ACUTE LYMPHOBLASTIC LEUKEMIA (ALL) CLASSIFICATION USING CNN-BASED MODEL

Sara Ahmad
Student# 100631263
sarayy.ahmad@mail.utoronto.ca

Mohammed Amir
Student# 1004814668
m.amir@mail.utoronto.ca

Michael Acquaviva
Student# 1007042308
michael.acquaviva@mail.utoronto.ca

Sarp Kose
Student# 1007022534
sarp.kose@mail.utoronto.ca

ABSTRACT

We present our findings on implementing a CNN-based deep learning model for leukaemia detection and classification. Currently, there exist models that effectively differentiate between normal and abnormal leukocytes obtained through blood collection. Such models further classify leukaemic cells into subgroups using transfer learning, k-means algorithm, and multi-class support vector machines (SVM). Making use of a limited dataset of various stages of Acute Lymphoblastic Leukaemia, we present a robust preprocessing process, along with a ResNet18-based convolutional feature extractor to group samples into one of four classes. Evaluating the design shows that employing the use of our CNN in diagnosing ALL greatly increases accuracy compared to an independent medical professional or baseline model.

—-Total Pages: 9

1 INTRODUCTION

Acute lymphocytic leukemia (ALL) is a blood cancer that starts from white blood cells in the bone marrow that become leukemic, these cells multiply uncontrollably and survive better than normal cells (Hirsch, 2021). ALL progresses rapidly throughout the body and can prove to be fatal within months if left untreated. (Hirsch, 2021). In 2020, leukaemia accounted for 3.1% of all cancer related deaths both globally and in the United States, (Junjie Huang, 2022) where it is the 10th most common type of cancer. (Markman, 2022).

A blood smear test is a common leukemia diagnostic procedure which looks for lymphocytes and does a white and red blood cell count (Mohamed Loey, 2020). This process is labour intensive and error prone due to its tedious nature and vulnerability to fatigue. Adding in the amount of time it takes to perform a single test makes building automated, fast, and low-cost systems a necessity.

A machine learning approach to diagnose leukemia will reduce the delays in cancer treatment by reducing the logistical delays concerned with turnaround time and diagnostic delay (Kumar et al., 2022). As the lymphocytes are characterized by their physical appearance including their shape and size, leukemia classification can be done using image classification models.

We built a CNN-based model, described in figure 1, that takes stained blood smears images as input and classify them based on whether they contain cancerous cells: cells that cause ALL. For those classified as containing cancerous cells, the model will detect at what stage of ALL the patient is. A deep learning solution is a good fit to this problem because blood smears can be easily transformed into images which will be used as input for the model. With their multi-level structures, deep learning algorithms can very successfully extract both basic, low-level and complicated, higher-level semantic information from images. This property of the deep learning models makes it possible for

them to yield much better results than traditional image recognition algorithms or humans. (Talluri, 2017)

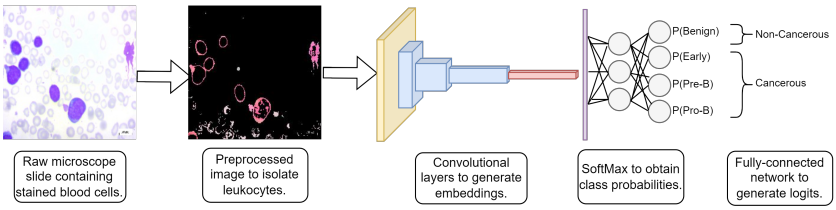


Figure 1: High-level depiction of the project's overarching goal. We begin with raw images of leukocytes, obtained from blood-smear sample. We then preprocess the samples to crop and enhance each image. This step serves to direct our model's attention to lymphocytes. Next, the deep-learning model classifies each sample into one of four categories, based on the presence or lack of cancer, and its stage

2 Background & Related Work

There are several relevant deep learning models that focus on biomedical image classification, including the diagnosis of brain tumors (Haq et al., 2022), malaria parasite detection (Umer et al., 2018), detection of specific gram stain morphologies (Smith et al., 2018), classify leukaemia from blood cells present(Abir et al., 2022), and some further classify the subtype of leukaemia present (Begum & Razak, 2017).

Smith et al. used a transfer learning approach using the Inception 3.0 CNN to recognize common gram stain morphologies in positive blood smears. They achieved a 92.5% whole slide classification accuracy on 100 000 classified images. This work is relevant to our project as the input for both is a blood smear and performs image classification.

Umer, Muhammad, et al. (2020) performed malaria detection with 100% precision, 99.9% recall, and 99% f1-measure using a novel stacked CNN with 22 layers, a ReLU activation function and ADAM optimizer was used.

Abir et al. (2022) presented a classification system that differentiates between normal and leukaemia cells using transfer learning. Firstly, the team used a class weight function to balance a dataset containing more acute lymphocytic leukaemia. To reduce additional effort, pre-trained convolutional neural networks (CNN) including the ResNet101V2, VGG19, InceptionResNetV2, and, InceptionResV3 were verified using TensorFlow history functions. To overcome the requirement of huge amounts of data, a transfer learning technique was utilized. Transfer learning The InceptionV3 achieved the highest test set accuracy of 80.02% on a dataset of 15,144 images.

A classification performed using a multi-class support vector machine, presented by Sajjad et al. (2016) , classified leukocytes into five classes which include neutrophil, eosinophil, basophil, lymphocyte, and monocyte. From blood smear images, a k-means algorithm was used to segment white blood cells (WBCs) and morphological operations to remove irrelevant components. Experiments on a dataset of 1030 blood smear WBC images produced an average accuracy of 98.6% for this framework.

A ResNet based CNN presented by Haq, Amin ul et la (2022) showed a 99.0% accuracy on the classification and diagnosis of brain cancer, showing that a ResNet-Cnn model could be applicable in other biomedical fields.

Doing a review analysis of the field, we built a classification model using a CNN and transfer learning approach that combines the functionality of the detecting whether leukaemia is present, and, if so, to further classify the subtype of leukaemia.

3 Data Processing

3.1 Datasets Used

The dataset we used to train and validate our model was the ALL dataset collected by Aria et al. containing over 3200 peripheral blood smear images of the following classes: Benign, Early, Pre-B, and Pro-B ALL (Aria et al., 2021). There are 500 Benign blood smear images that consist of lymphoblasts that exhibit tumorous qualities, but grow slowly enough to not be considered cancerous. The dataset also contains 980 images of Early ALL where cancer cells have not spread far into the surrounding tissue, but exhibit the precursor-B marker for ALL. 960 of the images are from the Precursor-B, or Pre-B, class where cancerous lymphocytes have effectively matured. Finally, the Pro-B ALL class contains images of a rare sub-type of ALL where B-cells remain immature but are still cancerous and harm nearby cells.

3.2 Pre-Processing

There are two main limitations of our dataset that must be addressed through pre-processing: one being the quality of the data, and the other being the quantity of the data. To circumvent these limitations we use cropping, enhancing and data augmentation as ways of cleaning the data and generating more data. Then we split the processed data into training, validation, and testing sets that each had an equal proportion of each class of ALL.

3.2.1 Cropping and Enhancing

Since the images within our dataset exhibit a wide range of resolutions, these images were cropped so that our input was of consistent dimensionality. The shape of lymphocytes is integral to the ALL classification process, therefore we could not simply resize the input images. Instead, the images had to be first manually cropped to the same proportions, ensuring that the lymphocytes were not cropped out of the blood smear. Then, the resize() function could be used to resize all images to the same resolution of 224 by 224 pixels.

To mitigate issues of poor staining, poor contrast between cells, poor image quality, and differences in slide colour, we used masking to enhance our cropped images and normalized them across each stain colour. This masking helped the model focus specifically on lymphocyte shape and relative density. It involved first creating upper and lower bound arrays for each pink, purple, and blue stain colours that could be applied to the images as masks. This took the form of an upper and lower numpy array for the pink, blue, and purple stains of (230, 170, 255) and (130, 70, 215), (180, 200, 160) and (90, 120, 100), and (200, 150, 200) and (100, 50, 100), respectively, in the LAB colour space. By then using a series of the cvt2.Color(), cvt2.InRange(), bitwise_or(), and bitwise_and() functions to combine and apply all three masks to the same image, we achieved a much higher contrast and more regularized image that our CNN could better classify (Mandala, 2023).

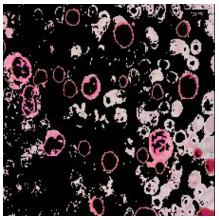
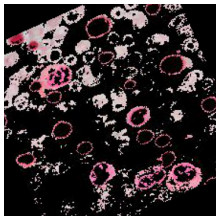
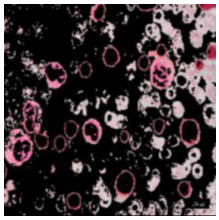
3.2.2 Augmenting

There is a need for a high degree of accuracy for a model that predicts cancer due to the fatality and rapidity of the disease's progression. As a result, we needed to train our model on a much larger dataset than the one we possessed. Since ALL blood smear datasets are so few and far in between, we increased the size of our dataset synthetically using pyTorch transformations, such as vertical flipping, horizontal flipping, randomized rotation, and the addition of Gaussian blur. As a result, we increased our dataset by a factor of five, turning a 3200 image dataset into one containing over 16000 images. We also better trained our model to accommodate imperfections and noise.

3.3 Splitting Between Training, Testing, and Validation

For splitting the dataset, we used a 60/20/20 split between training, validation and testing sets, as this was the optimal training/testing split for a classification problem (Dobbin & Simon, 2011). Specifically there were 1500 Benign, 1500 Early, 1500 Pre-B, and 1500 Pro-B blood stains used to train the model. This ensured that both the testing and training stages of the model experienced the same distribution of the stages of Leukaemia, establishing the model's ability to distinguish all

Table 1: The four main augmentations applied to images

Vertical Flip	Horizontal Flip	Randomized Rotation	Gaussian Blur
			

four of the classes with equal proficiency. The fact that the testing dataset was set aside and never used throughout training meant that the test accuracy achieved by the model would be an accurate representation of the model's performance in a realistic setting with never before seen images.

Table 2: Augmented Images Train/Validation/Test Split

Set	Benign Images	Early Images	Pre-B Images	Pro-B Images
Training Set	1500	1500	1500	1500
Validation Set	500	500	500	500
Testing Set	500	500	500	500

4 Baseline Model: Simple Convolutional Neural Network

Classifying the stage of leukemia into benign, early, pre, pro based on blood smear images lends itself well to using a convolutional neural network (CNN) which detects important features of images without supervision. The CNN takes blood smear images as inputs assigning weights to different aspects of an image, so they are able to differentiate one class from another. A CNN captures spatial and temporal dependencies which is important as the biomedical images like leukemic stem cells are not restricted to a specific region and are difficult to analyze Sarvamangala & Kulkarni (2011). Our primary model is a pretrained CNN with transfer learning, and this makes a simple CNN a good baseline for our project.

4.1 CNN Architecture

Our simple convolutional neural network has 2 convolutional layers followed by rectified linear activation function and uses max pooling to reduce dimensionality and for feature learning, as seen in Figure 2. The CNN has 2 convolutional layers, the first layer convolutes takes a 3x224x224 image with input of 3 channels and a kernel size of 5, outputting a 5 channel. A max pooling layer of stride 2 and size 2 is applied to reduce the dimensions in half from 5x220x220 to 5x110x110. An additional convolutional layer produces an output of 10 channels with dimensions 106x106. The

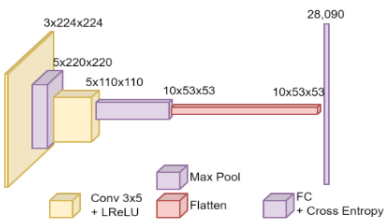


Figure 2: Architecture of the baseline CNN showing the feature extraction and classification stage

convolutional operation is again followed by max pooling layer to reduce to 10x53x53. 2 fully connected layers are used to flatten this input into a single vector and give final probabilities for each of the four labels.

4.2 Training & Testing

To train a cross entropy loss function and stochastic gradient descent optimizer with a learning rate of 0.01 and momentum of 0.9 was used. Using the pre-processed images, our baseline model has an accuracy of 83.6%.

5 Primary Model Architecture

Our final model consists of two main components: a convolutional feature extractor, and a deep-ANN fully-connected classifier. For the former component, we employed transfer learning to generate embeddings of each image in our dataset, using pre-trained models. The latter component was designed and constructed by our team. In this section, we will elaborate on the design decisions of both components, as well as the steps taken to perform end-to-end training and optimization.

5.1 Feature Extraction CNNs

Our team compared the performance of three pre-trained CNN models to perform feature extraction. We treated these models as hyperparameters, ultimately selecting the model which gave us the greatest validation accuracy. The three feature extractors we compared were: AlexNet, VGG, and ResNet18. The results from the first round of validation are listed in Table 3 which justifies our decision to select ResNet18. The training procedure will be described in more detail in Section 5.3

Model	No. Conv. Layers	Embedding Dimension	Validation Accuracy
AlexNet	8	4096	83%
VGG	16	8192	72%
ResNet18	18	512	87%

Table 3: Preliminary validation results used to justify the selection of ResNet18 as a feature-extractor.

Out of the above models, we noted that ResNet18 performed the best for us. ResNet18 is a very deep model which employs skip connections to help combat the diminishing gradient problem. As we have collected substantial amounts of data, we believe that a deep architecture is best suited for the task at hand. It is evident, from our experiments, that ResNet18 is sufficiently deep, with the propensity to perform well, even before fine-tuning its weights.

To implement this model, we first downloaded the pre-trained architecture from Ramzan et al. (2019). Once imported into PyTorch, we enumerated the layers of the model and constructed a new class consisting of all layers, except the fully-connected head. We instead replaced this with our own classifier, described in Section 5.2. A detailed diagram of the ResNet18 architecture can be found in Figure 3, below.

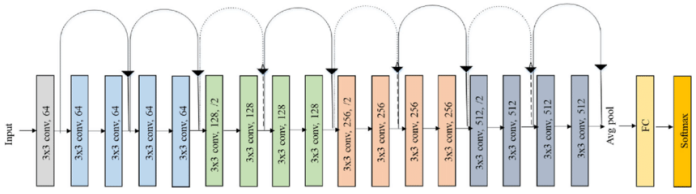


Figure 3: The ResNet18 architecture. We removed the FC and SoftMax layers, instead returning embedding vectors with dimension 512. These were then fed to our custom classifier.

To improve the embedding quality, the model was fine-tuned using methods described in Section 5.3.

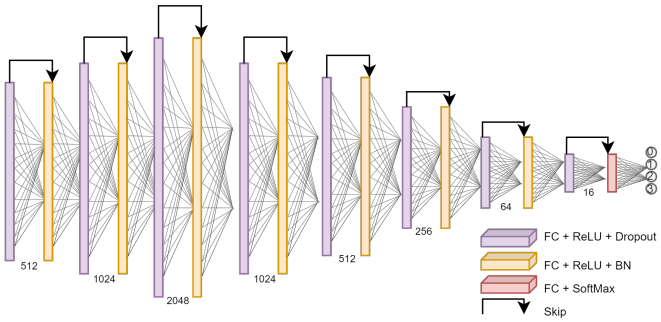


Figure 4: The classifier portion of our network. The input of the classifier is a 512-dimensional embedding vector produced by ResNet18. The output layer classifies the input into one of four classes.

5.2 Deep-ANN Classifier

With the feature extractor in place, our team implemented a classification head to map the 512-dimensional embedding space, produced by ResNet18, to a 4-dimensional vector for our prediction probabilities. Each one of these dimensions represents one of the following classes: *benign, early-stage leukemia, pre-stage leukemia, or pro-stage leukemia*.

Like our feature extractor, we opted for a deep residual classifier architecture. Unlike the one employed in our simple baseline, it is 16-layers deep and employs several advanced regularization methods. For example, after every second layer, we use a dropout layer, followed by a batch normalization.

In order to avoid vanishing gradients, we also use skip connections across linear layers of equal size. The nature of our residual network allows gradients to flow end-to-end, with little attenuation. This property is exploited during the end-to-end fine-tuning steps of the network, when we must update the CNN and ANN weights. CNN weights near input may be up to 40 layers away from the output layer – as a result, we rely on these skip connections for the un-impeded flow of information to guarantee effective backpropagation takes place (even at very low learning rates).

5.3 Training and Optimization Procedure

5.3.1 Loss Function

As this was a classification problem, we used the negative-log-likelihood (NLL) loss function between our computed output probabilities and the four ground-truth class labels. In addition to this, we also added a second term to our loss function. We grouped the probabilities of the cancerous classes together and constructed a binary cross-entropy (BCE) loss. The significance of this added term was to emphasize the importance of recognizing cancerous versus non-cancerous cases in our training. Ultimately, a misclassification of the sub-type of leukemia will not have as great of an effect on a patient's life as would classifying cancerous cases as non-cancerous.

As such, our total loss function was a linear combination of both terms:

$$L(x, y) = \lambda_1 \text{NLL}(\hat{x}, y) + \lambda_2 \text{BCE}(\hat{x}', y') \tag{1}$$

where $\hat{x}$ is the model prediction, y is the ground-truth label, $\hat{x}'$ is the model prediction which is post-processed into cancerous/non-cancerous superclasses, and y' is the superclass ground-truth label. λ_1 and λ_2 are regularization hyperparameters, which were set to 1 after a grid-search.

5.3.2 Training Procedure

The training of our model involved two main steps: hyperparameter tuning and fine-tuning. As aforementioned, we treated the three different feature extraction models as hyperparameters. To select the best model, we created embeddings for each input image using each model and used

these to train the classifier part only. Throughout the training, we used the Adam optimizer with a scheduled learning rate, exponentially decaying by a factor of 0.98 after each iteration. We also used mini-batches with a batch size of 256 (the most that CUDA could hold for us). We started with a learning rate of 0.1, which decayed to 3e-5 after 50 epochs.

After selecting the ResNet18 model as our feature-extractor, we unfroze the ResNet18 weights and proceeded to fine-tune the entire end-to-end on a small batch size of 32. We used a decaying learning rate, starting at 3e-4. We employed early-stopping with patience to return the best model, once our validation accuracy plateaued for at least 20 iterations.

6 Evaluating on New Data

To evaluate the performance of our model and its ability to generalize, we removed 20& of data from our datasets before the training of any models took place. This meant that, once we were confident that we had constructed a plausible model, we could try it on never-before-seen data to give us an unbiased estimate of the generalization power. We constructed our testing dataset by extracting 100 images from each of the four leukemia/benign classes. This gave us a balanced dataset of 400 images, which we used to generate the results described below. In total, these 400 images represented about 20% of our data. We also used 60% for training and another 20% for validation.

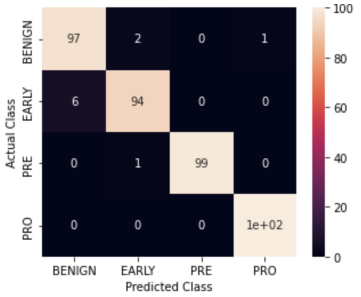
6.1 Quantitative Results

In the final iteration of our model, we tuned hyperparameters and added data augmentations to our pre-processing and obtained a validation set accuracy of 99% and thereby stopped further modifying our model.

When evaluating the primary model on the testing dataset, we observed an accuracy of 97%. This indicated that our model was successful at generalizing. To ensure the effectiveness of our model, we computed the precision, recall, and F1-Score values of our test results as seen in the confusion matrix, which are seen in figure 5.

The recall value of our model is one of the metrics that our team especially focused throughout the project timeline, considering a false-negative value could jeopardize the life of an actual ALL patient, having a high recall value is particularly important. Our model proves successful in that regard with a 98% recall value.

The primary model outperforms the simple baseline model significantly even when it performs on a never-before-seen blood smear images, thereby showing an effective leukemia classifier.



Class	Precision	Recall	F1-Score
Benign	94%	97%	96%
Early	97%	94%	95%
Pre	100%	99%	99%
Pro	99%	100%	100%
Average	**98%**	**98%**	**98%**

Figure 5: The Confusion Matrix and the table of Precision, Recall, and F1-Score values of our primary model, tested with never-before-seen testing dataset.

6.2 Qualitative Results

In order to better visualize the results of our feature extractor, a t-SNE Projection of the embeddings of the images in our dataset is shown in Figure 6. It is apparent that each class has their own region in the 2D plane, which explains how our model achieved high accuracies. There were some mix-ups

between the benign and early classes, this indicates that there were some similar features between these two classes that our model was not able to identify successfully. This is also apparent in the result of the confusion matrix as these 2 classes have the most misclassifications. Our model did the worst in classifying input belonging to these two classes.

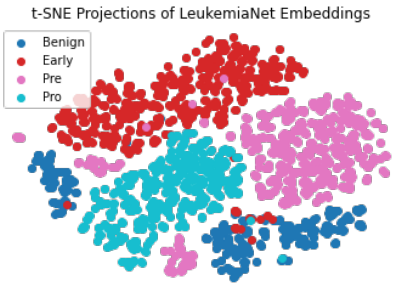
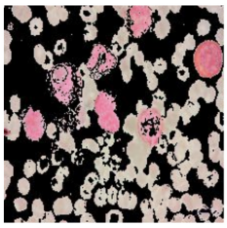


Figure 6: The t-SNE projection of the embeddings we extracted from our whole dataset.

Figure 7 below demonstrates how our model works depicting a sample input fed into our model and its corresponding output. In the output array, we see that the probability that the image belongs to the Benign Class is the highest, and that the probability of cancer is 0.0078. With this, our model successfully classifies the input image as benign.



```
Model Output: [9.9217564e-01 2.8899810e-03 4.0723130e-04 4.5271106e-03]
Probability of cancer: 0.0078
Prediced benign
Probability of early leukemia: 0.0029
Probability of pre leukemia: 0.0004
Probability of pro leukemia: 0.0045
```

Figure 7: Sample input fed into our model and the output our model produced.

7 Discussion

Where classification of ALL is concerned, typically the most difficult case to detect is early-stage leukemia. Oftentimes, early-stage cells closely resemble their benign counterparts – this is why the majority of leukemia cases remain undetected until they have progressed. Using our model, we have shown that early leukemia has a recall rate of 94% – the lowest out of all classes. While we would ideally want to see a higher recall score, this quantity is much higher than the recall of our baseline model and that typically expected for early leukemia.

In the t-SNE projection, we also see a high degree of overlap between some of the benign cases and the early cases. This further drives the conclusion that visually, early-leukemia and benign cases appear very similar. As a whole, however, our model provides very clear boundaries for the majority of data points in the projection. We believe that with more data, we could train the feature-extractor to perform embeddings which exploit the orthogonality between classes to a higher degree.

What is notable about the model we constructed, is its propensity for correct classifications. One study which used screening by medical professionals demonstrated misclassifications at a rate of 30 - 40%, which varied by specialist (Nagiub Abdelsalam et al., 2018). In contrast, our model gives significantly higher recall values, at 98%. Looking at the blood smear images with the naked eye, misclassification without computer aids is understandable as the differences between the classes cannot be easily inferred without a high degree of expertise and experience, showing that a medical image diagnostic problem like leukemia classification lends itself well to a deep learning approach.

8 Ethical Considerations

Issues of representation bias, aggregation bias, and authority must be considered to understand the limitations and the application of our model. These ethical considerations also provide insight into how to improve our model and the use of deep learning in the medical sector.

Representation biases and a lack of diversity in medical data can lead to statistical bias in diagnosis, particularly in deep learning models as they rely heavily on data for predictions(DeCamp & Lindvall, 2020). Patients with higher socioeconomic statuses are more likely to have access to medical care and they likely contribute more to blood smear datasets. Consequently, our dataset likely under-represents impoverished and remote populations. As a result the model might be skewed towards better generalizing towards populations with higher socioeconomic standing. There is an added level of measurement bias to consider since much of the collection, labeling, and representations of peripheral blood smear images is based on human biases (Parikh et al., 2019). Finally, our dataset is subject to aggregation bias, since the relative amounts of benign, early, and malignant blood smear images does not accurately represent the relative likelihood of ALL in a general population. Overall, the lifetime risk of ALL is 1 in 1000, yet the dataset, which was collected specifically for ALL, exhibits a relatively even distribution of early, benign, and malignant ALL (Wessel, 2023). This may cause the model to mistakenly overpredict ALL. Due to the rarity and complexity of diseases, there is a lack of data that can thoroughly capture the intricacies of every patient's case (Shuaib et al., 2020).

Special care must be taken to prevent "false negatives" considering the life-altering consequences of such an event. A doctor might inadvertently allow their patient's disease to quickly progress and claim their life by accepting a "false negative" classification of the model. (Clinton et al., 2020). Thus, special care must be taken to make sure that the model errs on the side of caution when deciding whether lymphocytes are malignant, even if unable to predict which type of malignancy is depicted. Furthermore, meticulous testing and lawmaking must be applied before the model can be deployed in the field. This legal framework must outline the level of accountability and accuracy that this model must have before being applied in the field.

Further building on this, the model's nature as an algorithm means that it cannot take liability for its mistakes. There is no existing legal framework that can reasonably assign authority or legal responsibility to an AI model (Lupton, 2018). Any ethical liability of the model falls on those that designed it or the physicians and hospital administrators that choose to employ the model. At best, ethically, the model could operate similar to a consulting physician, where the patient's human physician would be required to use their own expertise to approve or disregard the model's output. As such, the ALL model can not independently be reliably used as a method of diagnosis.

9 Project Difficulty and Quality

We believe that our project demonstrates that we went to great lengths to construct a model which performs exceptionally well on new data. Our model employs many of the methods we have explored in class, such as transfer learning, fine-tuning, regularization techniques, hyperparameter tuning, and very deep frameworks, to provide a robust detection model for leukemia.

It is critical to note that the scope of our project goes well beyond model training and optimization. Our team struggled to find quality datasets for this task and employed various techniques to increase the quantity and fidelity of data fed into the network as datasets in the biomedical field are scarce. Overall, this all paid off with an exceptionally-high 97% testing accuracy score. As a team, we enjoyed this project immensely, and are greatly satisfied with our achievements.

10 Link to Github

Preliminary Stage: https://github.com/mikeacquaviva/APS360-Leukaemia-Classification.git

Final Model: https://github.com/mikeacquaviva/APS360-LeukemiaNet

REFERENCES

Wahidul Hasan Abir, Md. Fahim Uddin, Faria Rahman Khanam, Tahia Tazin, Mohammad Monirujjaman Khan, Mehedi Masud, and Sultan Aljahdali. Explainable ai in diagnosing and anticipating leukemia using transfer learning method. *Computational Intelligence and Neuroscience*, 2022: 1–14, 2022. doi: 10.1155/2022/5140148.

Mehrad Aria, Mustafa Ghaderzadeh, Davood Bashash, Hassan Abolghasemi, Farkhondeh Asadi, and Azamossadat Hosseini. Acute lymphoblastic leukemia (all) image dataset, 2021. URL https://www.kaggle.com/dsv/2175623.

Jasmine Begum and Abdul Razak. Leukocytes classification and segmentation in microscopic blood smear: A resource-aware healthcare service in smart cities. *International Journal of Advanced Research in Computer Science*, 5(3), 2017.

Laurence P Clinton, Karen Somes, Yong Sik Chu, and Faiz Javed. Acute lymphoblastic leukemia detection using depthwise separable convolutional neural networks. 2020.

Matthew DeCamp and Charlotta Lindvall. Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association*, 27(12):2020–2023, 06 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa094. URL https://doi.org/10.1093/jamia/ocaa094.

Kevin K Dobbin and Richard M Simon. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, 4(1), 2011. doi: 10.1186/1755-8794-4-31.

Amin ul Haq, Jian P. Li, Shakir Khan, Alshara Ali Mohammed, Reemiah Muneer Alotaibi, and Cobbinah Bernard Mawuli. Dacbt: deep learning approach for classification of brain tumors using mri data in iot healthcare environment. *Scientific Reports*, 2022. doi: 10.1038/s41598-022-19465-1.

Larissa Hirsch. Acute lymphoblastic leukemia (all). 10 2021. URL https://kidshealth.org/en/parents/all.html.

Chun Ho Ngai Veeleah Lok Lin Zhang Don Eliseo Lucero-Prisno III Wanghong Xu Zhi-Jie Zheng Edmar Elcarte Mellissa Withers Martin C. S. Wong Junjie Huang, Sze Chai Chan. Disease burden, risk factors, and trends of leukaemia: A global analysis. 12, 7 2022. URL https://www.frontiersin.org/articles/10.3389/fonc.2022.904292/full.

Yogesh Kumar, Apeksha Koul, Ruchi Singla, and Muhammad Fazal Ijaz. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Nature Public Health Emergency Collection*, 18, Jan 2022. doi: 10.1007/s12652-021-03612-z.

Michael Lupton. Some ethical and legal consequences of the application of artificial intelligence in the field of medicine. *Trends in Medicine*, 18(4), 2018. doi: 10.15761/tim.1000147.

Sujith K Mandala. Pre-processing images for image classification, Feb 2023. URL https://www.kaggle.com/code/sujithmandala/pre-processing-images-for-image-classification.

Maurie Markman. Risk factors for leukemia. 05 2022. URL https://www.cancercenter.com/cancer-types/leukemia/risk-factors.

Hala Zayed Mohamed Loey, Mukdad Naman. Deep transfer learning in diagnosing leukemia in blood cells. 9, 4 2020. URL https://www.mdpi.com/2073-431X/9/2/29.

Eman M. Nagiub Abdelsalam, Khaled F. Hussain, Nagwa M. Omar, and Qamar Taher Ali. Computer aided leukemia detection using microscopic blood image based machine learning "convolutional neural network". *Clinical Lymphoma Myeloma and Leukemia*, 18, 2018. doi: 10.1016/j.clml.2018.07.246.

Ravi B. Parikh, Stephanie Teeple, and Amol S. Navathe. Addressing Bias in Artificial Intelligence in Health Care. *JAMA*, 322(24):2377–2378, 12 2019. ISSN 0098-7484. doi: 10.1001/jama.2019.18058. URL https://doi.org/10.1001/jama.2019.18058.

Farheen Ramzan, Muhammad Usman Khan, Asim Rehmat, Sajid Iqbal, Tanzila Saba, Amjad Rehman, and Zahid Mehmood. A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks. *Journal of Medical Systems*, 44, 12 2019. doi: 10.1007/s10916-019-1475-2.

D.R Sarvamangala and Raghavendr V Kulkarni. Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*, 4, 2011. doi: 10.1186/1755-8794-4-31. URL https://link.springer.com/article/10.1007/s12065-020-00540-3#citeas.

Abdullah Shuaib, Husain Arian, and Ali Shuaib. The increasing role of artificial intelligence in health care: Will robots replace doctors in the future? *International Journal of General Medicine*, 13:891–896, 2020. doi: 10.2147/IJGM.S268093. URL https://www.tandfonline.com/doi/abs/10.2147/IJGM.S268093.

Kenneth P Smith, Anthony D Kang, and James E Kirby. Automated interpretation of blood culture gram stains by use of a deep convolutional neural network. *Journal of Clinical Microbiology*, 44, February 2018. doi: 10.1128/JCM.01779-17.

Raj Talluri. Conventional computer vision coupled with deep learning makes ai better. 11 2017. URL https://www.networkworld.com/article/3239146/conventional-computer-vision-coupled-with-deep-learning-makes-ai-better.html.

Muhammed Umer, Saima Sadiq, Muhammad Ahmad, Saleem Ullah, Gyu Sang Choi, and Arif Mehmood. A novel stacked cnn for malarial parasite detection in thin blood smear images. *Journal of Clinical Microbiology*, 44, February 2018. doi: 10.1128/JCM.01779-17.

Megan Wessel. Key statistics for acute lymphocytic leukemia (all), 2023. URL https://www.cancer.org/cancer/acute-lymphocytic-leukemia/about/key-statistics.html.