I read the entirety of the article.

The authors of the study aimed to improve cancer classification, specifically by creating a general approach for classifying cancers based on gene expression by monitoring DNA microarrays. Their class discovery procedure specifically looked for and classified between acute myeloid leukemia and acute lymphoblastic leukemia (AML vs. ALL) as a test case. This choice of acute leukemias is significant because the distinction between the two has been well established, however there wasn’t a single test available at the time to distinguish between them. It’s important to know the specific diagnosis between the two as the chemotherapy regimens differ significantly. The authors used high density oligonucleotide arrays to measure gene expression levels of 6817 genes from 38 bone marrow samples taken from leukemia patients at the time of diagnosis. They used a few machine learning techniques for class prediction and class discovery. The main technique used for class prediction was an unsupervised learning approach called self-organizing maps (SOMs) which clustered the samples based solely in the gene expression data. The main technique used for class prediction involved a supervised learning approach that selected a subset of informative genes associated with the distinction between AML and ALL to build a predictive model. It actually used a form of ensemble learning involving weighted voting. In this method, each gene cast a vote based on expression level and degree of correlation with a class distinction between the two classes. They used cross-validation and tested the model on an independent dataset to validate the models further. The results showed that they could accurately classify AML and ALL cases with super high accuracy. Given this success, they furthered the method by showing that it could also distinguish between subclasses of ALL, B-cell and T-cell. All in all, great result.

First of all, this seems like somewhat of an older paper for machine learning to be used in a biological setting, so it seemed well done and novel especially for the time. One thing that particularly stood out about the success of the models was the fact that they only used gene expression, without any knowledge of other biological processes that were previously employed for the diagnosis and distinction between these two leukemias. This emphasizes the power of machine learning in biology and medicine. I also thought that their feature selection process was super interesting and particularly well done as they reduced the dimensionality of the data by selecting only a few particularly informative genes associated with the distinction between AML and ALL. This minimized noise, surely improving the accuracy of the model, as well as made it more computationally feasible. It was also cool that they decided to keep going with their model as it performed so will in their main task, that they further separated the leukemias into sub-subclasses.

This study was great for its time and employed some of the most cutting-edge machine learning techniques, but with the advancement of machine learning, there are definitely, at least now, a few ways to improve the way they went about things. The first thing I noticed was the size of the sample for the training of the model. They used 38 samples, which is relatively small for what is generally required for machine learning models currently to get the best results. There may have been some overfitting to the data they had access to because of this. Though I found the feature selection process interesting, choosing the top 50 genes based on correlation, I feel that maybe they could have used PCA to try to reduce dimensionality yet keep as much of the data as possible. Not sure how widespread PCA was at the time, but perhaps something to look at for the future. In addition, methods in deep learning like the ones we are learning about, SVM or RF could probably have gotten the most out of the data without having to cut it down so severely.

This was my own work