Hamed Hekmat 12/7/23

CS 279 Final Project

David vs Goliath: Comparing the Performance of ML Models for Leukemia Cell Classification

Introduction

There's been a recent increase in demand for non-invasive diagnostic support tools for Acute Lymphoblastic Leukemia (ALL). As a result, the Cancer Imaging Archive (TCIA) released a dataset of microscopy images of individual cells from blood clotting for a 2019 competition involving the classification of normal vs malignant cells (C-NMC). Many publications explored computational solutions through complex neural network architectures for classifying the labeled cells, achieving over 90% accuracy (Zakir Ullah et al. 2021). In order to quantitatively assess why this task is perceived to be so difficult, I also sought to conduct principle component analysis (PCA) of the malignant and normal cells. Additionally, as an individual lacking the computing power or formal machine learning training to develop a similarly sophisticated model of my own (relative to other entries in the competition), I was curious to see how well I could perform on this task utilizing out-of-the-box ML packages. Overall, I pursued this project to achieve a better understanding of the clinical implications of developing such a tool, as well as comparing how accessible ML tools and packages stack up against complex architectures designed intentionally for this specific task.

Background

Leukemia is a cancer of the blood-forming tissues and mechanisms of the body such as the bone marrow. In ALL, as mutations lead to the overproduction of lymphoblasts by the bone marrow, other essential blood cell types get crowded out (Yale Medicine n.d.). As of 2020,

estimates showed leukemia was among the top 15 causes of both cancer-related incidence and mortality worldwide. Projections indicate that age-standardized incidence and death rates will only continue to climb through 2030 (Du et al. 2022). To diagnose leukemia and distinguish ALL from other types, doctors often rely on invasive and painful procedures such as bone marrow aspirations/biopsies or lumbar punctures to get a holistic sense of blood cell type distribution/counts. Such procedures are often necessitated due to the difficulty of making these distinctions by simply having experts examine individual cells from blood samples under a microscope, as malignant immature lymphoblasts look very similar to healthy lymphocytes (Zakir Ullah et al. 2021).

However, early diagnosis and rapid intervention is key, so physicians would ideally have a tool to determine if individual cells are malignant before needing to view a whole bone marrow sample. This provides the motivation for the TCIA's releasing the dataset and creating a competition to determine if computer vision and computational classification methods can resolve this issue (Zakir Ullah et al. 2021).

Methods	Accuracy	Year
NASNet-Large with VGG19 [36]	0.965	2020
Hybrid model (VGG16 + MobileNet) [37]	0.961	2019
LeukoNet [38]	0.896	2018
Proposed Method	**0.911**	**2021**

Methods	F1-Score
SDCT-AuxNet [35]	0.948
Neighborhood-correction algorithm (NCA) [54]	0.910
Ensemble model based on MobileNetV2 [55]	0.894
Deep Multi-model Ensemble Network (DeepMEN) [50]	0.885
Ensemble CNN based on SENet and PNASNet [56]	0.879
Deep Bagging Ensemble Learning [57]	0.876
LSTM-DENSE [58]	0.866
Ensemble CNN model [59]	0.855
Multi-stream model [60]	0.848

The tables above present some of the top entries and their performance using two metrics from the original competition (Zakir Ullah et al. 2021).

Methods

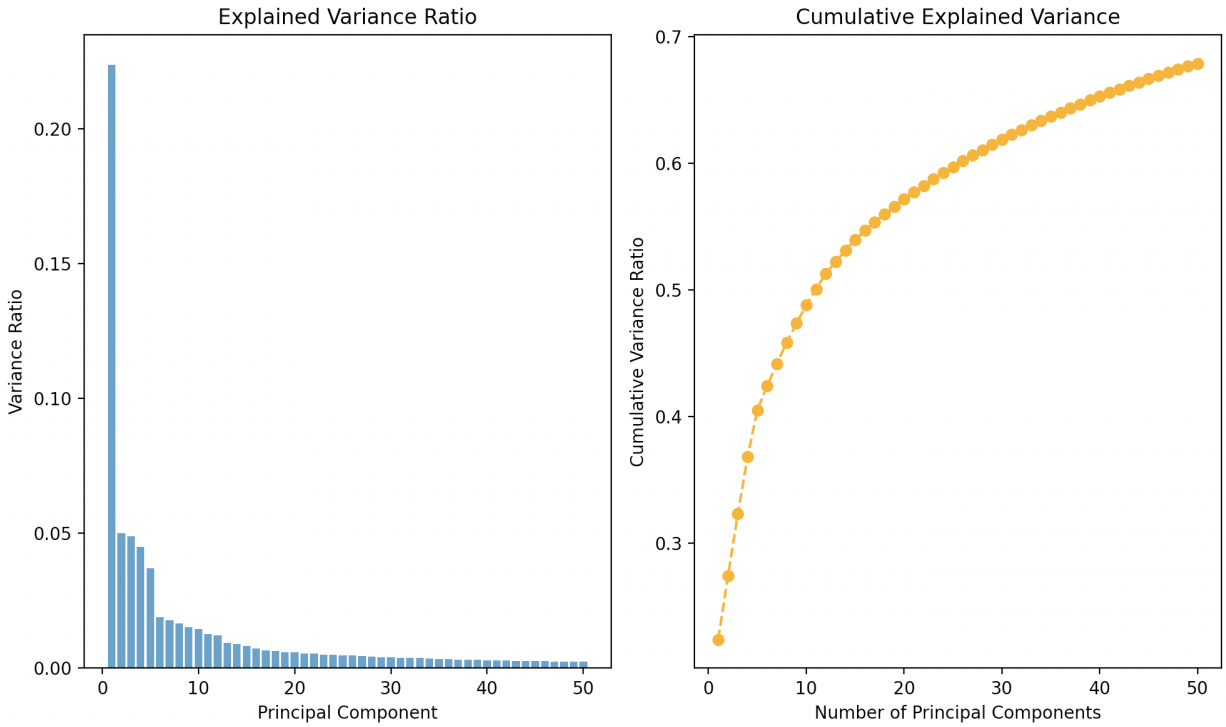
For the sake of brevity and to fit within the scope of this project, I narrowed my point of comparison to one paper, namely *An Attention-Based Convolutional Neural Network for Acute Lymphoblastic Leukemia Classification* by Zakir Ullah et al. published in *Applied Science* in 2021. I obtained the dataset from the TCIA competition through Kaggle, with final counts of 7272 malignant cell images and 3389 normal ones. In their work, the authors of the aforementioned paper employed a handful of data preprocessing and augmentation techniques, such as including rotated versions of the same image, with 4 rotations for healthy cells and 2 for cancerous ones in order to balance out the size of the dataset. I opted to forego such steps in order to assess the performance of the built-in models with minimal adjustments/tuning. Due to subpar performance from my laptop (I probably need a new one soon), I did have to use a smaller subset of the available data, so I chose to use 2500 malignant cells and 2500 normal ones as a proxy for Zakir Ullah et al.'s more systematic approach to remedying the data imbalance. Using cv2, I processed the images in grayscale (as I was also curious about whether color would have a noticeable effect on performance), flattened them, and converted them into NumPy arrays.

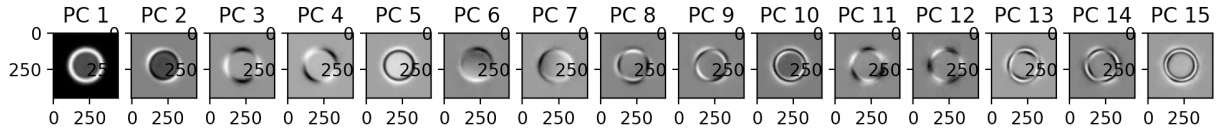
For PCA, I utilized the PCA module from sci-kit-learn's built in decomposition package, allowing the user to input how many components for the algorithm to compute. I ran my PCA function separately on the cancerous datapoints and the healthy datapoints. Additionally, I used matplotlib to generate charts containing the explained variance ratios and cumulative explained variance, along with to reconstruct the primary components (however many the user inputs) in image form.

As for my competing classification model, I used the built-in random forest classifier from sci-kit-learn's ensemble package. I deliberately selected this approach to contrast Zakir Ullah et al.'s deep learning approach. Simply put, random forest classifiers generate some number of randomized decision trees (100 by default) created from bootstrapped slices of the data, with each decision tree accepting a full datapoint and labeling it as 1 or 0 (cancerous or healthy). After passing a given datapoint from my test set through the entirety of the forest, the algorithm accumulates the "votes" from each tree, assigning a final label corresponding to the majority vote (IBM n.d.). I also began to implement a basic neural network using PyTorch to compare it to the much more complex deep learning architecture from Zakir Ullah et al., but I've decided to omit its results from this report for the sake of brevity.

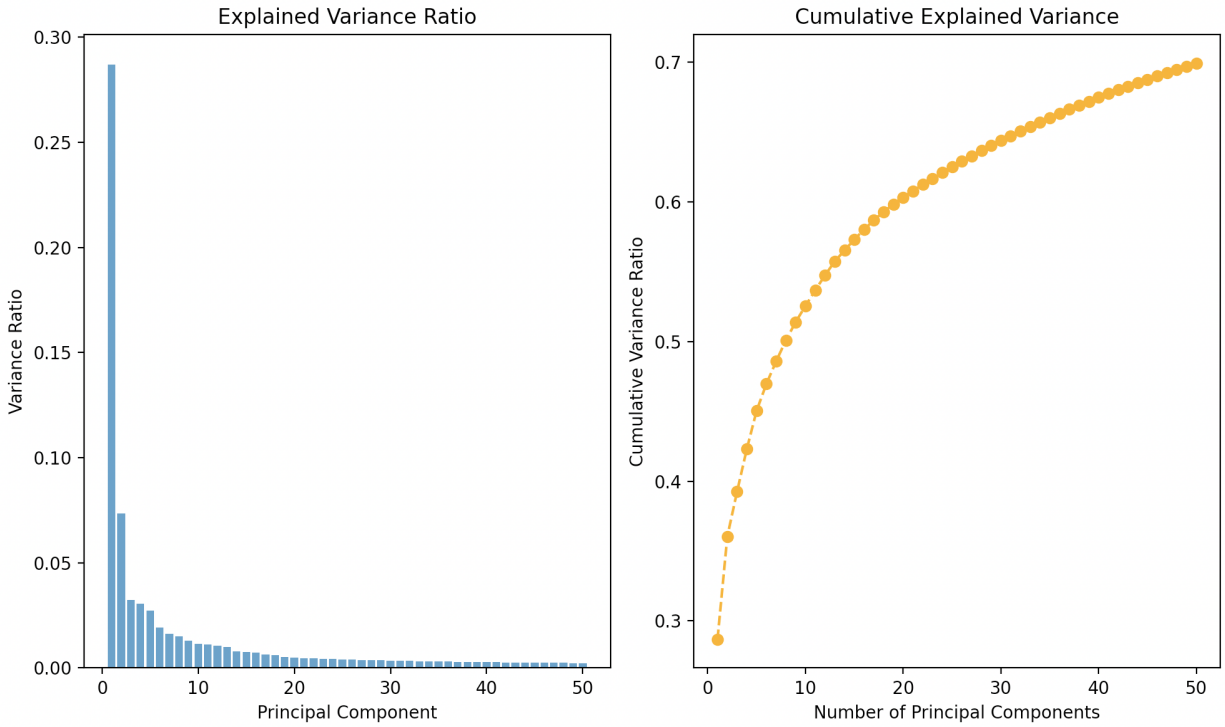
Results and Analysis

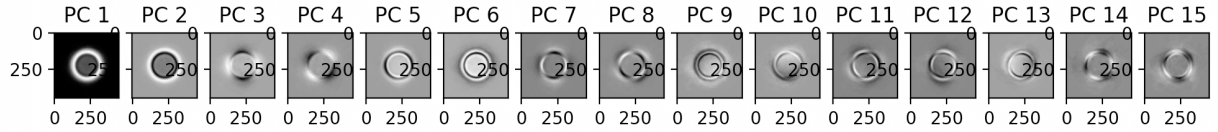
Beginning with PCA, below are the results for malignant cells.





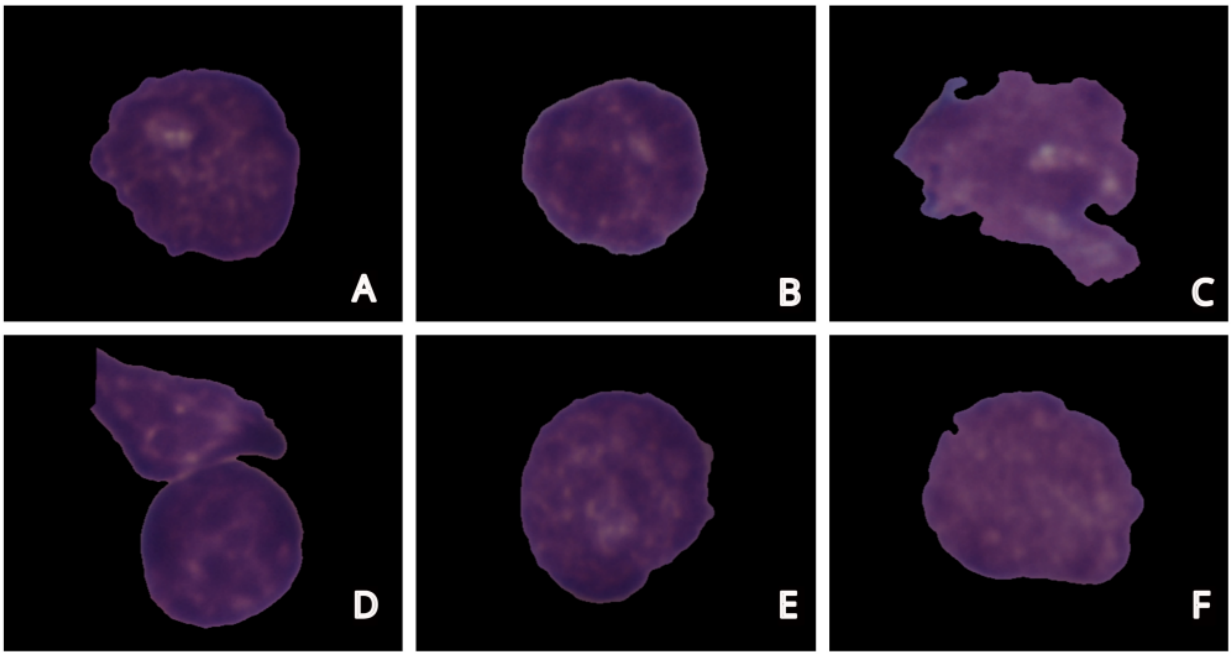
Note that the first 5 principle components seem all contribute more highly to the explained variance as there's an immediate dropoff in the ratio chart, while the first 10 account for 50% half of the cumulative explained variance. Now, here are the results for normal cells.





In this case, the dropoff appears to occur slightly earlier (after the second principle component), and it takes fewer (7) principle components to account for roughly 50% of the explained variance. This initial observation slightly confirms the intuitive assumption that the healthy cells

would likely possess more regular patterns in their structures, leading to less variability overall, while the cancerous cells demonstrate more complex and/or variable structures. Qualitatively glancing through the data confirms that this might be true, but the distinctions are not significant enough to imply that the classification task would be straightforward.



In the above image, A-C are all malignant, while D-F are examples of normal datapoints. While the majority of datapoints for healthy cells are more uniformly round, there are still a nonnegligible amount of instances like D that have irregular shapes. Simultaneously, while many cancerous cells exhibit obvious irregularities such as C, there are also many that could easily be mistaken as healthy cells to the untrained eye (Zakir Ullah et al. 2021). Examining the visual reconstructions of the first 15 principle components reaffirms the fact that these cells are extremely difficult to identify and classify by hand, validating the need for computational diagnostic support tools.

As for the results of the random forest classifier, the out-of-the-box model had an accuracy of 81%! Right off the bat, I was somewhat shocked to see how built-in methods that

don’t involve deep learning could already attain fairly high results without much tuning or modification. For context, not only did Zakir Ullah et al. use colored images and augment their dataset using rotated versions of samples for robustness, but they also split the data into various folds depending on which subject a given sample was obtained from to account for variability between subjects. They additionally included an attention module called Efficient Channel Attention (ECA) to further supplement their neural network architecture’s ability to extract deep features, boosting the mean accuracy across various folds to 91.1%. In fact, without ECA, their model had a mean accuracy of 83.9%, which isn’t significantly better than the much simpler random forest approach I employed (Zakir Ullah et al. 2021).

While my random forest classifier didn’t perform quite as well as their convolutional neural network model with ECA nor well enough to helpful in a clinical setting, it significantly outperformed my untrained human eye. I believe its ability to do so without all of the bells and whistles in Zakir Ullah et al.’s model relies on the random forest algorithm’s use of stochasticity. As an ensemble learning method, the multiple layers of randomness involved in generating the trees (picking subsets of data, bootstrapping) ensure that the model does not overfit to the datapoints it sees through the training process, as well as avoiding making any assumptions about the data’s distribution. Thus, I’d expect it to generalize fairly well when presented with unseen data, confirmed by it’s solid showing as far as accuracy goes. Pairing this aspect with the fact that decision trees inherently handle complex non-linear relationships (such as those often found in imaging data) much better than methods such as linear regression provides us with concrete reasons as to why the model performed fairly well.

To dive deeper into understanding its performance, below are the basic evaluation metrics of the trained model outputted by the classification report from sci-kit-learn:

	Precision	**Recall**	**F1**	**Support**
0 (Normal)	0.83	0.77	0.80	500
1 (Malignant)	0.79	0.84	0.81	500
Weighted Avg	0.81	0.81	0.81	1000

What's worth noting is that the model had a higher recall on malignant cells than normal ones, which means we had a lower rate of false negatives on the datapoints that'd indicate the cell/sample is cancerous. While it's true that having more false positives among malignant cells (indicated by the precision scores) isn't ideal as it'd lead to further testing/interventions when it isn't required, in a clinical setting I'd much rather prefer avoiding falsely labeling a sample from an individual with leukemia as healthy than vice versa. While the former concern relates to cost-effectiveness and efficient use of resources, the latter pertains to matters of life or death, in which case it's always better safe than sorry. Overall though, having an F1 score (the harmonic mean of precision and recall) that aligns with our accuracy indicates that the model has relatively balanced performance, and due to the aforementioned ethical considerations, we're okay with its slight leaning towards classifying a sample as malignant even when it's not rather than the other way around.

To conclude, my PCA results confirmed the fact that this classification problem originally presented as a C-NMC competition is by no means a straightforward one, but even out-of-the-box classification approaches that don't involve deep learning can achieve moderate performance without much fine-tuning. I've also gained an increased appreciation for the expertise of radiologists/hematologists, as the original dataset was hand-labeled by doctors of

those specialties! Most importantly, while I enjoyed learning about and experimenting with the random forest algorithm, there's an obvious gap between its effectiveness and that of methods that employ deep learning. For future steps, I'd first seek to max out the performance of the random forest algorithm by including color, introducing image filtering and augmentation methods, and tuning the parameter of the number of trees in the forest, before moving onto further experimentation building off my initial PyTorch neural network for this task.

References

Du, Mengbao, et al. "The global burden of leukemia and its attributable factors in 204 countries and territories: Findings from the global burden of disease 2019 study and projections to 2030." *Journal of Oncology*, vol. 2022, 2022, pp. 1–14, https://doi.org/10.1155/2022/1612702.

IBM. "What is random forest?" *IBM.* https://www.ibm.com/topics/random-forest.

Kudin, Alex. "C-NMC_Leukemia." *Kaggle*. https://www.kaggle.com/datasets/avk256/cnmc-leukemia.

Yale Medicine. "Acute Lymphoblastic Leukemia (ALL)." *Yale Medicine*. https://www.yalemedicine.org/conditions/acute-lymphoblastic-leukemia-all#:~:text=And%20because%20the%20bone%20marrow,are%20essential%20to%20good%20health.

Zakir Ullah, Muhammad, et al. "An attention-based convolutional neural network for acute lymphoblastic leukemia classification." *Applied Sciences*, vol. 11, no. 22, 2021, p. 10662, https://doi.org/10.3390/app112210662.