02-620 Machine Learning for Scientists

Prediction of Leukemia Subtypes from Gene-Expression Data: A Machine Learning Approach

Harihara Prakash, Wrootchit Mishra, Zubin Roy

28th April 2023

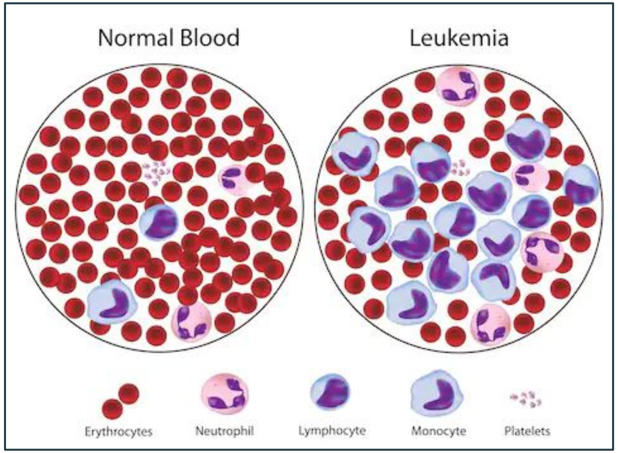
Introduction

Leukemia: A type of blood cancer which originates from the bone marrow

Classified into 4 primary types:

- Acute lymphoblastic leukemia (ALL)
- Acute myeloid leukemia (AML)
- Chronic lymphocytic leukemia (CLL)
- Chronic myeloid leukemia (CML)

Early prediction of leukemia is crucial to successful treatment.



Machine learning can play a critical role in predicting leukemia by analyzing large amounts of gene expression data to identify patterns and predict risk factors for developing cancer.

How can we use machine learning to our benefit to identify leukemia subtypes (or any cancer with subclasses) based on gene expression data?

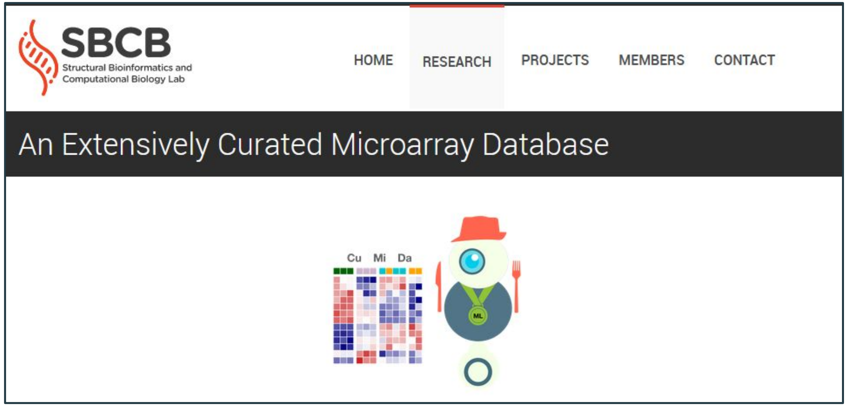
Selected Dataset

We used the Curated Microarray Database (CuMiDa), to identify prospective gene expression datasets for leukemia.

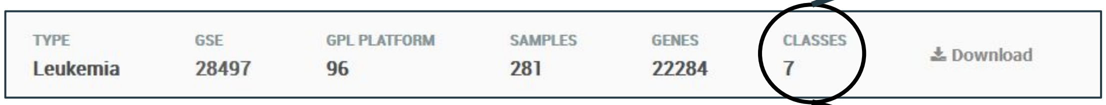
They have an extensive catalog of not just leukemia data, but various cancers including heart, brain, liver and many more.

The database offers manually and carefully curated datasets with background correction and normalization already done.

Our selected dataset has 22284 genes and 281 samples.



An Extensively Curated Microarray Database



TYPE	GSE	GPL PLATFORM	SAMPLES	GENES	CLASSES	
Leukemia	28497	96	281	22284	7	Download

GEO Accession: GSE28497

- B-CELL_ALL
- B-CELL_ALL-TCF3-PBX1
- B-CELL_ALL_HYPERDIP
- B-CELL_ALL_HYPO
- B-CELL_ALL_MLL
- B-CELL_ALL_T-ALL
- B-CELL_ALL_ETV6-RUNX1

Machine Learning Techniques

Logistic Regression

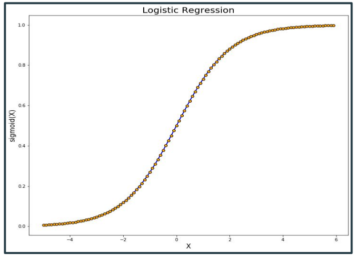
- Good for high dimensional datasets.
- Can carry out multi-class classification problems using several approaches such as one-vs-rest (OVR).

k-Nearest Neighbours (kNN)

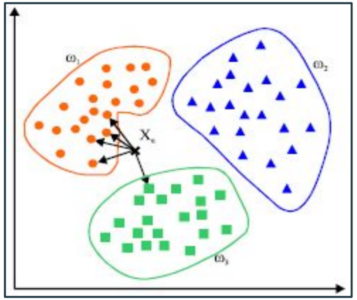
- Simple and easy to interpret.
- Versatile and can utilise various distance metrics.

Support Vector Machines (SVM)

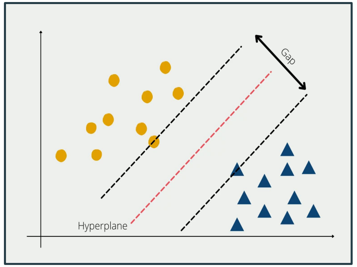
- Can handle high-dimensional data with a relatively small sample size.
- Effective in finding the best boundary between different classes.



LR



k-NN

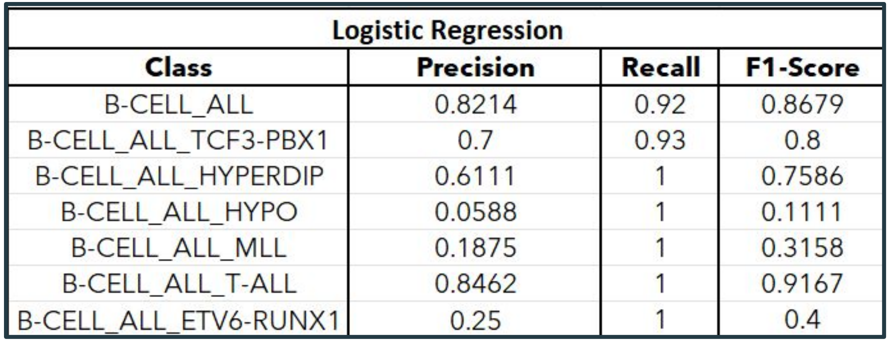


SVM

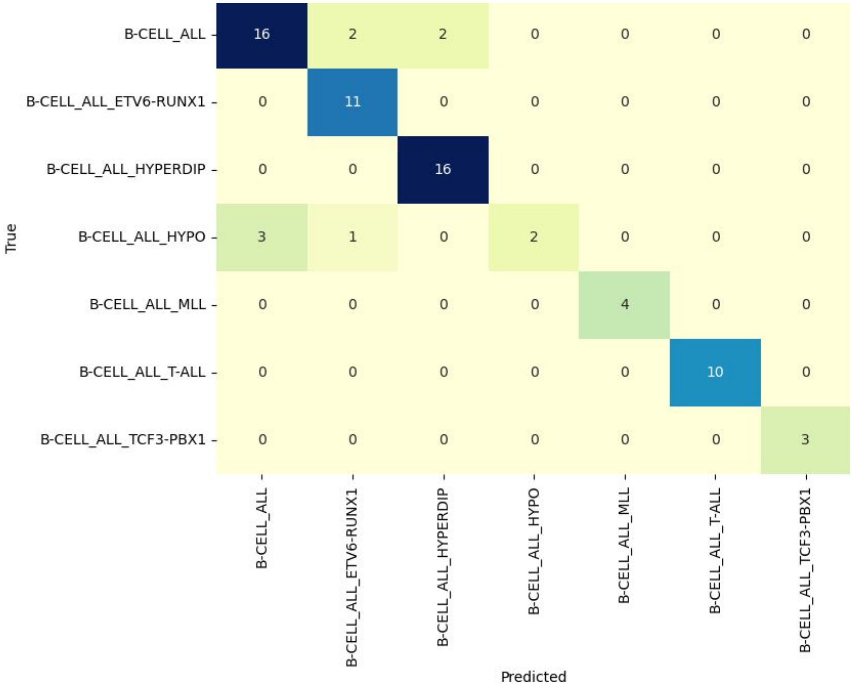
Results of Logistic Regression

The average accuracy across all subtypes for the OVR (one vs rest) LR classifier: 88.16%

The accuracy from 10-fold cross validation: 89.62%

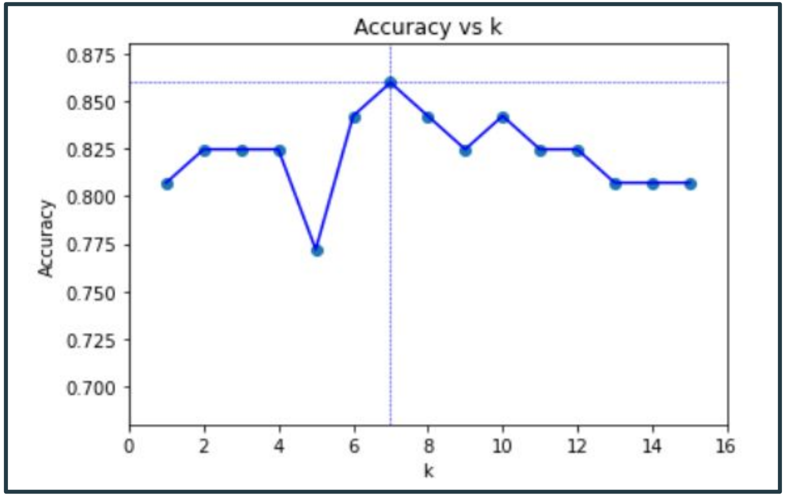


Logistic Regression			
Class	**Precision**	**Recall**	**F1-Score**
B-CELL_ALL	0.8214	0.92	0.8679
B-CELL_ALL_TCF3-PBX1	0.7	0.93	0.8
B-CELL_ALL_HYPERDIP	0.6111	1	0.7586
B-CELL_ALL_HYPO	0.0588	1	0.1111
B-CELL_ALL_MLL	0.1875	1	0.3158
B-CELL_ALL_T-ALL	0.8462	1	0.9167
B-CELL_ALL_ETV6-RUNX1	0.25	1	0.4



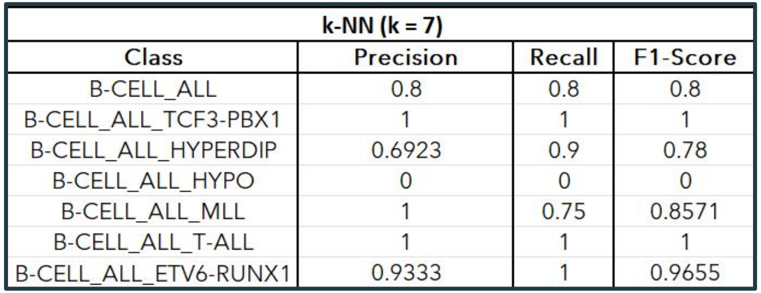
True \ Predicted	B-CELL_ALL	B-CELL_ALL_ETV6-RUNX1	B-CELL_ALL_HYPERDIP	B-CELL_ALL_HYPO	B-CELL_ALL_MLL	B-CELL_ALL_T-ALL	B-CELL_ALL_TCF3-PBX1
B-CELL_ALL	16	2	2	0	0	0	0
B-CELL_ALL_ETV6-RUNX1	0	11	0	0	0	0	0
B-CELL_ALL_HYPERDIP	0	0	16	0	0	0	0
B-CELL_ALL_HYPO	3	1	0	2	0	0	0
B-CELL_ALL_MLL	0	0	0	0	4	0	0
B-CELL_ALL_T-ALL	0	0	0	0	0	10	0
B-CELL_ALL_TCF3-PBX1	0	0	0	0	0	0	3

Results of K-Nearest Neighbours

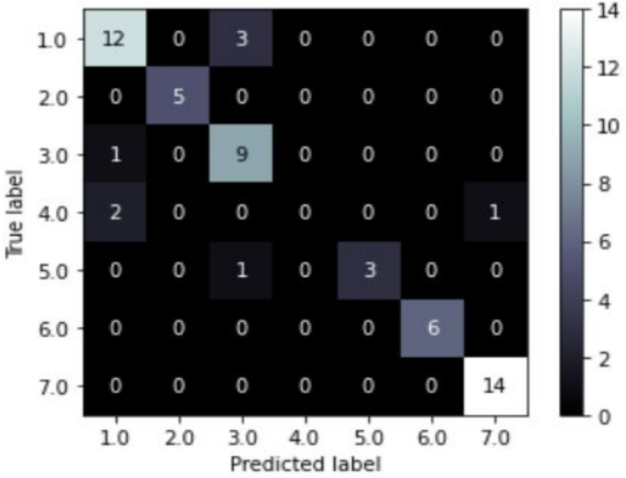


We observed the highest test accuracy of 85.96% for 7 nearest neighbours.

Accuracy obtained by 10-fold cross validation was 83.63%



k-NN (k = 7)			
Class	Precision	Recall	F1-Score
B-CELL_ALL	0.8	0.8	0.8
B-CELL_ALL_TCF3-PBX1	1	1	1
B-CELL_ALL_HYPERDIP	0.6923	0.9	0.78
B-CELL_ALL_HYPO	0	0	0
B-CELL_ALL_MLL	1	0.75	0.8571
B-CELL_ALL_T-ALL	1	1	1
B-CELL_ALL_ETV6-RUNX1	0.9333	1	0.9655



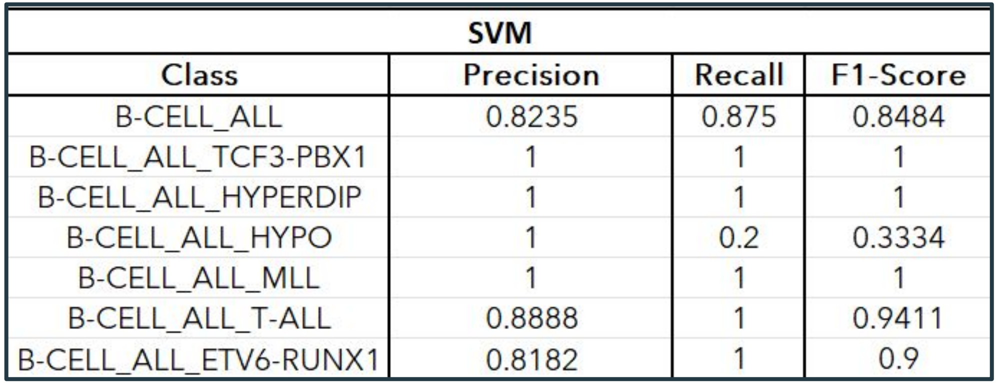
True label	1.0	2.0	3.0	4.0	5.0	6.0	7.0
1.0	12	0	3	0	0	0	0
2.0	0	5	0	0	0	0	0
3.0	1	0	9	0	0	0	0
4.0	2	0	0	0	0	0	1
5.0	0	0	1	0	3	0	0
6.0	0	0	0	0	0	6	0
7.0	0	0	0	0	0	0	14

Results of SVM

Scikit Class: sklearn.svm.SVC()

Kernel: Linear

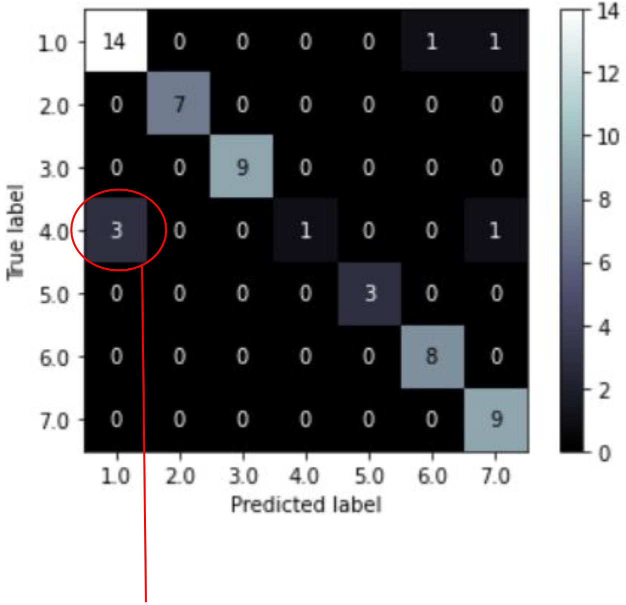
Decision shape function: One versus rest



SVM			
Class	Precision	Recall	F1-Score
B-CELL_ALL	0.8235	0.875	0.8484
B-CELL_ALL_TCF3-PBX1	1	1	1
B-CELL_ALL_HYPERDIP	1	1	1
B-CELL_ALL_HYPO	1	0.2	0.3334
B-CELL_ALL_MLL	1	1	1
B-CELL_ALL_T-ALL	0.8888	1	0.9411
B-CELL_ALL_ETV6-RUNX1	0.8182	1	0.9

Accuracy from the model = 89.47%

Accuracy from 10-fold cross validation = 88.60%



	1.0	2.0	3.0	4.0	5.0	6.0	7.0
1.0	14	0	0	0	0	1	1
2.0	0	7	0	0	0	0	0
3.0	0	0	9	0	0	0	0
4.0	3	0	0	1	0	0	1
5.0	0	0	0	0	3	0	0
6.0	0	0	0	0	0	8	0
7.0	0	0	0	0	0	0	9

Subtype 4 is most commonly misclassified as Subtype 1

Conclusion

All three models we implemented have fairly good accuracy possess equal potential in serving as a cancer subtype classifier.

Machine Learning Model			
	Logistic Regression	k-NN (k = 7)	SVM
Model Accuracy	88.16%	85.96%	89.47%
Cross Validation Accuracy	89.62%	83.63%	88.60%

Based on the results obtained from all three classifiers, we observe that B-CELL_ALL_HYPO (subtype 4) is most often mis-classified as B-CELL_ALL (subtype 1).

These models can be utilised for gene expression data of many other cancers with multiple subtypes.

Future Work

While our results support the idea that SVM, Logistic regression and k-NN are good classifiers for multiclass cancer subtype prediction, certain aspects cannot be overlooked.

- Relatively low sample to feature ratio (281 samples, 22284 features)
- Lack of feature selection and regularization

In addition, better machine learning classifiers are probable to exist and further testing to identify which other model works best for classification would be necessary.

It would be ideal to check the performance of other multiclass models such as random forest, decision trees etc.

References

- Coustan-Smith E, Song G, Clark C, Key L et al. New markers for minimal residual disease detection in acute lymphoblastic leukemia. Blood 2011 Jun 9;117(23):6267-76. PMID: 21487112
- Feltes, B.C.; Chandelier, E.B.; Grisci, B.I.; Dorn, M. CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research. Journal of Computational Biology, 2019.
- Rupapara V, Rustam F, Aljedaani W, Shahzad HF, Lee E, Ashraf I. Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model. Sci Rep. 2022 Jan 19;12(1):1000. doi: 10.1038/s41598-022-04835-6.
- Simsek, E., Badem, H., Okumus, I.T. (2022). Leukemia Sub-Type Classification by Using Machine Learning Techniques on Gene Expression. In: Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds) Proceedings of Sixth International Congress on Information and Communication Technology. Lecture Notes in Networks and Systems, vol 217. Springer, Singapore. https://doi.org/10.1007/978-981-16-2102-4_56