HIGH SCHOOL EDITION
Journal of Student Research

A Comparative Study of Transfer Learning Networks and Siamese Networks for Acute Lymphoblastic Leukemia (ALL) Diagnosis

Bhrugu Bhatt[1], Arjun Iyer[1], Duke Writer[#] and Geoffrey Schau[#]

[1]Academies of Loudoun, Leesburg, VA, USA
[#]Advisor

ABSTRACT

Acute Lymphoblastic Leukemia (ALL) is a type of blood cancer that primarily affects the white blood cells. ALL is a serious problem in our society, accounting for approximately 25% of all pediatric cancers. Another major issue of ALL is that the healthy and cancerous cells look extremely similar to the human eye, leading to high chance of misdiagnosis. A novel method was proposed to solve a two-fold problem using Few Shot Learning (FSL), a machine technique that uses a small support set to perform binary classification. Traditional machine learning methods require extremely large medical training datasets, which can lead to high computation expenses and are difficult to access due to patient privacy. Convolutional neural networks (CNNs) such as VGG19 and GoogLeNet were also compared to investigate if there was any significant improvement with FSL on the CNMC dataset. Currently, our Siamese Network has a 85% testing accuracy when training it on 10 Epochs. However, transfer learning has also shown to be a way to use a small amount of data to receive a high accuracy rate, and require a less computationally intensive training process.An accuracy rate of 99% and 96% was achieved on GoogLeNet and on VGG-19, which were trained for 50 epochs with image transformation preprocessing. Overall, traditional CNNs continue to outperform FSL methods, but methods are still a viable option for second opinions.

Introduction

Since the invention of the first computer, the goal has always been to put computers at par with human intelligence and learning efficiency with computational ease. In the past few decades, the incredible development in the field of computer science has brought about the ideas of machine learning (ML) and artificial intelligence, which allow computers to imitate some of the learning processes of humans. Machine learning models have the ability to take extremely large amounts of data and find trends for various tasks such as image segmentation, image classification, regression-based predictions, and finding correlations between the hundreds of parameters (Wang et al., 2020). In recent years, the applications of machine learning have expanded to medical sciences in which it has helped medical professionals with various diagnoses based on various scans and images of cells. These diagnosis methods allow for greater accuracy and a more robust analysis of medical reports. Due to these many factors, numerous machine learning studies have been specifically conducted to improve the medical diagnosis of Acute Lymphoblastic Leukemia (ALL).

Using machine learning applications in ALL has been of interest as it is extremely difficult to distinguish between leukemic lymphoblasts from normal healthy cells because under a microscope; they morphologically look very similar (Shafique & Tehsin, 2018). This greatly increases the chance of misdiagnosis and human error, which could be fatal to the patient. ALL is caused by genetic mutations that tell the cells to continue growing and dividing instead of following the standard process. It causes rapid creation of immature white blood cells (lymphoblasts) inside the bone marrow instead of mature white blood cells (lymphocytes). There are two types of acute lymphoblastic leukemia: B-lymphoblastic leukemia and T-lymphoblastic leukemia.

Traditional approaches to solving this problem include looking at bone marrow and blood samples under a microscope to try and classify using defined lymphoblast characteristics in consideration with patient history and lab tests. The defining characteristics of a lymphoblast include insufficient and sometimes tennis racket shaped cytoplasm, always absent auer rods, rarely present granulation, always absent nucleoli, dispersed condensation chromatin in large lymphoblasts and very condensed in small lymphoblasts, and central and mainly round nucleus with occasional indentations. So far, the French-American-British (FAB) classification method has been used to differentiate between the three subtypes of lymphoblasts: L1, L2, L3. Classification by eye is extremely prone to human error due to the simple number of factors that affect a person's perception, decision making, and thought process.

A comparative study was conducted between the diagnosis accuracies of Siamese Networks for FSL and GoogLeNet and VGG-19 networks for traditional transfer learning. The aim was to identify the effectiveness of a FSL approach compared to the traditional methods and create a proof of concept of if it is viable for future research and should be pursued for ALL and other cancer classification.

Present Research

Over the years, numerous machine learning studies have attempted to solve this image classification problem using large ALL cell image datasets such as ALL-IDB, ASH image bank, and the ISBI 2019 - CNMC challenge dataset. The CNMC dataset is the largest available single cell ALL image dataset. This is extremely suitable for machine learning purposes because the larger the dataset, the more variation in patient cases can be taken into account for diagnosis. Majority of these studies have taken machine learning concepts a step further and applied them in the newer field of deep learning (DL). Deep learning uses many of the machine learning algorithms like support vector machine (SVM) algorithm, naive bayes algorithm, k-nearest neighbors (KNN), random forest algorithm and applies them to the concept of deep neural networks (DNN) (Liu et al., 2017). Deep neural networks form the basis for computers to "learn" like humans. Of these DNNs, convolutional neural networks (CNN) have grown to become the most widely used neural network due to the increased automatic detection accuracy without human supervision. It is also extremely efficient for dimensionality reduction for complex multidimensional images. This helps increase computational efficiency. The number of parameters needed in the model is also reduced, which reduces the number of computations (Liu et al., 2017).

In the case of the CNMC dataset, there are mismatching amounts of data in each class. To solve for this, image augmentation techniques can be used to increase the number of images in one of the classes. Liu et al. employed a bagging ensemble method to overcome the problem of ALL cancerous images being double the number of images for the healthy cell image class. The healthy cell images class would have to get augmented with various techniques to match the number. Bagging ensembles can be used to increase variance in the data for a more robust data analysis, along with increasing the images in one of the classes (Liu & Long, 2019). J. E. Mauricio de Oliveira and D. O. Dantas (2021) proposed to augment the data to increase the image amount in the healthy cell class by reflecting vertically and horizontally, 60 degree rotations, 17 × 17 pixel Gaussian blur, salt and pepper noise, and shear factor of 0.3 (Oliveira & Dantas, 2021); however, Kasani et al. (2020) augmented data with brightness fluctuations, intensity, and flips along with resizing to 380 × 380 pixels. The resizing of images is to crop unnecessary black bordering around the cells of the original 450 × 450 pixel images.

There are numerous types of CNNs that were initially created based on their performance in the image classification task on the largest image database, ImageNet. The most popular types of CNNs used are ResNet, AlexNet, and GoogLeNet (Liu et al., 2017). J. E. Mauricio de Oliveira and D. O. Dantas (2021) attempted to use more traditional neural networks such as Xception and VGG16 instead of very complex architectures to achieve a high accuracy. Kasani et al. (2020) used multiple ensemble model combinations to find the most accurate prediction model. Based on prediction performance, the NASNetLarge model was used for aggregated ensemble architecture which achieved accuracy of 96.58%. Honnalgere and Nayak (2019) aimed to decrease output time by using shared weights and using a pretrained CNN. They used transfer learning to adapt a VGG16 model with batch normalization, pretrained on the ImageNet dataset. Weights are the biases or conditions applied to the input variable to get the output variable. The weights change as the model trains to improve the accuracy.

In VGGNet models, the output layer is made of 1000 neurons using the softmax function. Oliveira and Dantas (2021) replaced the neurons global average pooling layer followed by two fully connected layers with 512 neurons using rectified linear units (ReLU), then linked to a prediction layer with two neurons using softmax function. A dropout layer was also added to prevent overfitting on the data. Kasani et al. (2020) used five layers extracted from the MobileNet and average pooling layers were applied to each of them. They achieved an overall accuracy of 96.17%. Prellburg and Kramer's (2019) model had five convolutional stages with spatial downsampling by a factor of two between stages, followed by global pooling and linear classifier and were able to achieve an overall accuracy of 89.91%. Honnalgere and Nayak (2019) used a combined model architecture that was employed for training for each stage of training with two Inception ResNets pre-trained on ImageNet. The outputs of each model were then combined and used for classification with two neurons.

All of these studies employed a learning model on the full extent of the dataset of thousands of images. In reality, these models would not be feasible in terms of real applications in medical facilities due to the sheer massive computational power necessary and mainly because most medical conditions do not have a massive image database for machine learning. Extremely large and labeled medical image datasets are hard to collect and annotate due to the rarity of certain cases and patient privacy laws; therefore, it would be more feasible to have a machine learning approach that can use a smaller image sample to make an equivalently valid diagnosis.

Recently a study was published using few-shot learning (FSL), which was able to classify cervical cancer cells into four classes. Yarlagadda et al. (2019) used the Inception-v3 and ResNet models with R-MAC global descriptors representing CNN layer features. The Inception-ResNet model utilizes residual connections and cheaper inception blocks to make it more accurate while still being efficient. The R-MAC global descriptors are used to create an image representation of the CNN features using spatial regions. The CNN was trained on the dataset for 940 epochs. They used nearest neighbors to find the closest similarities. With this method, they achieved an accuracy of 94.6% (Yarlagadda et al., 2019).

Few shot learning paves the way for pushing research forward. It is a newer machine learning approach that is intended to closely follow human learning by learning how to distinguish between similar and different images and then using that knowledge to classify queries with a small support sample. Few-shot learning can train the parent model on a well known large dataset (e.g., ImageNet) and then use a way smaller dataset for support samples in the child model which can decrease computational time and cost (Yarlagadda et al., 2019). The weights or biases in the parent model are transferred to the child model to apply to the support dataset. Weights and biases are what cause a machine learning model to evolve overtime as it learns. It uses this learned knowledge to correctly classify the query image. It decreases the amount of incorrectly classified rare cases because it doesn't use the same dataset for training and testing. This greatly improves the effectiveness of the child model. It allows usage of a small support sample set of images for rare cases (Wang et al. 2020).

FSL problems are described with a N-way K-shot notation in which the way is the number of classes a query can be classified into, and the shot is the number of each type of image in each class (Yarlagadda et al., 2019). For example, in the ALL problem, it would be two-way K-shot as the query is being classified into either ALL or Healthy and the number of shots are subject to change depending on the model efficiency. Some of the applications of FSL include few-shot classification which predicts the label for an input, few-shot regression which estimates a trend based on a few given input-output pairs, and few-shot reinforcement learning which uses a reward system as incentive to get the model to move closer to the correct end result with every epoch on the training set (Wang et al., 2020).

The most relevant research in the FSL field is the concept of weakly supervised learning that the model learns from its pre training experience with a very limited amount of samples with labeled information (Yarlagadda et al., 2019). The most suitable type of weakly supervised learning is transfer learning which gradually learns generic information across a large pre-training dataset. Transfer-learning results can be used as the prior knowledge necessary for the model in FSL (Wang et al., 2020). The empirical risk is minimized when meta-learning is used as prior knowledge is being used. Empirical risk is the expected risk based on the average loss during model iteration (Wang et al., 2020).

While past research in ALL image classification has been able to yield great results, their model accuracy depends on having an extremely large dataset, which may not be feasible in real life as discussed previously. The past research mostly uses a combination of supervised learning with some sort of CNN on the huge datasets and then fine tuning it or using an ensemble technique to make it more accurate. If the classes look extremely similar, like in this case, simple supervised learning models have a high likelihood of incorrectly classifying rare or edge case scenarios (Wang et al., 2020). This can lead to frequent misdiagnosis, which has direct negative implications on patient health. Therefore, I would like to take the first few-shot learning model approach on the ALL image classification problem with the CNMC dataset. While the CNMC dataset is one of the largest available datasets, it is still very useful for FSL, as a small sample of images can be pulled from it. These images have already been preprocessed, which makes them optimal for machine learning purposes. This approach has previously never been taken for classification of ALL cells from healthy cells. In general, applying a FSL approach for ALL classification is very useful, despite having large datasets available, because if the model successfully completes the difficult task of classifying very similar looking images, it can be expanded with slight modification for other medical conditions.

Methods

The dataset that was used in this study was produced by the Children's National Medical Center for the ISBI 2019 ALL Challenge Event(Gupta, Gupta, 2019). The dataset contains approximately 10,000 images of both lymphocytes and lymphoblasts from children affected with ALL.

Siamese Networks are specifically designed for Few Shot Learning as they only require a small support set of images. Siamese Networks have a special preprocessing step that requires the creation of a support set of images in the form of positive and negative image pairs. Positive image pairs are if the support images are of the same class. Negative image pairs are if the support images have different classes. These image pairs are then fed into the Siamese Network for training. They have two identical Convolutional Neural Networks (CNN) that train and update weights together. Each CNN then extracts the features from one image from the pair and converts them to image encoding that are scalar quantities. The Euclidean distance between these two scalar quantities can be calculated to create a similarity score. If the similarity score is greater than 0.5, it can be classified as the same class as the support image and if it is below 0.5, it can be classified as the opposite class of the support image. Lastly, the hyperparameters were fine tuned to maximize the test accuracy with minimal overfitting.

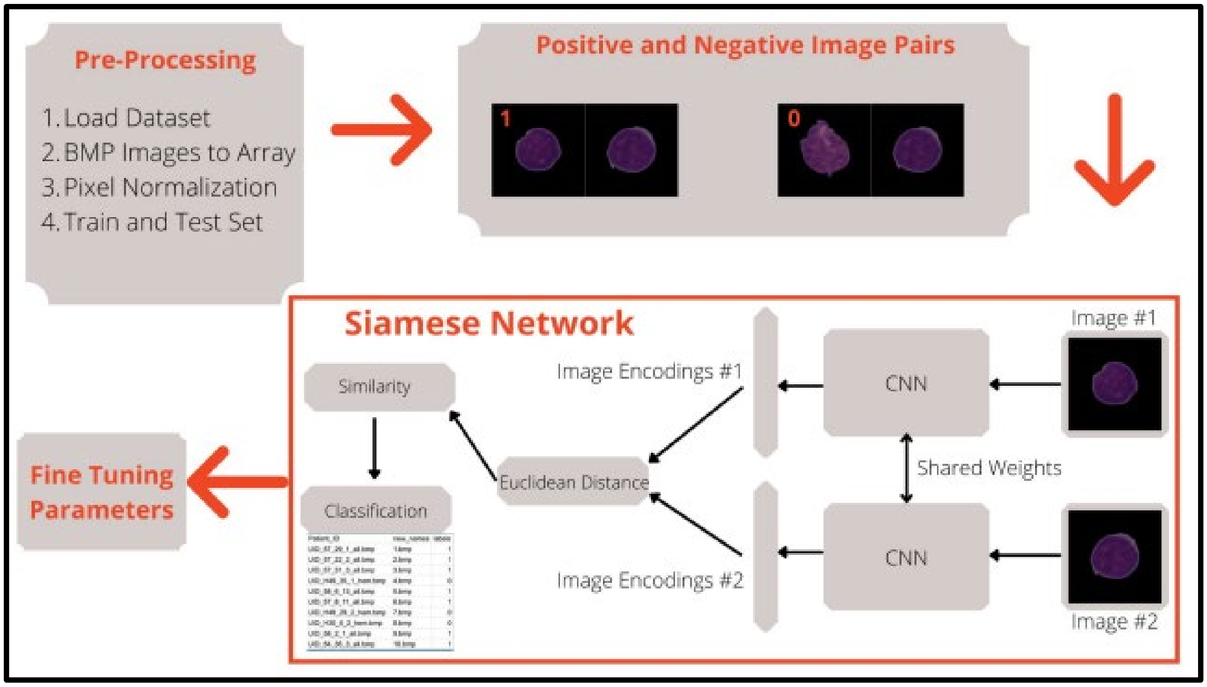

Figure 1. Siamese Network Implementation

GoogLeNet was proposed as a high performance solution to Deep Neural Network problems such as overfitting caused by additional layers. The introduction of GoogLeNet's Inception module uses feature detection at different scales through convolutions with different filters and reduces the computational cost of training an extensive network through dimensional reduction. In contrast to Siamese Networks, the preprocessing for GoogLeNet has different steps, such that the cell images go through transformations such as random rotations, resizing/crops, flips, and normalized coloring. Each of these images are passed into the architecture for training through a mix of 22 convolution, pooling, filter, and dropout layers. The last of these layers is a softmax function that predicts the probability of an input being in the ALL or Healthy class. The output is then compared and validated against the given values to calculate a validation accuracy. It then moves onto a testing phase which compares the prediction to a completely new set of inputs that the model has never seen before. Lastly, the hyperparameters were fine tuned to maximize the test accuracy with minimal overfitting.

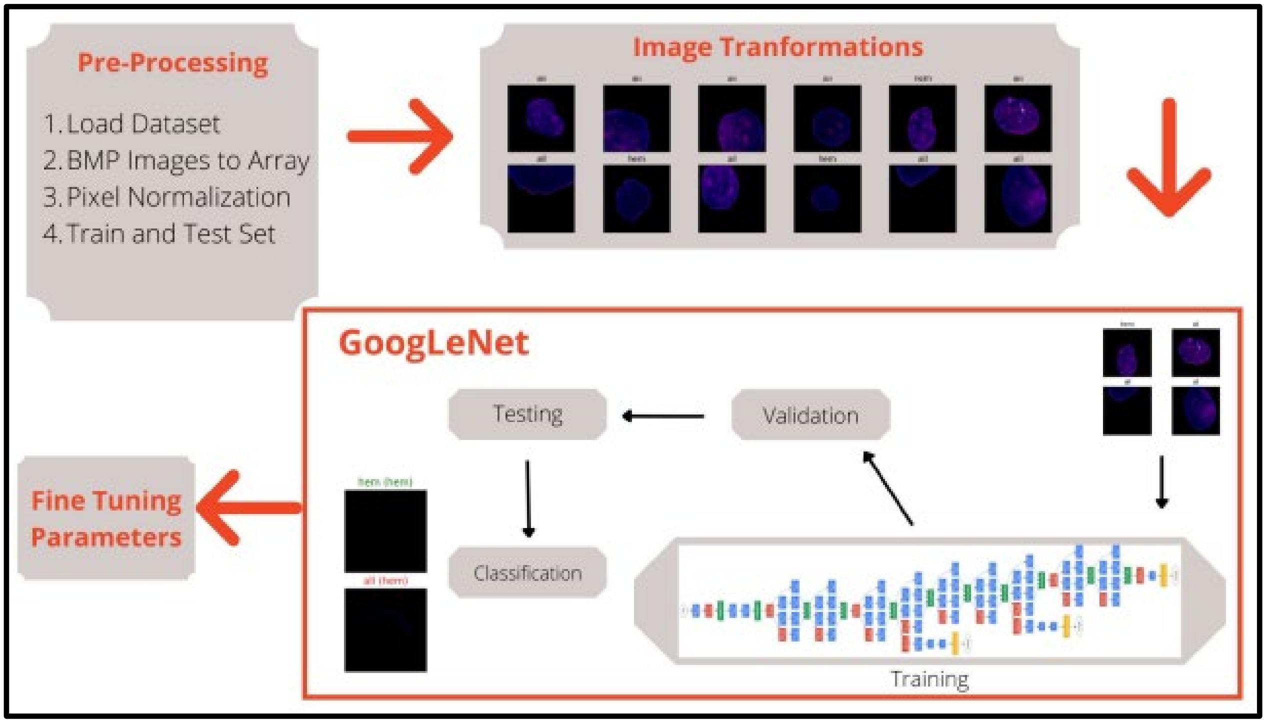

Figure 2. GoogLeNet Implementation

The VGG-19 model was designed to reduce the number of parameters in the convolution layers and decrease train time for image classification. The model has the same preprocessing and image transformation steps as the GoogLeNet model with the transformations such as random rotations, resizing/crops, flips, and normalized coloring. The model architecture includes a mix of 19 convolution and pooling layers of various dimensions. The last of these layers is a softmax function that predicts the probability of an input being in the ALL or Healthy class. The output is then compared and validated against the given values to calculate a validation accuracy. It then moves onto a testing phase which compares the prediction to a completely new set of inputs that the model has never seen before. Lastly, the hyperparameters were fine tuned to maximize the test accuracy with minimal overfitting.

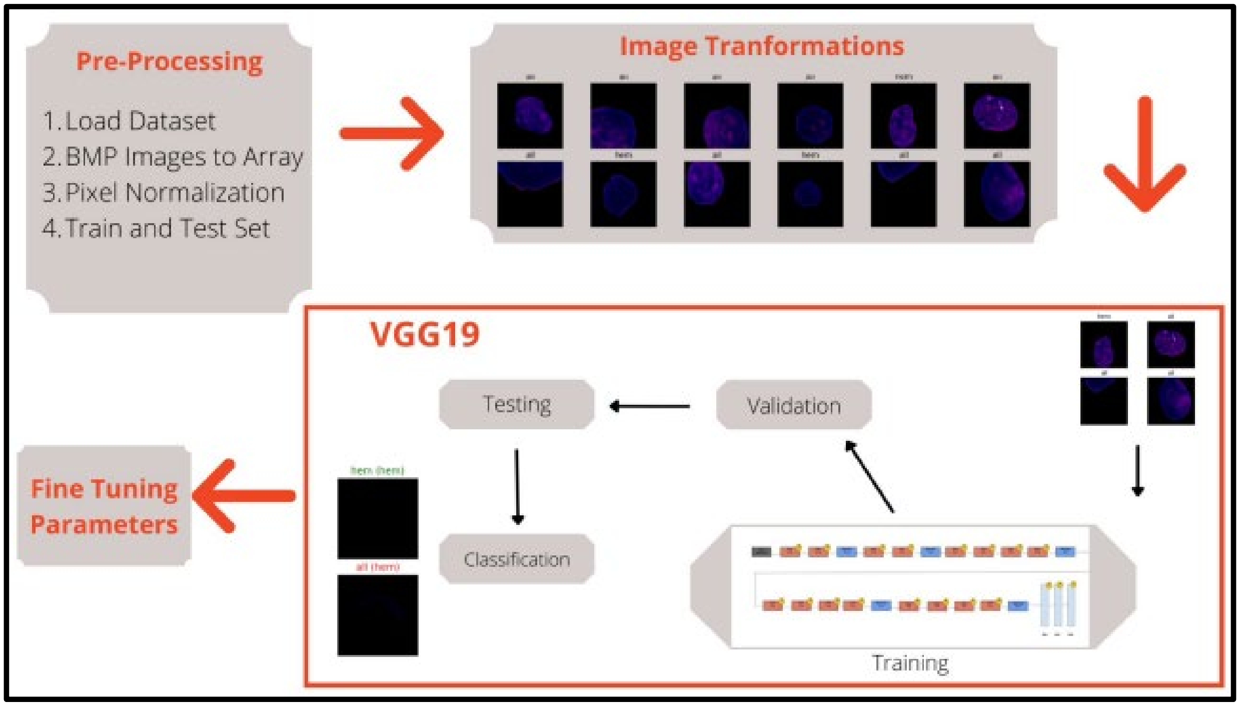

Figure 3. VGG-19 Implementation

Results

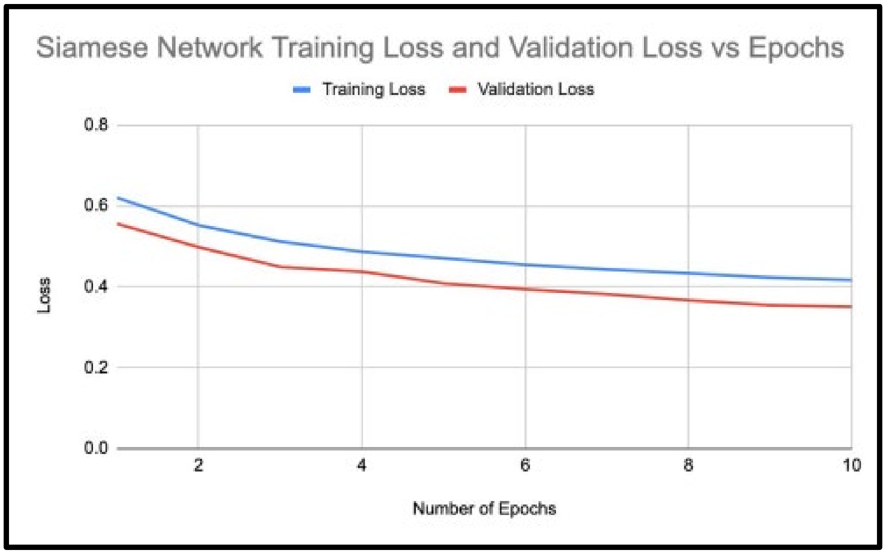

Figure 4. Siamese Network Loss/Epoch Graph

For the Siamese Network, the training loss and validation loss graph that indicated that there was little to no overfitting of the data for a 85% accuracy, which means that the results can be generalized to other cancer datasets, meaning that our initial procedure for few shot learning has potential to be expanded to other types of pediatric cancers and solid tumor cancers such as lung and cervical cancer. Traditional CNN's seem to show signs of overfitting or a small validation set, which indicates that the accuracy received by traditional CNN's may not be applicable to other ALL-based datasets. However, further testing is necessary to determine to what extent overfitting is present in both the Siamese Network and the traditional CNN's that were tested.

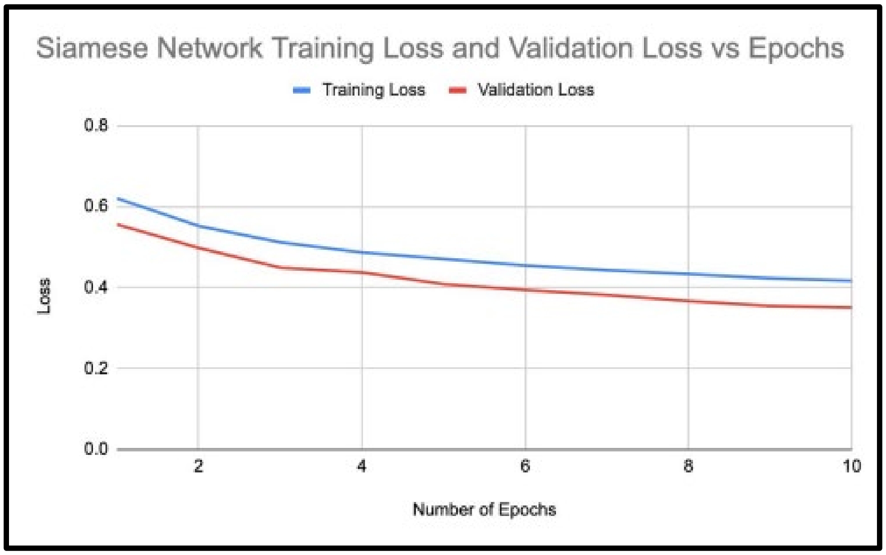

Figure 5. VGG-19 Network Loss/Epoch Graph

The results received by training a VGG-19 CNN with the CNMC dataset indicate that traditional CNN's are still able to receive a higher accuracy rate compared to FSL. An accuracy rate of 96.6% was received, and there were less issues with the training features not generalizing well to the validation set, as there were no validation loss values greater than 1. Overfitting was likely not an issue with this model, as a Dropout Layer was implemented to prevent overfitting, but with a larger validation set, our validation loss graph trend would appear more uniform than it does currently. Between the three models that were tested, the VGG-19 has the highest variances in its training loss graph, which would indicate that there may need to be some adjustment to the training set to finetune the model.

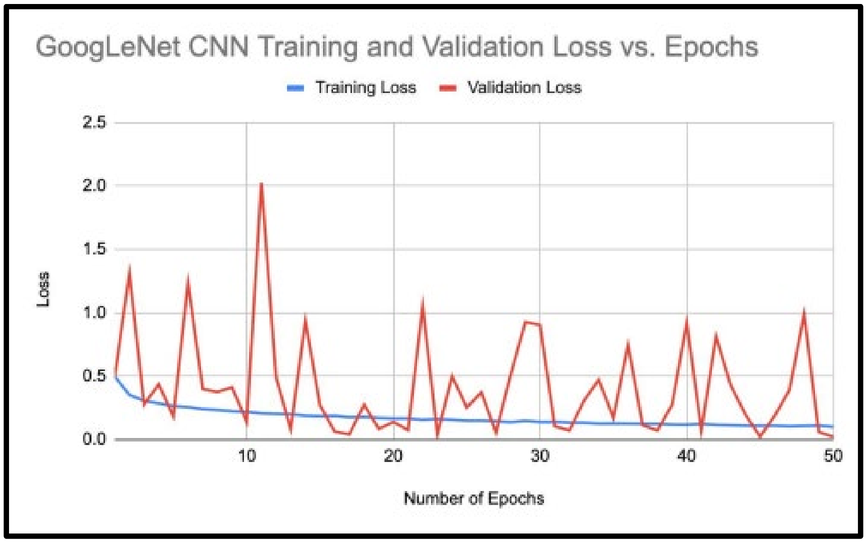

Figure 6. GoogLeNet Loss/Epoch Graph

The results shown by the GoogLeNet Training and Validation graph show that the model was successful in the binary classification of ALL, represented by the decreasing values of Training Loss as the number of epochs increases. However, there is a fluctuation in the results received by the GoogLeNet Validation Loss graph, indicating that the validation set may have been smaller than the optimal amount. Some of our validation loss values were much higher than 1, indicating that the training features are not generalizing well to the validation set. With the 99% accuracy rate that was achieved, it can be concluded that transfer learning is an effective method for ALL diagnosis, but the validation set may need to be expanded to prevent the high validation loss values from occurring.

Conclusions

Currently results for three CNN's have been gathered: GoogLeNet, VGG-19, and a Siamese CNN. The same normalization procedure was applied for the pre-processing for the GoogLeNet and the VGG-19. Some overfitting mitigation strategies such as a Dropout Layer and a relatively smaller amount of epochs were implemented for the comparison to past studies involving this dataset. The highest accuracy was received with the GoogLeNet's transfer learning approach, indicating that the integration of past CNN architectures and the usage of transfer learning provides a higher accuracy than either FSL and traditional CNN's.

Comparing our research to some past papers utilizing the CNMC dataset, our VGG-19 model received a lower accuracy rate of 96.6% compared to other comparable traditional CNN's, such as AlexNet, which received an accuracy rate of 98.7%. Our Siamese Network approach has not been used on the CNMC dataset, and it shows potential for being a breakthrough in cancer research, but more fine tuning and augmentation may need to be performed to receive better results. Methods using One-Shot

Learning has received a higher accuracy rate compared to our model, so changing some variables may be necessary to improve future iterations of a FSL model for cancer diagnosis.

Some potential ways to improve/explore this research topic include testing more CNN's, using different activation functions for FSL such as the Triplet Loss Function, and some data augmentation utilizing the Laplace operator to improve 3 channel images compared to its past usage in improving gray scale images.

Currently, it can be concluded with a high degree of certainty that GoogLeNet is the most effective CNN architecture that was tested. However, while an accuracy of 99% was received for the GoogLeNet model, there were issues with a small validation set, which may have caused the accuracy to be higher than it actually would be with a larger validation set. An 85% accuracy was achieved with the Siamese Network architecture, showing that is can be successfully on cancer datasets as it has been used on other datasets in the past such as the Omniglot and the ImageNet dataset, showing that this model architecture can be used with some adaptations for the binary classification of ALL. VGG-19 provided a high accuracy of 96.6%, but there was some evidence of overfitting or the validation set being smaller than what would be necessary to receive a declining trend in the Validation Loss graph. Overall, the results received support the conclusion that FSL is a credible solution to existing ALL classification issues.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Honnalgere, A., & Nayak, G. (2019). Classification of normal versus malignant cells in B-ALL white blood cancer microscopic images. In A. Gupta & R.Gupta (Eds.), *ISBI 2019 C-NMC challenge: Classification in cancer cell imaging* (pp. 1-12). Springer. https://doi.org/10.1007/978-981-15-0798-4_1

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141. https://doi.org/10.1109/cvpr.2018.00745

Kasani, P. H., Park, S.W., & Jang, J.W. (2020). An aggregated-based deep learning method for leukemic B-lymphoblast classification. *Diagnostics*, Article 12. https://doi.org/10.3390/diagnostics10121064

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*, 11–26. https://doi.org/10.1016/j.neucom.2016.12.038

Liu, Y., & Long, F. (2019). Acute lymphoblastic leukemia cells image analysis with deep bagging ensemble learning. In A. Gupta & R. Gupta (Eds.), *ISBI 2019 C-NMC challenge: Classification in cancer cell imaging* (pp. 113-121). Springer. https://doi.org/10.1007/978-981-15-0798-4_12

Maurício de Oliveira, J. E., & Dantas, D. O. (2021). Classification of normal versus leukemic cells with data augmentation and convolutional neural networks. *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, *4*, 685–692. https://doi.org/10.5220/0010257406850692

Prellberg, J., & Kramer, O. (2019). Acute lymphoblastic leukemia classification from microscopic images using convolutional neural networks. In A. Gupta & R. Gupta (Eds.), *ISBI 2019 C-NMC challenge: Classification in cancer cell imaging* (pp. 53-61). Springer. https://doi.org/10.1007/978-981-15-0798-4_6

Shafique, S., & Tehsin, S. (2018). Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. *Technology in Cancer Research & Treatment*, Article 17. https://doi.org/10.1177/1533033818802789

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples. *ACM Computing Surveys*, *53*(3), 1–34. https://doi.org/10.1145/3386252

Yarlagadda, D. V. K., Rao, P., Rao, D., & Tawfik, O. (2019). A system for one-shot learning of cervical cancer cell classification in histopathology images. *Medical Imaging 2019: Digital Pathology*. Article 34, https://doi.org/10.1117/12.2512963