Efficient Diagnosis of Osteosarcoma in Adolescents With Convolutional Neural Network for Histopathology Images

Amy Li[1], Navya Rawal[2], and Aviva Wang[3]

[1]Valley Christian High School San Jose, niuniu95135@gmail.com
[2]Dougherty Valley High school, navyarawal2016@gmail.com
[3]Cary Academy, smartaviva@gmail.com

ABSTRACT

Osteosarcoma, a rare and aggressive form of bone cancer, predominantly affects adolescents and poses significant challenges when providing an accurate and timely diagnosis. Distinguishing between necrotic, viable, and non-tumor in histopathology images is a daunting task for pathologists due to the diverse patterns of osteosarcoma cell morphology. Although prior work on automated multi-classification of osteosarcoma has produced accurate results, those models require immense amounts of computational power which are not widely available. Our work found an efficient machine learning-based approach that produced an 82.41% accuracy rate in classifying the type of osteosarcoma present in a patient through the analysis of whole slide-stained histopathology images. Our distinctive model architecture incorporates activation functions, convolution, max pooling, and fully connected layers. Our simple neural network design substantially reduces the amount of computational power required to reach a conclusive result compared to the existing models. This novel approach holds the potential to aid pathologists in efficiently diagnosing and classifying osteosarcoma. Data and source code.

Keywords: Osteosarcoma, machine learning, classification model, viable tumors, artificial intelligence, Convolutional Neural Network

I. INTRODUCTION

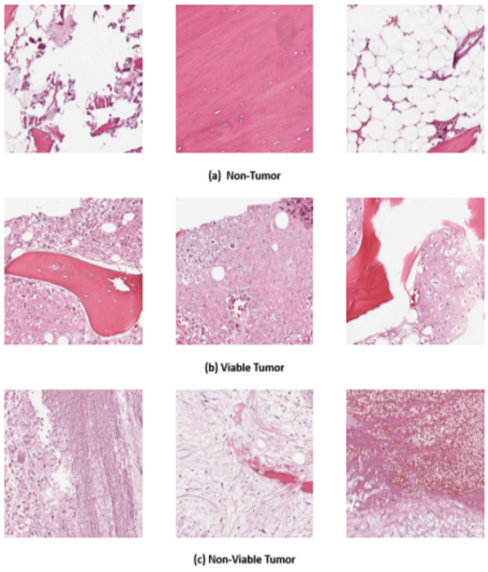

Fig. 1: These images are examples of the three different types of osteosarcomas that our model attempted to classify.

Osteosarcoma is a type of bone cancer that arises from the cells that form new bone tissue. It predominantly affects adolescents and often develops in the long bones of the body, such as the arms or legs, but it can also occur in other bones.

Various diagnostic techniques such as biopsies, which are tissue samples obtained from the tumor site, are used to classify osteosarcoma. Biopsies are examined by a pathologist who specializes in analyzing cells and tissues to confirm the diagnosis of osteosarcoma. (Taupin et al.) Based on these diagnostic assessments, the classification of osteosarcoma can be determined as a non-tumor, viable tumor, or necrotic tumor. The classification of osteosarcoma into these categories plays a crucial role in determining the appropriate treatment strategy, including surgery, chemotherapy, and radiation therapy. It helps medical professionals understand the extent of the tumor, plan interventions, and assess the response to treatment. With an overall 5-year survival rate of just 59% and a distant stage survival rate of 24%, early classification is all the more important (“Survival Rates for Osteosarcoma”). It helps medical professionals understand the extent of the tumor, plan interventions, and assess the response to treatment.

SEER* stage	5-year relative survival rate
Localized	76%
Regional	64%
Distant	24%
All SEER stages combined	59%

*SEER = Surveillance, Epidemiology, and End Results

Fig. 2: Five-year relative survival rates for osteosarcoma based on people diagnosed with osteosarcoma between 2012 and 2018.
"Survival Rates for Osteosarcoma." American Cancer Society, www.cancer.org/cancer/types/osteosarcoma/detection-diagnosis-staging/survival-rates.html. Accessed 25 July 2023.

Unfortunately determining the difference between the three classes is extremely difficult. The cell morphology within a class varies and different classes share similar features. Due to the nuances in osteosarcoma histological images, pathologists struggle to diagnose patients accurately. Additionally, it is an extremely time-consuming process for the pathologist. A large number of whole-slide histopathology images need to be analyzed to reach a conclusive result (Amdt and Crist).

With the advancement in image processing technology and the different methods of creating neural networks, it is now possible to use CNN networks to do the computational labor of analyzing the image. Unfortunately, current models face a similar issue.

Model	Weighted Average Precision	Weighted Average Recall	Weighted Average F1-Score	Accuracy	Epochs	Layers
Our Model	0.81	0.77	0.78	0.8241	50	12
VGG19	0.94	0.94	0.94	0.939	1500	19
MobileNet V2	0.90	0.90	0.89	0.91	100	53

Fig. 3: A chart comparing the results of our model, amount of epochs we ran, and layers in our model to current state of the art models in three class osteosarcoma classification.

The model presented by Anisuzzaman et al reached an impressive accuracy rate of 93.91% for the three class classifications. However, their model required 1500 epochs to achieve this rate. Additionally, their model had a very deep neural network, meaning that it would require great amounts of computing power. (Anisuzzaman et al.)

Recently, Md. Tarek Aziz et al. published research on a novel method to classify histology images of osteosarcoma into categories of nontumor, viable tumor, and necrosis. Their method used deep learning-based feature extraction with processed images. Image slides of 1024 by 1024 pixels were resized to 224 by 224 or less. The images were all converted to tensors and normalized. In this research, a mix of multiple ML models was tested for feature extraction, including VGG16, VGG19, DenseNet 121, and ResNet50. Between these, DenseNet121 performed best and was implemented for feature extraction. RFE was used for binary and multiclass classification. This method achieved a 95.2% accuracy for multiclass classification and a 99.4% accuracy for binary classification. (Aziz et al.)

Mishra et al. have also done research on the classification of osteosarcoma using convolutional neural networks. The approach taken was a standard convolutional neural network. Since different depths in CNN architecture have different strengths and weaknesses, the researchers tried out four different network architectures: AlexNet, LeNet, VGG, and their own proposed architecture. The proposed model is an extended version of a baseline model. The baseline model was an extremely simple model composed of 3 convolutional layers. With baseline architecture achieved an average accuracy of 0.84. To make the model deeper, 5 x 5 filters were switched to two 3x3 filters stacked on each other. Of these four, the proposed architecture performed significantly better in all categories, with 0.924, 0.97, 0.94, and 0.95 for accuracy, precision, recall, and F1 score respectively. (Mishra et al.)

Zhi Li et al. took a very different approach to solving this problem of classification. The experimental procedure used healthy control groups and tumor groups of different types and compared them with each other. Instead of separating tumors directly into necrotic, viable, or non-tumors, a multilayer method was used. First, healthy control patients were compared to tumor patients. Then the tumor patients were compared with each other to distinguish benign tumors from osteosarcoma cases. Experiments proved the Random Forest classifier to be the most effective model overall for this task compared to the logistic regressions and SVM model. The random forest classifier model achieved an average accuracy of 0.97, sensitivity of 1.00, specificity of 0.93, and F statistic of 0.96. (Li et al.)

This paper presents a simple neural network that is able to reach an accuracy rate of 82.4% through a small duration of training. This model requires less computational power and is much more efficient compared to current ones which makes it much more accessible for pathologists to use.

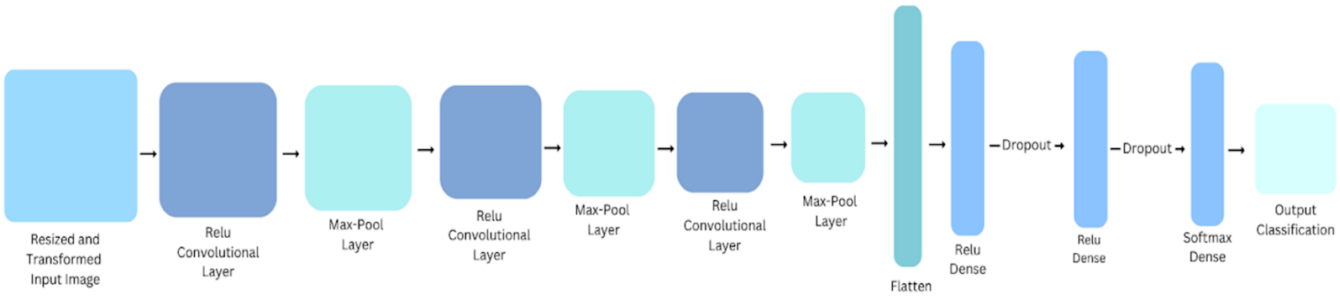

Fig. 4: A simplified visual representation of our unique model.

II. METHODS

A. Data Set

Our dataset comes from the Cancer Imaging Archive. It is composed of 1144 images of Hematoxylin and eosin (H&E) stained osteosarcoma histology images which are 1024 pixels by 1024. The data collection was conducted by a team of clinical researchers associated with the University of Texas Southwestern Medical Center in Dallas. They used samples obtained from 50 patients who received treatment at Children's Medical Center in Dallas between 1995 and 2015. Each image in the dataset was categorized as either Non-Tumor, Viable Tumor or Necrosis (Non-Viable Tumor) based on the predominant cancer type in the image. The data had 536 (47%) non-tumor images, 263 (23%) necrotic tumor images, and 345 (30%) viable tumor images.

B. Data Preprocessing

Manual adjustments were made to ensure the data was appropriate for our model. Since the images in the dataset were originally split into numbered sets, with each set containing 50 images and a CSV file with the respective labels, we manually extracted images from every folder and manually merged the CSV files. Our model required a training, validation, and testing set so we split the data in the ratio of 7:2:1, respectively.

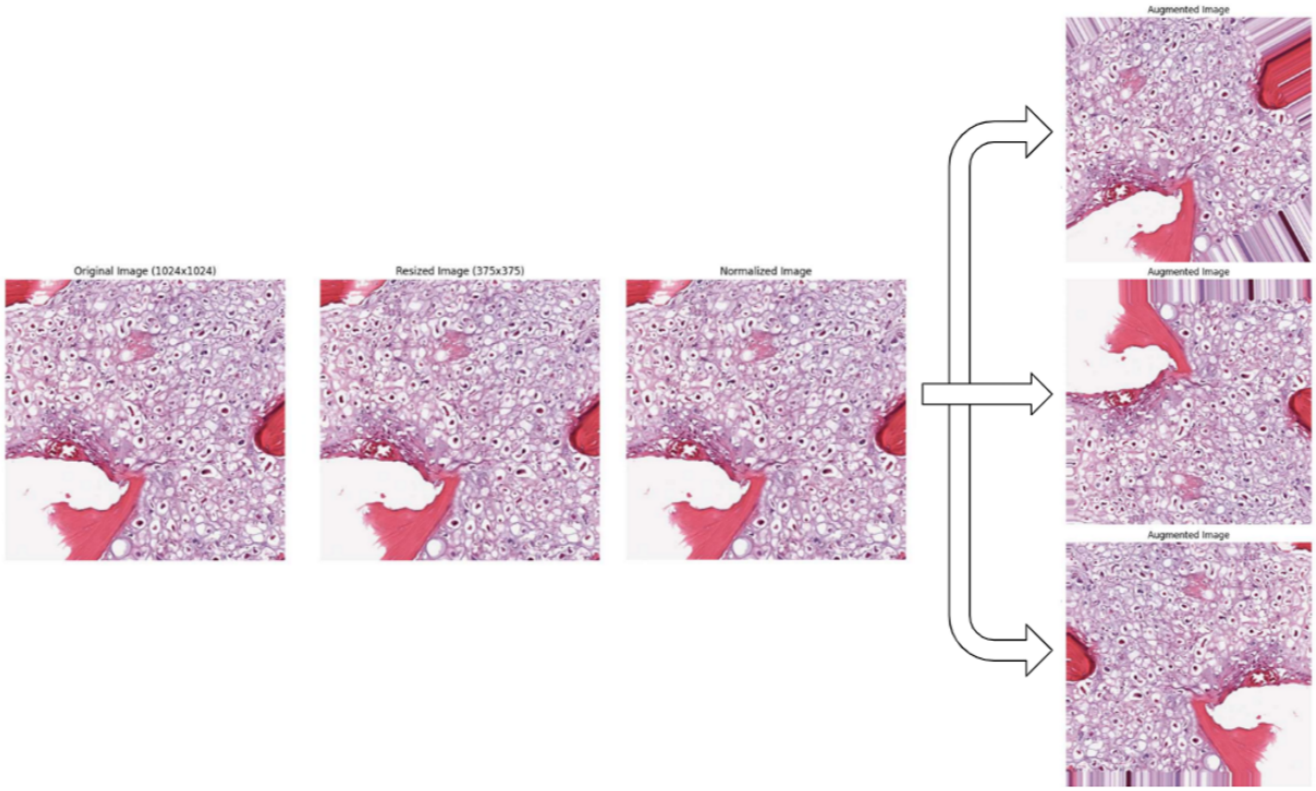

Fig. 5: How our images change when going through preprocessing.

Due to memory and computing power restraints, we resized the histopathology images from 1024 by 1024 to 375 by 375. We then rescaled the image intensities levels to in between the range of 0 to 1. We performed data augmentation, including rotation, width shift, height shift, vertical flip, and horizontal flip, on the images randomly. This was used to expose our model to a variety of transformed versions of the original images during training. This created diverse training samples that help the model generalize better to unseen data and improve its performance on the target task.

C. Challenges Faced

When creating our model, we faced many challenges. At first, we attempted to use transfer learning with VGG 19 because the current state-of-the-art model used it. We planned on applying VGG 19 and adding a few layers to hopefully create a more accurate model. However, when we actually implemented this, we did not have enough computing power to run enough epochs to obtain substantial results. When running this modified VGG 19 model, we would run out of GPU on Google Colaboratory, and simply not have enough time to run enough epochs with CPU. So, we decided to design our own simple convolutional neural network. With this new model that we tested, we found that it was not learning, just guessing one class for every image. This was because we resized the original 1024 by 1024 images to 224 by 224, reducing the dimensions by nearly a fourth and thus losing a significant amount of information. To combat this, we tried not reducing the dimensions at all, however, we faced the same problem of not having enough computing power. After trial and error, we found that resizing the image to 375 by 375 was the perfect balance between not losing significant information and having small enough dimensions to run through our model. When we ran this new model with the adjusted dimensions, we faced the challenge of overfitting. To fix this, we added dropout layers between each fully connected layer. Finally, we achieved a working model that was both accurate and efficient.

D. Model Architecture

Our model was run in the Google Colaboratory environment. To implement our proposed model, the Python 3.10.6 programming language with PyTorch, TensorFlow, Keras, and various libraries such as Numpy, Pandas, Matplotlib, and more were used.

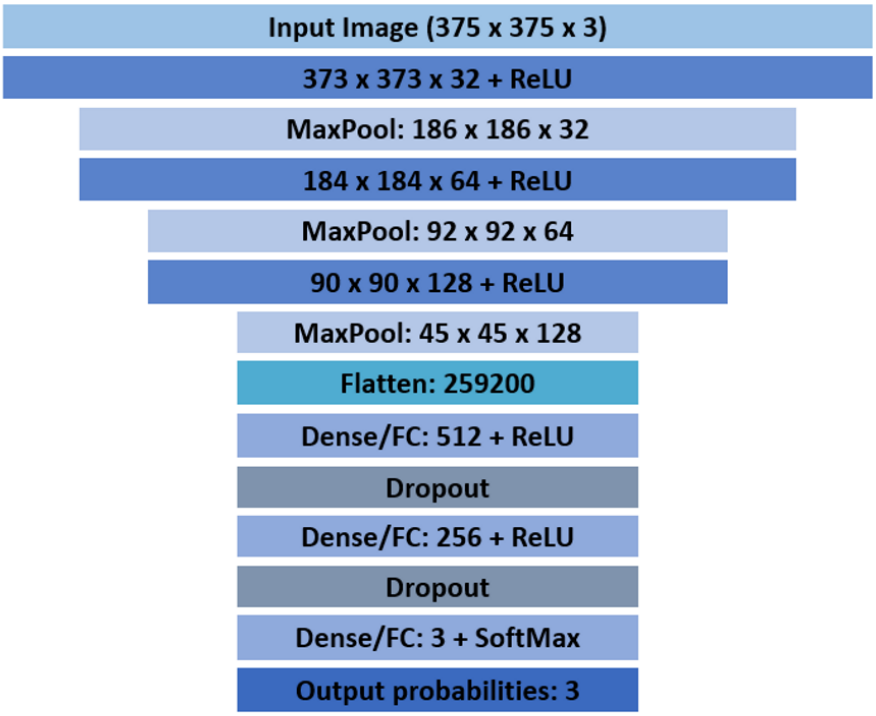

Fig. 6: A diagram of our network architecture showing the dimensions after each layer.

Our model starts by alternating between a convolutional layer with ReLU activation and max pool. It goes through this process 3 times. After that, a flattened layer was used to create a 1D vector and transition the convolutional layers to fully connected layers. Between each fully connected layer, there is a 0.7 dropout layer. Finally, a softmax activation function was used to calculate the probabilities of the 3 classes, producing the final classification.

As you can see in Figure 7, each of the convolutional layers used filters of size 3 x 3. During each convolution layer, the out channels doubled in order to extract more features. Each of the max pool layers used filters of size 2 x 2. We implement dropout layers in between the fully connected layers to combat overfitting.

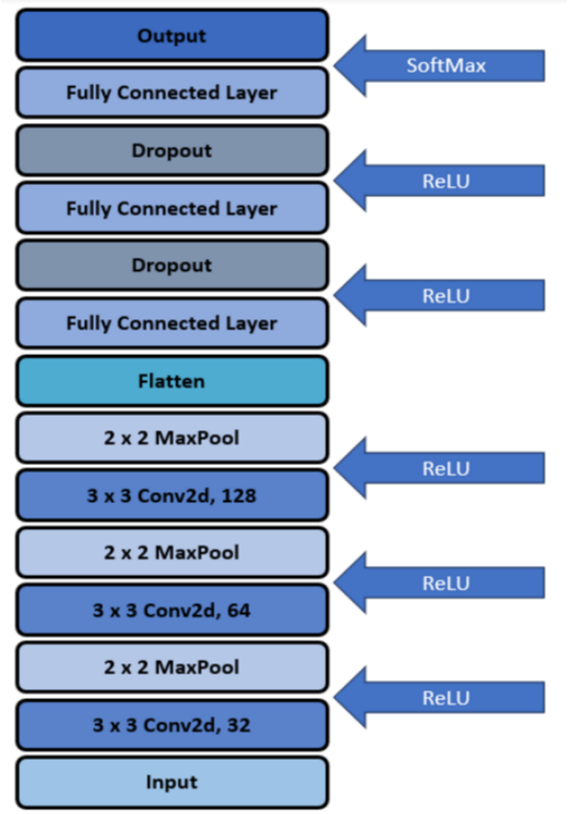

Fig. 7: A diagram of our network architecture that shows the layers used.

III. RESULTS

A. Model Results

$$Accuracy = \frac{Correct}{Total}$$

$$P_A = \frac{TP_A}{TP_A + FP_A} \qquad WAP = \frac{P_A \times A + P_B \times B + P_C \times C}{A + B + C}$$

$$R_A = \frac{TP_A}{TP_A + FN_A} \qquad WAR = \frac{R_A \times A + R_B \times B + R_C \times C}{A + B + C}$$

$$F1_A = \frac{P_A \times R_A}{P_A + R_A} \qquad WAF1 = \frac{F1_A \times A + F1_B \times B + F1_C \times C}{A + B + C}$$

Fig. 8: Formulas for the metrics we used to evaluate the success of our model.

To evaluate the success of our model, we used the metrics accuracy, weighted average precision, recall, and F1-score. We used the weighted average because we were dealing with an unbalanced data set, and the weighted average takes into account the number of images in each class. As you can see in Fig. 9, our model performed decently well in these metrics, but not as well as the current state-of-the-art models. Our model has fewer layers and uses fewer epochs as shown in Fig. 3. This shows how our model performs more efficiently than the current state-of-the-art models.

Metric	Weighted Average Precision	Weighted Average Recall	Weighted Average F1-Score	Accuracy
Our Model	0.81	0.77	0.78	0.8241

Fig. 9: The weighted average results of the metrics we chose to use.

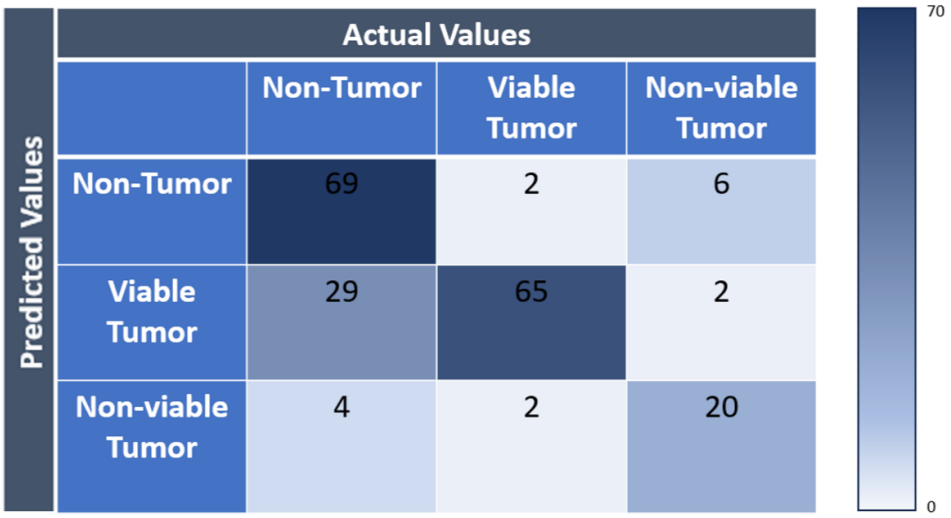

Fig. 10: The confusion matrix from our model with a heat map. Our model performs pretty well, however it confuses Non-Tumors for Viable Tumors most frequently.

B. Limitations

Although our efficiency was much greater than that of current models, our accuracy did not compare. Using our model, we only reached an accuracy rate of 82.41%, which is much lower than the current state of the art. Due to computing restraints, we had to resize the images in the data set from 1024 by 1024 to 375 by 375, thus losing key information the model could have used to train.

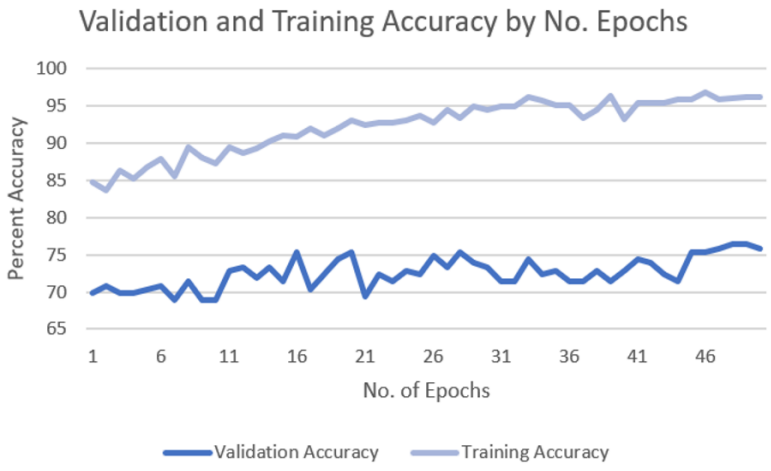

Fig. 11: Validation and training accuracy by number of epochs ran.

Additionally, we did not have the resources to run enough epochs. Previous models ran more than a thousand epochs to achieve their accuracy rates. Fig. 11 indicates our accuracy was steadily increasing, suggesting with more epochs we could have had similar results. Lastly, our data set was unbalanced. Although we had done some work to augment the data and split it into three categories so that it would best fit our model, having a balanced data set would have aided our model during training. We believe that with a better data balance, our model could perform at a better rate. The unavailability of memory and computing power made our model have a comparatively low accuracy, however, we believe that with additional resources, our model could perform efficiently and have a high accuracy rate.

IV. CONCLUSION

With the hard and tedious nature of diagnosing osteosarcomas, as well as the low survival rate of necrotic tumors, it is essential to automate the accurate diagnosis of these tumors. In this study, we created our own simple convolutional neural network that is able to efficiently and accurately classify a whole slide histopathology image of bone cells into either non-tumor, viable tumor or non-viable tumor. Although our accuracy was not comparable to current state-of-the-art models, reaching only 82.41%, our model used less computing power and was more efficient than current models, which is where our novelty lies. We hope that in future works we are able to build off the efficiency of our model to create a model with a higher accuracy rate while still maintaining .

REFERENCES

Amdt, C. A., and W. M. Crist. “Common musculoskeletal tumors of childhood and adolescence.” *The New England Journal of Medicine*, vol. 341, no. 5, 1999, pp. 342-52. *PubMed*, https://pubmed.ncbi.nlm.nih.gov/10423470/.

Anisuzzaman, D. M., et al. “A Deep Learning Study on Osteosarcoma Detection from Histological Images.” 2020. *Arxiv*, https://arxiv.org/pdf/2011.01177.pdf.

Aziz, Tarek, et al. “A Novel Hybrid Approach for Classifying Osteosarcoma Using Deep Feature Extraction and Multilayer Perceptrom.” *Diagnostics 2023*, vol. 13, no. 12, 2023, p. 2106.

“Confluence Mobile.” *Confluence Mobile - Cancer Imaging Archive Wiki*, 9 February 2016, https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=52756935. Accessed 30 July 2023.

Huang, Shigao, et al. “Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges.” *Cancer Letters*, vol. 471, 2020, pp. 61-71.

Li, Zhi, et al. “Classifying osteosarcoma patients using machine learning approaches.” *Annual International Conference of the EEE Engineering in Medicine and Biology*, 2017, pp. 82-85. *PubMed*, https://pubmed.ncbi.nlm.nih.gov/29059816/.

Mishra, Rashika, et al. “Convolutional Neural Network for Histopathological Analysis of Osteosarcoma.” *Journal of computational biology: a journal of computational molecular cell biology*, vol. 25, no. 3, 2018, pp. 313-325. *PubMed*, https://pubmed.ncbi.nlm.nih.gov/29083930/.

“Osteosarcoma | Osteosarcoma Information.” *American Cancer Society*, https://www.cancer.org/cancer/types/osteosarcoma.html. Accessed 25 July 2023.

“Survival Rates for Osteosarcoma.” *American Cancer Society*, 1 March 2023, https://www.cancer.org/cancer/types/osteosarcoma/detection-diagnosis-staging/survival-rates.html. Accessed 30 July 2023.

Taupin, Tatiana, et al. *Diagnostic and Interventional Imaging*, vol. 97, no. 3, 2016, pp. 327-331. *Science Direct*, https://www.sciencedirect.com/science/article/pii/S2211568415003678.

Vezakis, Ioannis A., et al. “Deep Learning Approaches to Osteosarcoma Diagnosis and Classification: A Comparative Methodological Approach.” Edited by Shinji Miwa. *Cancers*, 2023. *National Library of Medicine*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10136449/.

ACKNOWLEDGEMENTS

We thank Shailja (University of California Santa Barbra), Arthur Caetano (University of California Santa Barbara), and Satish Kumar (University of California Santa Barbra) for teaching us about neural networks and providing mentorship throughout our research. We would also like to thank Dr.Lina Kim (University of California Santa Barbara) and Dr.Teresa Holden (University of California Santa Barbara) for hosting Summer Research Academy and giving us the opportunity to perform this research.

AUTHOR CONTRIBUTION STATEMENT

N.R. and A.W. both experimented in creating a model, N.R. created the successful model and experimented with it, N.R. wrote the abstract and introduction, A.W. and N.R. created the figures, A.L. wrote the references, A.L. reviewed current work, A.W. and N.R. wrote the methods, A.W. and N.R. wrote the results, A.W wrote the conclusion. All authors reviewed the manuscript.