lower dimensions, while preserving a large portion of the variance within a dataset. PCA from 60,488 dimensions into 2 was able to capture 22% of the variance in the dataset, and with 3 dimensions this increases to 33%. Whilst low, to get 80% of the total variance in the dataset captured needs only 400 dimensions compared to the original 60,488.

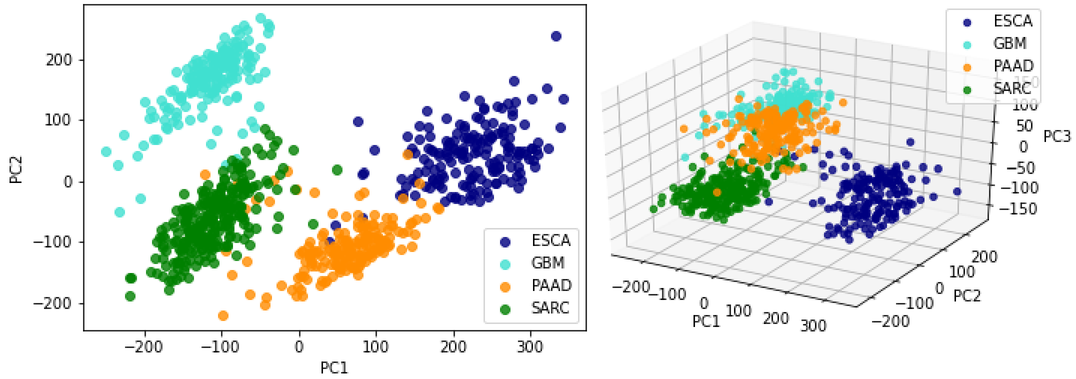

Figure 4: PCA Results

While there is significant overlap between some classes in Figure 4, the classes do form succinct groupings on their own. It is important to note that the axes in these graphs are along the principal components. Any analysis of each principal component sheds light on the relationships within that axis, but these can reveal a lot due to how intrinsic the variance between classes is.

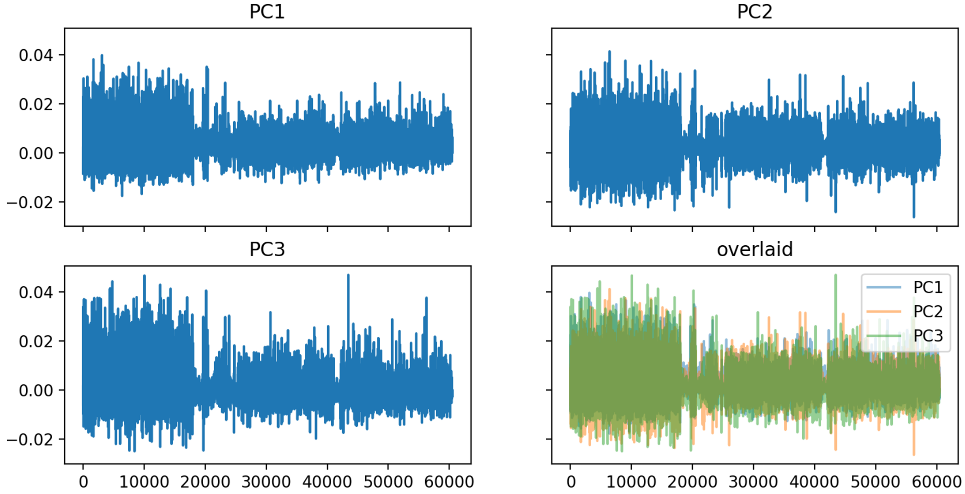

Figure 5: Effect on Principal Component of each protein

By finding the values that contribute greatest to a principal component (Figure 5), a measure can be found of the greatest influences to variation along that axis, and therefore the proteins that greatest affect the variation in that axis.

	1	2	3	4	5
PC1	CEACAM5	CEACAM6	LCN2	AGR2	TMPRSS4
PC2	GFAP	PMP2	AQP4	MIR9-1HG	NCAN
PC3	AC087379.1	SST	TTR	GJB1	REG1A

Table 1: Top 5 scaling proteins in each dimension

Principal Component 1:

1. CEACAM5 - Part of the carcinoembryonic antigen (CEA) gene family, a gene family used to diagnostically test for pancreatic cancer as well as showing expression in esophageal cancer.
2. CEACAM6 - Another member of the CEA gene family.
3. LCN2 - An iron trafficking protein, has been shown to be expressed in high amounts in pancreatic cancer.

This principle component seems to focus greatly on pancreatic cancer, with the top 4 proteins directly having ties to this type of cancer.

Principal Component 2:

1. GFAP - Discussed earlier in the geneotype decision tree, GFAP is expressed in high amounts in CNS cells such as in the brain.
2. PMP2 - A protein involved in the myelin part of nervous system cells, present in small amounts in CNS cells but more common among peripheral nervous system cells. Expressed in some gliomas.
3. AQP4 - A water channel protein, it can be particularly over-expressed in glioblastomas and this protein has become the subject of research for treatment.

Principal component 2 particularly focuses on glioblastoma. All of 5 most influential proteins for PC2 have direct influence or link to glioblastoma.

Principal Component 3:

1. AC087379.1 - Not much was found on this gene, except for it being expressed in many places including the brain and pancreas.
2. SST - precursor to a hormone, disorders related to this gene affect both the pancreas and the esophagus.
3. TTR - A protein found in cerebrospinal fluid, interestingly also a marker for lung cancer.

This principal component is not as concise in how it reflects the proteins involved in these cancers, proteins 4 and 5 are heavily involved in cancers and the pancreas respectively.

4 Conclusion

Decision trees worked well for classifying between different cancers, both with phenotypes and genotypes. But to gain a deeper understanding into the difference between these cancers and how they are so different principal component analysis is key to draw out the main protein differences between the classes, with so many proteins relevant to cancers studied found via PCA, it is clear that this method is useful in narrowing a search for the proteins involved in these cancers. Not all the cancers studied in this investigation were prevalent throughout, this may be due to the non-specific nature of Sarcoma. It would be interesting to test further if this is an outlying example or if PCA was only successful for this subset.