Imperial College London

Prediction of blood cancer using blood count phenotypes

Group 5

Abstract

Blood cancer is the fifth most common cancer in the UK, with more than 40,000 people diagnosed annually. Earlier prediction of blood cancer could lead to improved patient outcomes. Previous studies have investigated prediction using gene expression and imaging data. Blood count data is more accessible and could be used for prediction. Therefore, we used machine learning to explore predictors of blood cancer, both in terms of blood characteristics as well as lifestyle factors. We used data from the UK Biobank, one of the largest biomedical databases in the world. We aimed to answer the following: 1) Are there clusters of people with shared blood characteristics? 2) What is the contribution of each group to blood cancer risk? 3) Can we identify lifestyle variables and/or other biomarkers that account for a large proportion of blood phenotypes? We clustered our data using K-means and GMM clustering and used random forest and logistic regression models to predict blood cancer. One cluster was distinct from the others, showing high levels of white blood cell count, basophil count and monocyte count. C-reactive protein was significantly associated with membership in this cluster. However, prediction of blood cancer was poor for raw blood counts using random forests (AUC = 0.67, F1-score = 0.23) and logistic regression (AUC = 0.63, F1-score = 0.23). Prediction using K-means clustering plus risk factors was also poor for random forest (AUC = 0.75, F1-score = 0.16) and logistic regression (AUC = 0.70, F1-score = 0.11) models. We suggest using blood counts at regular time intervals instead of only at baseline in order to improve prediction.

Contents

1 Introduction 2

2 Methods 2
- 2.1 Subjects 2
- 2.2 Pre-processing 2
- 2.3 Statistics 3

3 Results: Descriptive Statistics 4

4 Results: Clustering (R1) 5
- 4.1 K-means clustering 5
- 4.2 GMM clustering 8

5 Results: Predicting risk of blood cancer from clusters (R2) 9
- 5.1 Binary logistic regression 9
- 5.2 Multinomial logistic regression 9
- 5.3 Binary random forest 10
- 5.4 Multi-class random forest 10

6 Results: Predicting cluster membership (R3) 12
- 6.1 Describing our clusters 12
- 6.2 Cluster prediction from lifestyle variables 12

7 Discussion 14

1 Introduction

Blood cancer is the fifth most common cancer in the UK, and the most common type of childhood cancer. It is caused by mutations in the DNA of blood cells. Common types of blood cancer include leukaemia, lymphoma and myeloma. Anually in the UK, more than 40,000 people are diagnosed with some form of blood cancer [1].

A number of studies have been published in recent years aiming to predict blood cancer, however these have mainly focused on using gene expression ([2], [3]) or diagnosis using microscope imaging ([4], [5]) data . This data may, however, be time consuming and costly to obtain on a large scale. Therefore, we would like to investigate whether blood count data, which can be obtained from routine blood samples, could be used as an alternative. This would significantly improve efficiency and cost, leading to much greater access to blood cancer diagnosis. UK Biobank was launched in 2006 in order to study both genetic and environmental factors contributing to disease development. It currently has over half a million participants, making it one of the largest biomedical databases in the world, and a reliable source for blood count and lifestyle data. Therefore, we aimed to use Biobank blood count and risk factor data in the prediction of blood cancer. Additionally, clustering blood count data into representative phenotypes may improve utility of the models by medical professionals when predicting cancer risk in patients, as lower-dimensional phenotypes will be more intepretable.

In this study, data from the UK Biobank was used to research the following questions: 1) Are there clusters of people with shared blood characteristics? 2) What is the contribution of each group to blood cancer risk? 3) Can we identify lifestyle variables and/or other biomarkers that account for a large proportion of blood phenotypes?

2 Methods

2.1 Subjects

Subjects were selected from the UK Biobank database. Predictor data (blood counts, lifestyle factors and additional risk factors) were collected at the time of participant recruitment. Cases were included if they were diagnosed with blood cancer at least 6 months after recruitment, but excluded if they had myeloma or multiple blood cancers. Controls are excluded if they had any ICD-9 or ICD-10 codes in order to remove possible confounding.

2.2 Pre-processing

The data was extracted and recoded from the Biobank. It was then split into a training set (80%) and a test set (20%) using the `caret` package in R. The training set was imputed, and subjects were matched in a ratio of 3:1 for controls vs cases by age and sex using the nearest neighbours matching method. After imputation and matching, there were $3,936$ cases and $11,808$ controls in the training set. Finally, categorical predictors were one-hot encoded and predictors standardised prior to fitting all models.

2.3 Statistics

Research question 1: Clustering

K-means clustering and Gaussian mixture clustering (GMM) were used to identify clusters of people with shared blood characteristics. Consensus clustering and the silhouette score were used to optimise the number of clusters for K-means due to its sensitivity to the initial parameters. Silhouette scores were calculated using euclidean distances. For GMM clustering, we used the Bayesian Information Criterion (BIC) in order to determine the optimal number of clusters. The BIC measures the goodness of fit of a statistical model, and penalises model complexity. It is more reliable to use for GMM due to the fact that it considers the likelihood of the model, and takes into account the fact that the data is assumed to come from Gaussian distributions. The silhouette score does not take this fact into account, and is a more general measure of clustering quality.

Research question 2: Predicting risk of blood cancer from clusters

Logistic regression and **random forest** were used to predict the risk of blood cancer from cluster membership. SMOTE (Synthetic Minority Oversampling Technique) oversampling was used to generate equal sample sizes between classes in the training data for random forest. Hyperparameter tuning was performed for the random forest model using 5-fold cross validation. The tuned hyperparameters were the number of trees/estimators, minimum samples per leaf node, max number of features, criterion, max depth and minimum samples per split. To evaluate the how well cluster membership predicts blood cancer, the results of the random forest and logistic regression were compared using the following feature sets as inputs:

1. Covariates (baseline model)
2. Covariates + cluster K-means
3. Covariates + cluster GMM
4. Covariates + blood counts
5. Covariates + [best cluster] + lifestyle variables

In order to compare the models, we used metrics such as AUC, F1 score (given a 0.5 threshold), precision and recall. Multiclass models were then conducted in a subgroup analysis in order to predict cancer type (leukaemia and lymphoma) using the different feature sets. For logistic regression, multiple binary one vs rest models were fit. For random forest, a multiclass model was fit, with AUCs reported as the mean of one vs rest results.

Research question 3: Predicting cluster groups

The clustering algorithm which performed best in the supervised models was chosen to be investigated further. Clusters were described by lifestyle and biological factors using both descriptive analysis, as well as logistic regression. Separate univariate logistic regression models were run in order to estimate the effect of the lifestyle factors on cluster membership for each cluster. The lifestyle variables used were alcohol status, housing score, traffic intensity on the nearest major road, C-reactive protein and mood swings.

3 Results: Descriptive Statistics

Summary statistics of covariates and important blood count data are shown in table 1. As expected, the mean age of cases compared to controls is very similar, as is the ratio of males to females. Density plots comparing the matched versus the unmatched data are given in figure 1. The mean BMI is also virtually identical between cases and controls. However, current or previous smoking status is 4%-5% higher amongst cases. A list of all variables used is given in table 2.

Variable	Case	Control
	(N=3920)	(N=10657)
Age		
Mean (SD)	60.9 (6.70)	60.1 (6.61)
Sex		
Male	1718 (43.8%)	4837 (45.4%)
Female	2202 (56.2%)	5820 (54.6%)
BMI		
Mean (SD)	27.7 (4.73)	27.5 (4.37)
Smoking status		
Never	1902 (48.5%)	6086 (57.1%)
Previous	1572 (40.1%)	3786 (35.5%)
Current	446 (11.4%)	785 (7.4%)
Mood swings		
No	2271 (57.9%)	7048 (66.1%)
Yes	1649 (42.1%)	3609 (33.9%)
C-reactive protein		
Mean (SD)	3.05 (5.09)	2.33 (3.79)
Traffic int nearest major road		
Mean (SD)	24500 (22200)	23500 (21000)
Housing score		
Mean (SD)	19.4 (11.0)	19.6 (11.2)
Alcohol drinker status		
Never	174 (4.4%)	388 (3.6%)
Previous	154 (3.9%)	233 (2.2%)
Current	3592 (91.6%)	10036 (94.2%)
Cancer		
Healthy	0 (0%)	10657 (100%)
Leukaemia	1699 (43.3%)	0 (0%)
Lymphoma	2221 (56.7%)	0 (0%)

Table 1: Summary statistics of demographic variables of the study population. Cases were defined as having blood cancer at least 6 months after recruitment.

Blood counts	**Covariates**	**Lifestyle risk factors**
Lymphocyte count, Monocyte count, Reticulocyte count, White blood cell count, Red blood cell count, Hemoglobin concentration, High light scatter reticulocyte count, Haematocrit percentage, Platelet count, Basophil count, Eosinophil count, Neutrophil count, Immature reticulocyte fraction	Age, BMI, Sex, Smoking status	Mood swings, C-reactive protein, Alcohol drinker status, Traffic intensity on nearest major road, Housing score

Table 2: Predictors used in blood cancer predictions.

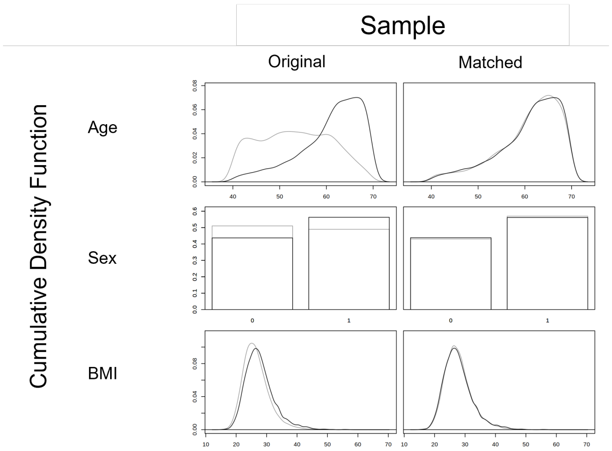

Figure 1: Distribution of age, sex and BMI before and after matching. Matching was performed using nearest neighbours at a 1:3 case control ratio.

4 Results: Clustering (R1)

4.1 K-means clustering

Figure 2 shows the silhouette scores for K-means clustering. K=3 has the maximal silhouette score, indicating that it is the optimal K.

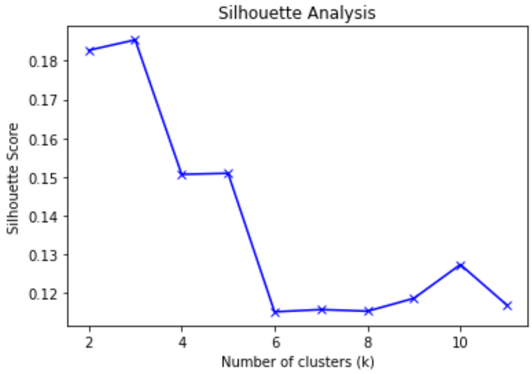

Figure 2: Silhouette scores of K-means clusters over different values of K. Silhouette scores were calculated using euclidean distances.

Clustering method results vary between iterations due to their random initialization. This is especially significant for the K-means algorithm, where initial centroid choice can have a significant impact on the final clusters. To control for this, consensus clustering was used. The consensus CDF in figure 3 shows the cumulative frequency of consensus values for the consensus matrix. The optimal value of K has a minimal increase in the middle section of the graph (i.e where the consensus is not close to 0 or 1). This suggests that higher values of K are more suitable.

Figure 4 shows the consensus matrix heat map, where more colour indicates higher consensus values. These matrices suggest that the best values are K=6 and K=7, since the number of blocks on the diagonals line up with the number of clusters in the algorithm. Due to the analysis of the consensus CDF graph, silhou-

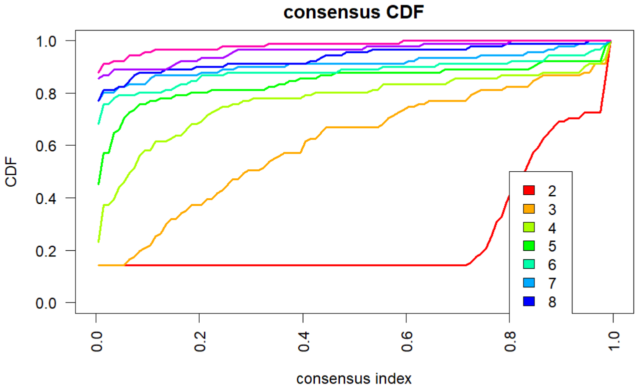

Figure 3: Consensus CDF for K-means clustering.

ette score graph and consensus matrix heat maps, we will use K=3, K=6 and K=7 and compare our results.

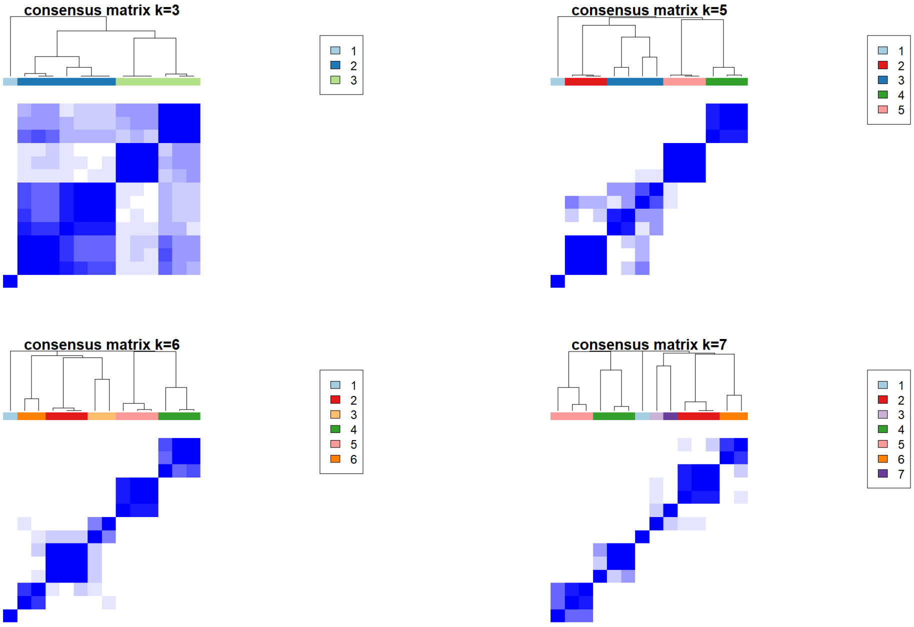

Figure 4: Consensus matrices for K-means clustering for different values of K. Subsampling was performed using 100 resampling iterations.

After clustering our data, we perform dimensionality reduction, as can be seen in the t-SNE (t-Distributed Stochastic Neighbor Embedding) plot in figure 5. This allows us to visualise our high dimensional dataset in two dimensions, whilst preserving both its local and global structure. In these graphs we can see clearly separable clusters. An interesting observation is that for K=3, there is a small green point on the graph corresponding to the third cluster. We see this phenomenon for the other clusters too, whereby there is a cluster which is significantly smaller than the others. We will investigate the composition of the clusters further in section 6.

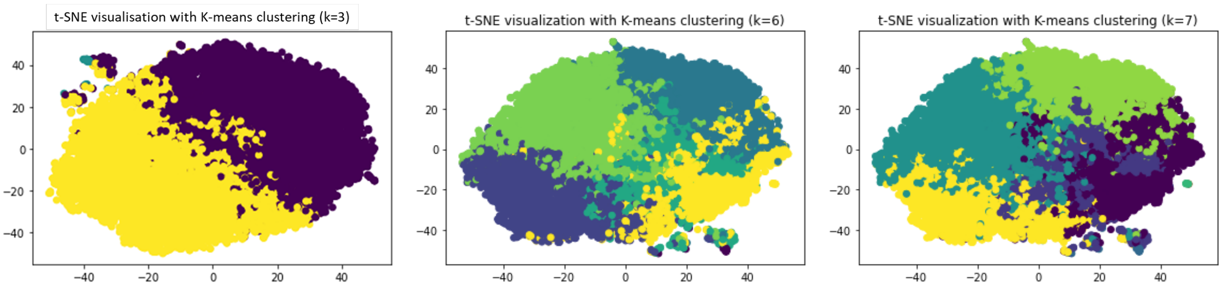

Figure 5: t-SNE visualisation for K-means clustering for different K values. Abbreviations: t-SNE - t-distributed stochastic neighbor embedding

4.2 GMM clustering

Figure 6A shows the BIC scores for GMM clustering, where lower scores indicate better fit. BIC scores decreases significantly up to n=8 clusters, and has diminishing returns thereafter. Therefore, 8 clusters was chosen for GMM clustering. A t-SNE plot of the clustering is also shown in 6B. Unlike K-means clustering, the GMM clusters do not show clear separation, which may indicate a lack of structure to the generated clusters.

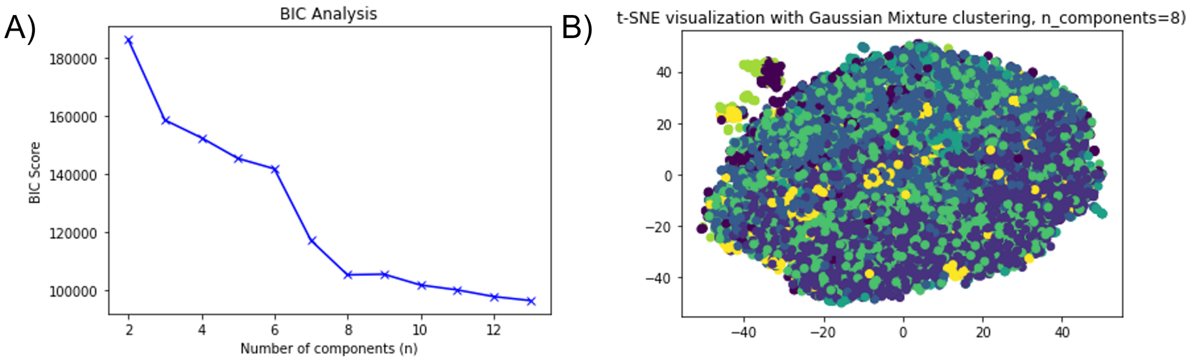

Figure 6: GMM clustering of blood counts. A) BIC scores for GMM clustering over different numbers of components. B) t-SNE of GMM clustering with 8 components. Abbreviations: BIC - Bayesian Information Criterion, GMM - Gaussian Mixture Model.

5 Results: Predicting risk of blood cancer from clusters (R2)

5.1 Binary logistic regression

The results for blood cancer prediction using logistic regression are shown in table 3. AUC, F1 score, precision and recall are used in order to compare and analyse the models.

AUC test results are very similar across all feature groups, approximately 0.60, making it difficult to compare the performance between the models by this metric. It is also possible that these scores have been inflated due to the class imbalance. Hence, it is necessary to analyse the F1, precision and recall scores. The GMM$_2$, GMM$_8$ and blood count models have the highest F1 test results, and they are all very close to one another, around 0.23. However, this is still relatively low. This is due to the fact that all of these models have high recall, but low precision. This means that although the model is able to identify many of the cases, it also has a high false positive rate.

The AUC scores for the training set are similar to those of the test set, and the corresponding ROC curves included in the Appendix for reference in figure 12. The training set also has low F1 scores, which suggests that the poor performance of the models on the test set is likely not due to overfitting. The odds ratios for the logistic regression are also shown in box plots in figure 11 in the Appendix.

Model	Train-Recall	Train-Precision	Train-F1	Train-AUC	Test-Recall	Test-Precision	Test-F1	Test-AUC
Covariate	N/A	N/A	N/A	0.71	N/A	N/A	N/A	0.57
K-means + Covariate (k=3)	0.95	0.01	0.02	0.58	0.95	0.01	0.02	0.58
K-means + Covariate (k=6)	0.79	0.02	0.03	0.61	0.79	0.02	0.03	0.61
K-means + Covariate (k=7)	0.77	0.02	0.04	0.61	0.77	0.02	0.04	0.61
GMM + Covariate (k=8)	0.06	0.53	0.11	0.53	0.69	0.15	0.24	0.62
Blood counts + Covariate	0.32	0.20	0.25	0.66	0.84	0.13	0.23	0.63
K-means (k=3) + Covariates + Risk Factors	0.63	0.03	0.05	0.61	0.53	0.06	0.11	0.70

Table 3: Results of logistic regression for cancer prediction using different feature sets. Covariates are age, sex, BMI and smoking status. Risk factors are mood swings, alcohol drinking status, traffic intensity on the nearest major road and housing score. In each feature set, blood counts are represented either in their raw form, or in a clustered form using K-means or GMM clustering. Abbreviations: F1 - F1-score, AUC - Area under the receiver operator curve, GMM - Gaussian Mixture Model.

5.2 Multinomial logistic regression

Following on from section 5.1, the multi-class predictiveness of the models was compared by splitting the cases group into leukemia and lymphoma. Table 4 shows the result of a multinomial logistic regression. The AUC test scores for the multinomial logistic regression are generally higher than for the binary logistic regression. This may be due to the fact that leukaemia F1 scores are much higher than the lymphoma scores. This suggests that the model is better at predicting leukemia. Although the leukemia F1 test scores are higher than the binary F1 test scores, they are still very low, once again likely due to poor precision on the training set. In conclusion, although taking only leukemia cases into account may improve the model, this is unlikely to improve the prediction significantly enough.

Model	Case	Train-Recall	Train-Precision	Train-F1	Train-AUC	Test-Recall	Test-Precision	Test-F1	Test-AUC
Covariate	Leukemia	N/A	N/A	N/A	0.6	N/A	N/A	N/A	0.77
	Lymphoma	N/A	N/A	N/A	0.55	N/A	N/A	N/A	0.66
K-means + Covariate (k=3)	Leukemia	0.97	0.02	0.03	0.62	0.85	0.07	0.12	0.77
	Lymphoma	0.89	0.00	0.01	0.56	0.38	0.01	0.01	0.65
K-means + Covariate (k=6)	Leukemia	0.96	0.02	0.03	0.65	0.89	0.04	0.07	0.76
	Lymphoma	0.89	0.00	0.01	0.58	0.60	0.01	0.01	0.63
K-means + Covariate (k=7)	Leukemia	0.96	0.02	0.03	0.65	0.89	0.04	0.07	0.76
	Lymphoma	0.89	0.00	0.01	0.59	0.60	0.01	0.01	0.64
GMM + Covariate (k=8)	Leukemia	0.66	0.21	0.31	0.68	0.04	0.40	0.07	0.52
	Lymphoma	0.55	0.01	0.02	0.57	N/A	N/A	N/A	0.53
Blood counts + Covariate	Leukemia	0.91	0.16	0.28	0.71	0.42	0.27	0.33	0.76
	Lymphoma	0.78	0.02	0.03	0.61	0.13	0.01	0.02	0.60
K-means (k=3) + Covariate + Risk Factors	Leukemia	0.86	0.02	0.04	0.64	0.78	0.07	0.12	0.77
	Lymphoma	0.63	0.01	0.02	0.59	0.24	0.01	0.02	0.64

Table 4: Results of multinomial logistic regression for cancer prediction from various features, split by cancer type. Covariates are age, sex, BMI and smoking status. Risk factors are mood swings, alcohol drinking status, traffic intensity on the nearest major road and housing score. In each feature set, blood counts are represented either in their raw form, or in a clustered form using K-means or GMM clustering. Abbreviations: F1 - F1-score, AUC - Area under the receiver operator curve, GMM - Gaussian Mixture Model.

5.3 Binary random forest

Similarly to the previous section, the main results for blood cancer prediction by random forest are shown in table 5 using the AUC, F1 score, precision and recall.

AUC and F1 scores are generally higher than for logistic the regression. However, similarly to logistic regression, the recall is much higher than the precision, indicating a high false positive rate. This is likely due to overfitting since the training F1 scores are much higher than the test F1 scores. Therefore, although the random forest method seems to be superior to logistic regression, the performance of the model is still relatively low with a high false positive rate.

Feature importance for random forest model

Figure 7 shows the directional SHAP plots for the K_3-means random forest models. The K_3-means clustering model is driven mostly by covariates, and the clusters have little effect on the outcome of the model. Covariates such as age, mood swings and current alcohol drinking were highly predictive of case status. Control status was predicted by decreasing housing score. Of the clusters, membership in group 1 was predictive of case status, but the number of samples in this cluster group was small (37 samples), reducing its total SHAP importance.

5.4 Multi-class random forest

Analogously to section 5.2, we repeat the random forest model, but this time for leukemia and lymphoma separately, shown in table 6. The AUC shown is the mean AUC for leukemia and lymphoma. Although splitting cancer type improved the logistic regression model, both the AUC scores and the F1 scores for multi-class random forest are lower than binary random forest.

Model	Train-Precision	Train-Recall	Train-F1	Train-AUC	Test-Precision	Test-Recall	Test-F1	Test-AUC
Covariate	0.59	0.38	0.46	0.58	0.16	0.72	0.26	0.74
K-means + Covariate (k=3)	0.59	0.52	0.55	0.62	0.13	0.8	0.23	0.76
K-means + Covariate (k=6)	0.56	0.66	0.61	0.61	0.08	1	0.15	0.71
K-means + Covariate (k=7)	0.59	0.54	0.56	0.62	0.17	0.71	0.27	0.74
GMM + Covariate (k=8)	0.64	0.44	0.52	0.62	0.12	0.83	0.21	0.73
Blood counts + Covariate	0.75	0.33	0.46	0.66	0.15	0.5	0.23	0.67
K-means (k=3) + Covariates + Risk Factors	0.59	0.64	0.61	0.64	0.09	0.95	0.16	0.75

Table 5: Results of random forest for cancer prediction from various features. Covariates are age, sex, BMI and smoking status. Risk factors are mood swings, alcohol drinking status, traffic intensity on the nearest major road and housing score. In each feature set, blood counts are represented either in their raw form, or in a clustered form using K-means or GMM clustering. Abbreviations: F1 - F1-score, AUC - Area under the receiver operator curve, GMM - Gaussian Mixture Model.

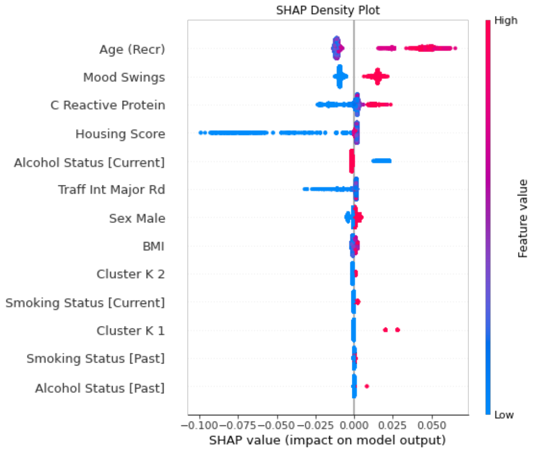

Figure 7: Directional SHAP value plots of the K_3-means random forest model for blood cancer prediction

Model	Case	Train-Precision	Train-Recall	Train-F1	Train-AUC	Test-Precision	Test-Recall	Test-F1	Test-AUC
Covariate	Leukemia Lymphoma	0.41 0.42	0.52 0.08	0.46 0.14	0.57	0.07 0.03	0.77 0.05	0.13 0.04	0.68
K-means + Covariate (k=3)	Leukemia Lymphoma	0.40 0.36	0.55 0.13	0.46 0.19	0.58	0.07 0.03	0.77 0.08	0.13 0.05	0.69
Blood counts + Covariate	Leukemia Lymphoma	0.64 0.43	0.32 0.35	0.42 0.39	0.66	0.11 0.06	0.48 0.36	0.18 0.10	0.64
K-means (k=3) + Covariate + Risk Factors	Leukemia Lymphoma	0.42 0.44	0.35 0.11	0.38 0.18	0.60	0.09 0.05	0.72 0.09	0.16 0.06	0.68

Table 6: Results of multinomial logistic regression for cancer prediction from various features, split by cancer type. Covariates are age, sex, BMI and smoking status. Risk factors are mood swings, alcohol drinking status, traffic intensity on the nearest major road and housing score. In each feature set, blood counts are represented either in their raw form, or in a clustered form using K-means. Abbreviations: F1 - F1-score, AUC - Area under the receiver operator curve.

6 Results: Predicting cluster membership (R3)

For our final research question, we investigate which lifestyle variables are associated with cluster membership. We will consider K-means clustering for K=3. However, before we do this, we describe our clusters in terms of their blood counts in order to understand the composition of the clusters.

6.1 Describing our clusters

Table 7 shows the prevalence of controls and cases (stratified by blood cancer type) in each K_3-means cluster. Cluster 0 and 2 showed similar prevalences, whereas cluster 1 was much smaller (37 samples) and consisted almost entirely of cases (97%), with a higher proportion of leukemia cases (78%) than the other clusters (43%). The contingency tables for the other clustering algorithms are given in the Appendix.

	Cluster 0	Cluster 1	Cluster 2
Controls	5379 (75.43%)	1 (2.70%)	5277 (71.22%)
Cases	1752 (24.57%)	36 (97.30%)	2132 (28.78%)
Leukemia	750 (42.81%)	28 (77.78%)	921 (43.20%)
Lymphoma	1002 (57.19%)	8 (22.22%)	1211 (56.80%)
Total	7131	37	7409

Table 7: Prevalance of cases, controls, leukemia and lymphoma status in each K_3-means cluster.

Cluster 1 is made up of a very small proportion of controls compared to cluster 0 and 2. Such a cluster also exists for the other clustering algorithms which can be seen in tables 8, 9 and 10 in the Appendix. Therefore, cluster 1 is analysed in more detail. In order to understand why cluster 1 differs from the other clusters, box and whisker plots for various blood counts were generated. Figure 8 shows that mean white blood cell count, basophil count and monocyte count are higher in cluster 1 than in the other clusters, most prominently the white blood cell count. Analogously, the same cluster with a high white blood cell count is present in figure 10 in the Appendix for K_6 and K_7.

6.2 Cluster prediction from lifestyle variables

9 shows the association between lifestyle variables and cluster membership is calculated by logistic regression. All variables showed a significant cluster association except for traffic incidence. C reactive protein was significantly associated with membership of group 1. Alcohol intake and housing score was associated with membership in the other clusters.

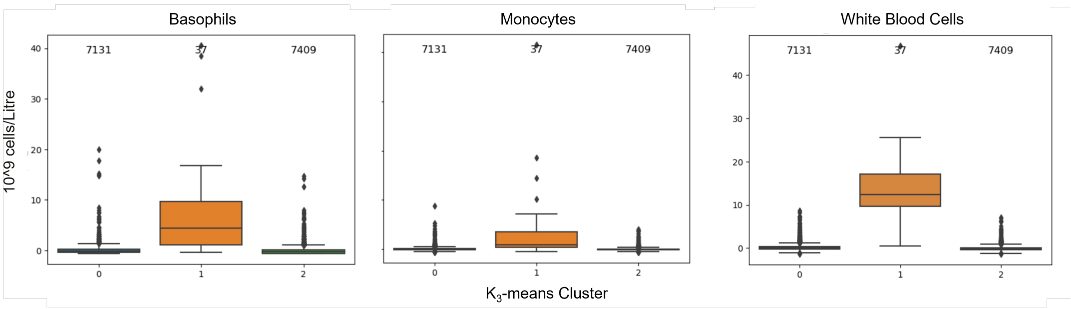

Figure 8: Box plots for the composition of K_3 clusters by blood counts

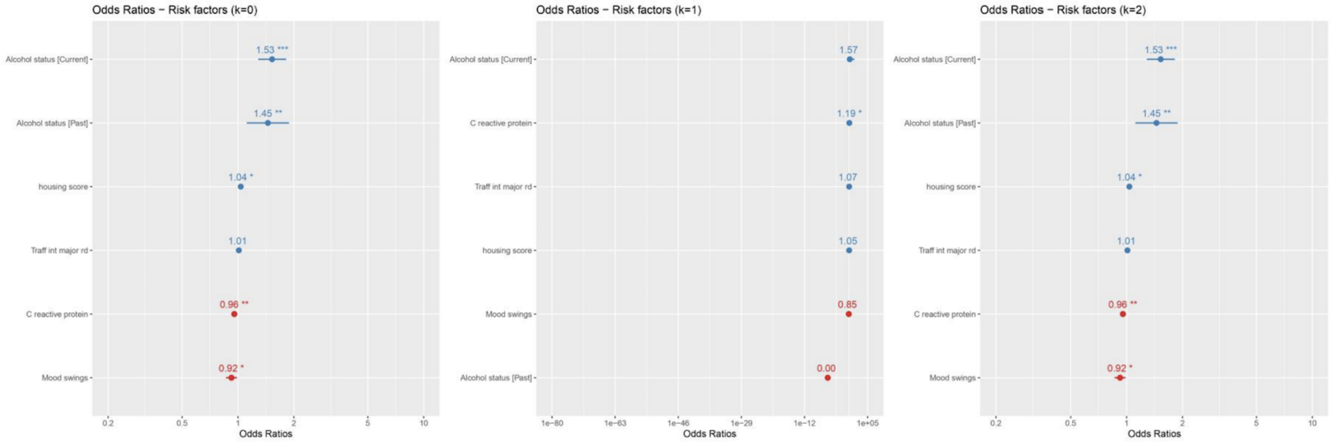

Figure 9: Odds ratios of cluster membership by lifestyle factors. Results were obtained using three multivariate binary logistic regression models. K=0, 1 and 2 indicates cluster membership for cluster 0, 1 and 2 (respectively) of the K$_3$-model. Abbreviations: Traff int major rd - Traffic intensity on the nearest major road. Associations are significant at an alpha of 0.05 (*), 0.01 (**) or 0.001 (***).

7 Discussion

Clustering

We identified clusters of people with shared blood characteristics through K-means and GMM clustering. For K-means, we used the silhouette score to identify K=3 as the optimal number of clusters, and consensus clustering to identify K=6 and K=7. For GMM on the other hand, we used the BIC score, BIC takes into account the fact that the data is assumed to come from Gaussian distributions by using the likelihood function. The elbow method determined that K=8 was optimal for GMM. The silhouette scores for K-means were low (<0.2) and the BIC scores for GMM were high (>100,000), indicating that the results may not have been robust. This may be partially explained due to the fact that there was a failure in the Hopkins test.

Blood cancer prediction

We used both logistic regression as well as random forest models to predict blood cancer from cluster membership, raw blood counts and covariates. Binary logistic regression models had similar AUC scores of approximately 0.6, but low F1 scores. Recall was high, but precision was low, indicating high false positive rates. Both clustering and blood count models performed similarly in terms of F1 scores. Random forest showed slightly higher and F1 scores but still had high false positive rates. For our multi-class model, leukaemia results were better than lymphoma.

Cluster description and prediction

We examined the K_3 clusters in more detail, and noticed that cluster 1 was significantly smaller than cluster 0 and 2, and consisted almost entirely of cases. We saw that the other K-means algorithms also had a similar cluster with these properties. When analysing cluster 1 in more detail, we noticed a higher mean white blood cell count, basophil count and monocyte count than cluster 0 and 2. This suggests that there may be a small subset of cases that are distinct from the other cases.

Logistic regression was performed in order to determine which lifestyle factors affect cluster membership. We found that alcohol status and housing score were significantly associated with cluster 0 and 2, but not with cluster 1. On the other hand, C reactive protein was positively associated with membership of cluster 1, and negatively associated with 0 and 2. Mood swings were significantly associated with clusters 0 and 2, but not 1.

Key findings

Our main findings of our research were as follows:

1. It is possible to group people with shared blood characteristics together, however the prediction of blood cancer from these groups is weak.
2. There is a small group of blood cancer cases that have different characteristics to other cases (mean white blood cell, basophil and monocyte count).
3. Leukaemia may be easier to predict than lymphoma.
4. Raw blood counts taken at baseline are not strongly predictive of blood cancer.

Limitations

The choice of dataset (i.e. Biobank) comes with certain limitations such as lack of diversity and selection bias. Moreover, we were unable to remove batch effects on biomarker measurements. One major constraint of the biobank is the fact that blood count measurements were only taken at baseline. In some cases, it would have been a number of years before cancer developed, meaning that these blood samples were not reliable. In order to avoid potential confounding, we used controls without any ICD codes present, however this means that the controls may not have been representative of the general population. We also had difficulty removing the effects of covariates via matching. For example, age still showed high feature importance according to SHAP.

Improvements and suggestions for future research

We suggest conducting similar research, but use blood samples at regular intervals rather than only samples taken at baseline. We expect this to improve the predictive performance of the models. Moreover, more advanced modelling techniques such as neural networks could be used in order to better capture nonlinear relationships in the data. Lastly, we suggest that future researchers consult with domain experts to incorporate existing knowledge about blood cancer mechanisms into the model.

Further Information

All code used in the analysis can be found on the following GitHub repositories: Haileygu99/Blood-cancer-risk-and-prediction.

References

[1] *Blood Cancer UK*, 2023. Accessed on: 24/02/2023. `https://bloodcancer.org.uk/understanding-blood-cancer/what-is-blood-cancer/`

[2] Lee, Jaewoong and Cho, Sungmin and Hong, Seong-Eui and Kang, Dain and Choi, Hayoung and Lee, Jong-Mi and Yoon, Jae-Ho and Cho, Byung-Sik and Lee, Seok and Kim, Hee-Je and et al., *Integrative analysis of gene expression data by RNA sequencing for differential diagnosis of acute leukemia: Potential application of machine learning*, Frontiers in Oncology, vol. 11, 2021. `https://doi.org/10.3389/fonc.2021.717616`

[3] Hwang, KB., Cho, DY., Park, SW., Kim, SD., Zhang, BT., *Applying Machine Learning Techniques to Analysis of Gene Expression Data: Cancer Diagnosis*, In: Lin, S.M., Johnson, K.F. (eds) Methods of Microarray Data Analysis, Springer, Boston, MA, 2002. `https://doi.org/10.1007/978-1-4615-0873-1_13`

[4] Ghaderzadeh, Mustafa and Asadi, Farkhondeh and Hosseini, Azamossadat and Bashash, Davood and Abolghasemi, Hassan and Roshanpour, Arash, *Machine learning in detection and classification of leukemia using SMEAR BLOOD IMAGES: A systematic review*, Scientific Programming, 2021, pp. 1–14, vol. 2021. `https://doi.org/10.1155/2021/9933481`

[5] Jagadev, P. and Virani, H. G., *Detection of leukemia and its types using image processing and machine learning*, 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, India, 2017, pp. 522-526. `https://doi.org/10.1109/ICOEI.2017.8300983`

Appendix

Cluster	0	1	2	3	4	5
Controls	1 (2.78%)	2200 (68.43%)	2226 (78.88%)	1114 (65.34%)	3544 (77.57%)	1572 (70.49%)
Cases	35 (97.22%)	1015 (31.57%)	596 (21.12%)	591 (34.66%)	1025 (22.43%)	658 (29.51%)
Leuk	27 (77.14%)	419 (41.28%)	247 (41.44%)	300 (50.76%)	396 (38.63%)	310 (47.11%)
Lymph	8 (22.86%)	596 (58.72%)	349 (58.56%)	291 (49.24%)	629 (61.37%)	348 (52.89%)
Total	36	3215	2822	1705	4569	2230

Table 8: K_6-means contingency table

Cluster	0	1	2	3	4	5	6
Controls	1826 (72.78%)	1096 (64.97%)	1 (2.78%)	3576 (77.09%)	6 (1%)	2441 (79.15%)	1711 (65.41%)
Cases	683 (27.22%)	591 (35.03%)	35 (97.22%)	1063 (22.91%)	0 (0%)	643 (20.85%)	905 (34.59%)
Leuk	313 (45.83%)	295 (49.92%)	27 (77.14%)	423 (39.79%)	0 (0%)	251 (39.04%)	390 (43.09%)
Lymph	370 (54.17%)	296 (50.08%)	8 (22.86%)	640 (60.21%)	0 (0%)	392 (60.96%)	515 (56.91%)
Total	2509	1687	36	4639	6	3084	2616

Table 9: K_7-means contingency table

Cluster	0	1
Controls	262	10395
Cases	508	3412
Leuk	376	1323
Lymph	132	2089
Total	770	13807

Table 10: GMM contingency table

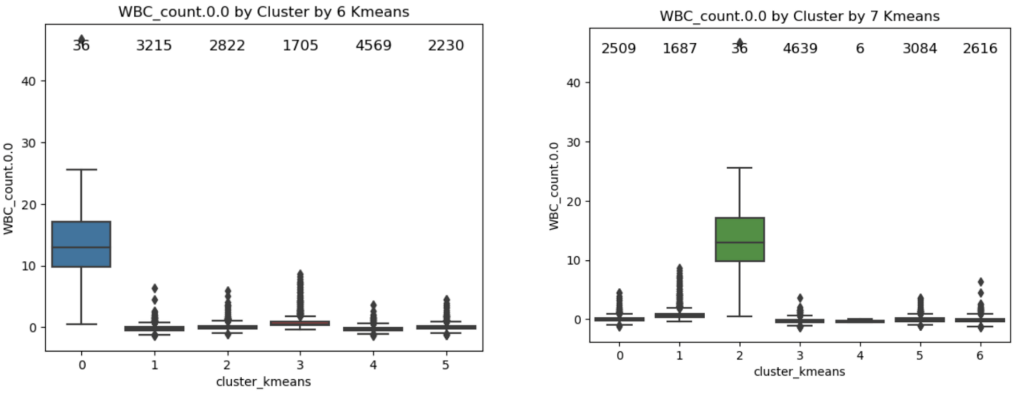

Figure 10: Box plots demonstrating composition of white blood cells in K-means clusters

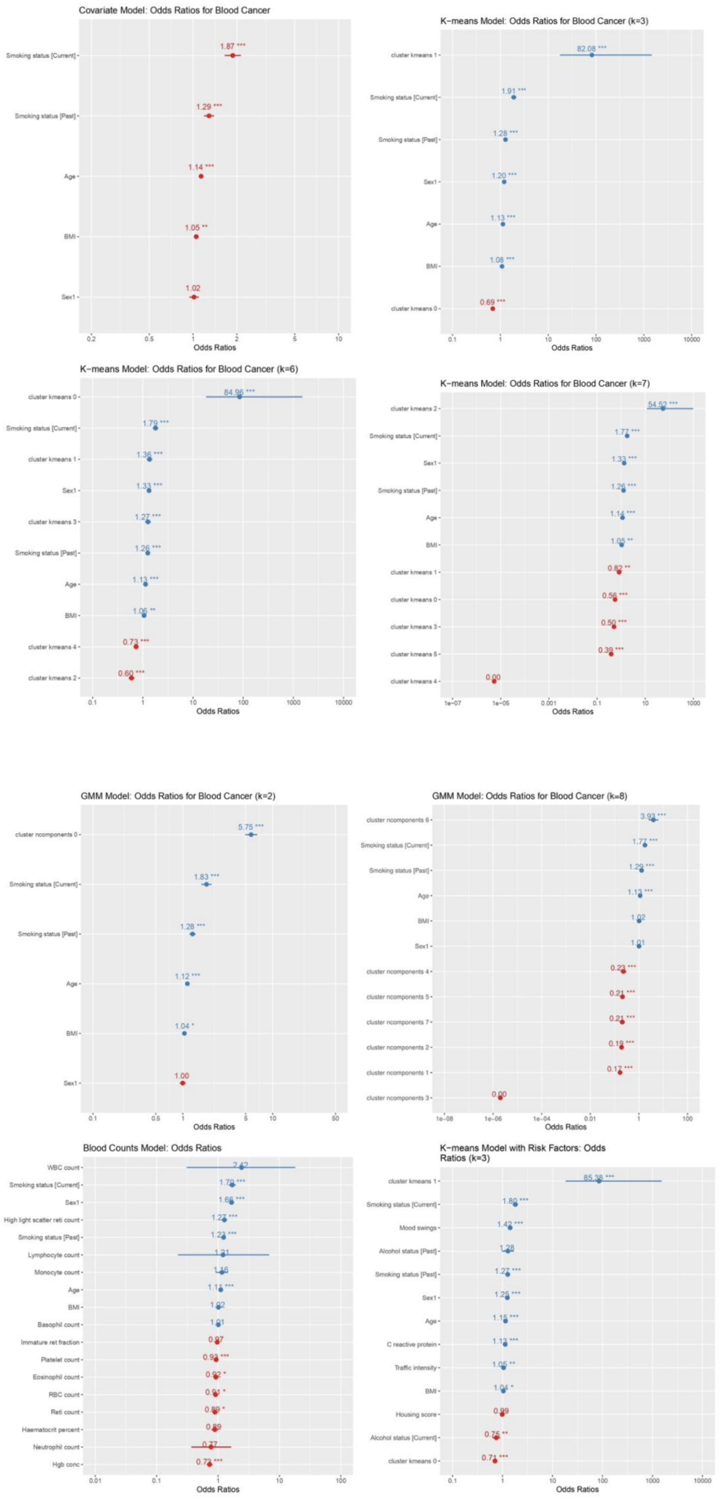

Figure 11: Odds ratios for logistic regression on training set

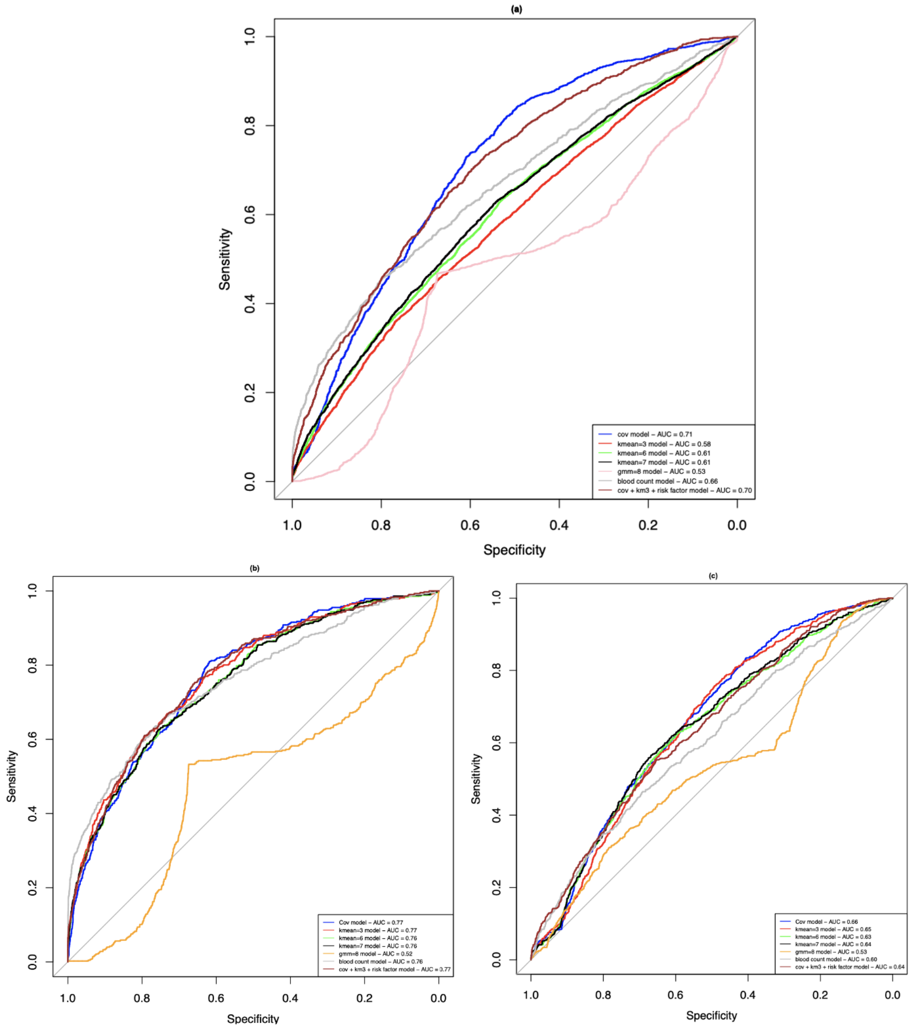

Figure 12: ROC curves for logistic regression on training set (a)case/control (b)leukemia (c)lymphoma