International Journal of Scientific Research and Engineering Development-– Volume 6 Issue 4, July- Aug 2023
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A Comparative Analysis of Supervised and Semi-Supervised Learning Models for Malaria Classification Using Explainable AI Techniques

Aryan Agrawal*, Conrad Hsu**, Stephen Tan***
*(Dougherty Valley High School
Email: aryanrajagrawal@gmail.com)
** (Carlmont High School
Email:conrad.m.hsu@gmail.com)
***(Homestead High School
Email: stephenctan2@gmail.com)

----------**********----------

Abstract:
Accurate and timely detection of malaria is crucial for effective disease management. Researchers have been implementing computational models to help doctors in diagnosing malaria. There are two important challenges: 1) collecting labeled data is expensive, and 2) many models remain as black boxes and are not inherently transparent or interpretable to humans. Using the Malaria Cell Images Dataset from Kaggle, we first explore the use of semi-supervised learning to obtain high accuracy with limited labeled data and examine different methods to efficiently select candidates for labeling. Second, we employ Explainable AI tools to improve the transparency of these models. We use Explainable AI on a fully supervised model. For semi-supervised learning, we adapted the state-of-the-art FixMatch model for our data and achieved 96% accuracy with 4000 (25%) labeled data, which is comparable to the best result on the same dataset. For Explainable AI, we apply Grad-Cam and SHAP to give detailed insights into our model. Removing the transparency in our model will allow for its improvement and further understanding for classifying malaria for researchers and patients.

Keywords **—Convolutional Neural Networks (CNN), Malaria, Explainable Artificial Intelligence (XAI), Grad-CAM, SHAP, supervised, semi-supervised.**

----------**********----------

I. INTRODUCTION

Malaria is adisease spread through the Anopheles mosquito, who transmits the malaria parasites in the blood cells from an infected person to a new host [1], and itaffects nearly four billion people worldwide [10]. In the absence of automated medical diagnosis, malaria cases were often tested using thin blood smears under a microscope, while visually searching for infected cells. A clinician then manually counts the number of infected blood cells. The problem with this method is that manually counting often has high error rates and is a slow and tedious process [11]. In addition, the knowledge of malaria is limited to the common people, and many do not have access to deeper knowledge on malaria, receiving information only through their doctors. Therefore, we ask the question: How can we classify malaria more efficiently and how can we improve interpretability? We look at Convolution Neural Networks to solve this problem.

The advent of convolutional neural networks (CNNs) has greatly improved the process of malaria

diagnosis [2, 3, 4]. Jane Hung et al. used a 16-layer CNN model for malaria detection, achieving 97% accuracy in 2016. Zhaohui Liang et al. extended malaria diagnosis to use a fast region-based CNN, allowing detection for pictures containing hundreds of cells. In 2020, Gautham Shekar et al. compared the performance of a basic CNN (VGG-19) and a Fine Tuned VGG-19 for malaria detection and obtained accuracy rates ranging from 92% to 96%. It is important to note that all of these models are supervised models, where every image was manually labeled.

However, a challenge of using these models and supervised models in general is efficiently collecting labeled datasets. Labeled data often requires a large amount of human labor, and this can become costly when such labeling is done by experts in the specific field of study. One potential solution is to use semi-supervised learning, which exploits the abundance of unlabeled data for training [5]. State-of-the-art semi-supervised learning methods such as MixMatch [6] and FixMatch [7] employ pseudo-label generation for unlabeled data, followed by training using strong augmentation. MixMatch achieved 89% accuracy on CIFAR-10 by using mixing as augmentation. FixMatch further improves accuracy to 94.9% by using strong augmentation. We adopt FixMatch for our task of malaria cell classification and obtain a 97% accuracy with fully supervised learning.

In addition to the implementation of a semi-supervised model, former models fail to implement interpretability for patients and do not explain exactly how their models work. This is why we implement explainable AI. Explainable Artificial Intelligence (XAI) has recently emerged as a popular field, due to its potential to finally dispute the “black box” problem, which refers to not being able to understand how an AI algorithm arrives at a particular conclusion [13]. Especially in the medical domain, transparency of AI models are held to an even higher standard when it comes to healthcare and making critical medical decisions that may impact patients [14]. In the context of medical imaging, a common approach to analyzing these images is through Convolutional Neural Networks (CNN), due to its ability to employ multiple layers of convolution and pooling operations that can extract meaningful features at different levels of abstraction [15]. However, neural networks often get opaque with a multitude of dense layers that attempt to model the neural architecture of the human brain; thus, as the models get more complex, the accuracy of the model tends to increase, while the interpretability or explainability of the model tends to decrease, also known as the well-known accuracy-interpretability-trade off [16]. Current popular approaches in XAI of medical imaging involve visualization techniques such as the use of Grad-CAM, a visual explanation algorithm that generates a heatmap overlay on the input image, highlighting the regions of the image that contributed the most significantly to the output; Grad-CAM is used primarily for visual explanations in CNN networks [17]. In addition to Grad-CAM, perturbation-based explanations, which involve altering input features on the output of the model, are popular as seen in SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) [18]. Since black box models must be explained globally (explaining the whole model) or be explained locally (explaining a single instance), LIME is more suitable for local explanations while SHAP provides both global and local explanations [18]. All the methods mentioned above are model-agnostic, meaning that these methods are not a model itself, but rather a method that is applied on top of an already trained model, such as a CNN in our research. In our research, we apply both Grad-CAM, a more visual based explanation, and SHAP, a framework that is not limited to visual data or CNNs, to create a comparative analysis between two popular XAI frameworks and compare their explainability on two differently trained models. By doing so, we can get a greater insight into supervised versus a semi-supervised CNN model and compare the XAI explanations of these models to relevant biological data on malaria.

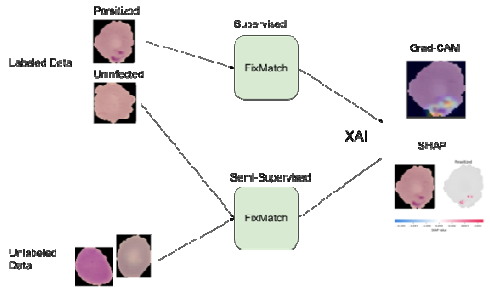

Fig. 1 Overall Architecture of our project. Both the supervised and semi-supervised are run through the FixMatch model, then SHAP and Grad-Cam are used.

So overall, our contributions to improving efficiency and interpretability of models for identifying malaria are:

- Implementing a semi-supervised model using the FixMatch model in order to reduce manual labor.
- Creating a fully supervised model that achieves similar accuracy to the state-of-the-art.
- Using XAI methods Grad-CAM and SHAP to view the models and gain a deeper understanding between how the models work, where the general outline can be seen in **Fig. 1.**

II. METHODS

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

A. Supervised and Semi-Supervised Methods

FixMatch is a state-of-the-art model for semi-supervised learning, which we also adapted for supervised learning. It is a simplified model that uses two main components: consistency regularization and pseudo-labeling. Consistency regularization uses unlabeled data to ensure that the model recognizes images that are modified to have the same labels, while pseudo-labeling uses the model itself to obtain artificial labels for the augmented images. Using these two approaches, the model first selects unlabeled images that it is confident with the guessed label, and then strongly augments these images and assigns it the same guessed label. The artificial label is only kept if the model assigns a high probability to either class. This approach is applied to CIFAR-10 and achieves a 94.93% accuracy with 250 labels and 88.61% with 40 labels.

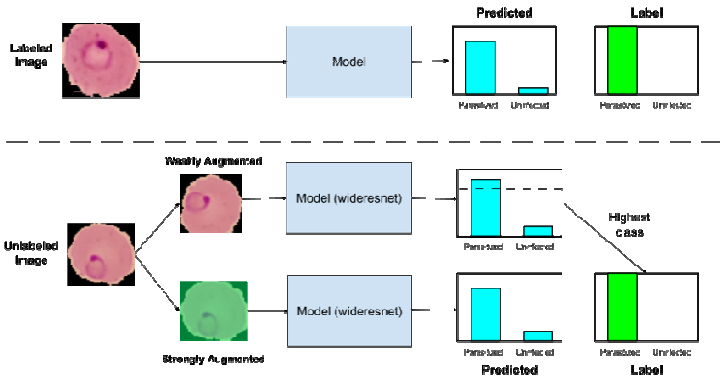

Fig. 2. FixMatch model outline. It takes both the labeled data for the supervised learning and unlabeled data for semi-supervised.

We use the WideResNet model used by FixMatch and adapt it for the use of malaria cell classification. We then train the model for 40 epochs for several different times: once fully supervised with 16,500 labeled data, and once for 4000, 1000, 200 and 100 labeled data each.

We then train using labeled data for 100 samples that is manually selected in four different ways:

- Selecting the largest 100 samples.
- Selecting 100 samples that are most similar to each other.
- Selecting 100 samples that are most dissimilar to each other.
- Randomly selecting 80 samples and selecting 20 samples that are both dissimilar and hard to classify.

B. XAI Methods

We utilized Grad-CAM, a class activation visualization XAI technique that visualizes where the CNN model is looking. After training the supervised CNN model and saving it to a HDF5 file (TensorFlow file), it takes the last convolutional

layer in the pretrained model, which for our supervised model, was a convolutional layer with 147,584 parameters. We inputted the file path for the malaria image we wanted to look at, and the Grad-CAM model shares the same input as the original pre-trained model but has two distinct outputs. The first output represents the activations of the last convolutional layer in the original model, capturing the essential features extracted from the input image. The second output provides us with the model's final prediction, indicating the probabilities of the image belonging to different classes. Next, gradients are computed from the predicted class score of the output of the last convolutional layer through the feature maps produced by the last layer of our CNN.

$$\alpha_k^c = \frac{1}{Z}\sum i \sum j \frac{\partial y^c}{\partial A_{ij}^k} \quad (1)$$

$$L_{Grad-CAM}^C = ReLU\left(\sum k\, \alpha_k^c A^k\right) \quad (2)$$

Equation 1 calculates the importance weight kc by averaging the partial derivatives over all spatial locations in the feature maps for the specific channel k. It indicates how sensitive the predicted class score $y^c$is to changes in the activations A_{ij}^k of the k-th channel. Equation 2 represents the class-discriminative localization map for class c, and the equation takes the weighted combination of feature maps A^k based on the importance weights α_k^c. $\sum k\, \alpha_k^c A^k$)then performs the weighted sum on the feature maps, and the ReLU is applied to the weighted sum, highlighting the regions in feature maps with a positive influence on predicting class while suppressing the negative contributions. Finally, the heatmap is normalized from 0 to 1, with varying colors indicating the intensity of the regions. Finally, we superimposed the heatmap onto the original image using the superimposed function built into Grad-CAM.

To utilize SHAP, we used our pre-trained supervised model and a custom image for interpretation is again prepared for input into the model. After resizing the image to the required dimensions and converting it into a numpy array, we imported the SHAP library. A binary masker was created using a function built into SHAP, which hides parts of the input image during the SHAP value computation. SHAP values are calculated by calculating the average marginal contributions of each feature to all possible coalitions in a game. Below is the how SHAP values are calculated:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M-|z'|-1)!}{M!}\left[f_x(z') - f_x(z' \setminus i)\right] \quad (3)$$

Equation 3 at a high level, calculates the prediction of the model without feature iand also the prediction of the model with feature i, then calculates the difference. The difference is known as the marginal value. This process is repeated for each permutation of subsets and each of those is additionally weighted depending on the number of features M out of the total number of features present in that particular subset. In the context of our CNN model, for each feature in input image x, a perturbed image x' is created. The difference in the model's prediction for x' and x are calculated, and the SHAP values quantify the impact of each pixel on the model's prediction, considering all possible combinations of pixel values in the image. The final SHAP values are then computed by averaging the differences in model predictions. In our model, we had parameters max_eval which specifies the maximum number of samples used to approximate the SHAP values and batch_size, which controlled the size of the batches used during SHAP computation. We used a max_eval of 5,000 and a batch_size of 5. Once the SHAP values are calculated, they are visualized by plotting it over the original image, highlighting the regions that pushed the model towards classifying the specific image for the particular class and the regions that contributed to not classifying that image for that particular class.

III. RESULTS

C. Supervised and Semi-Supervised Results

Unsurprisingly, we find that the model has higher accuracy when there is more labeled data. We can also see that with the supervised model that precision, ROC, Recall, and F1 score are all relatively high, seen in Table 1, suggesting that the supervised model performs better than the semi-supervised.

	Precision	ROC AUC	Recall	F1 Score
Supervised	0.938	0.986	0.971	0.955
Semi-supervised (4000 labels)	0.902	0.967	0.935	0.923

Table 1. Metrics used to measure supervised and semi-supervised model, includes precision, ROC Area Under Curve(AUC), Recall, and F1 Score.

It turns out that semi-supervised learning is a good trade-off between accuracy and number of labeled samples. The fully supervised model (in green) has a 97% accuracy, while the accuracy for the model trained on 6% labeled data (yellow curve) has approximately a 95% accuracy. When the model is trained on only 6% data, its accuracy only drops by 2% despite losing 94% of its labeled data, as shown in Fig. 3. Therefore, semi-supervised learning is a promising approach in the context of malaria detection.

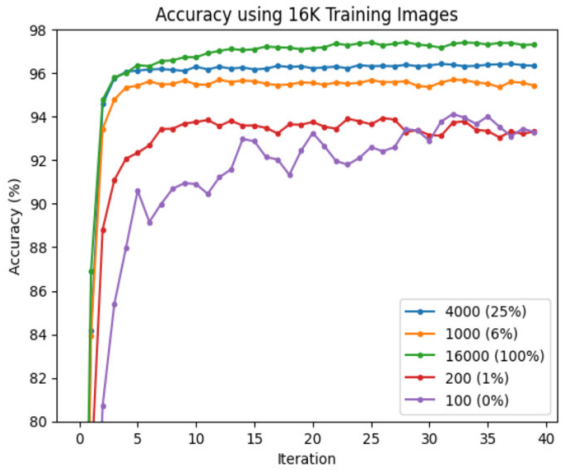

Fig. 3. Graph of accuracy as a function of number of epochs.

Fig. 4. Graph of loss as a function of number of epochs.

The result of training our fully supervised model on FixMatch is comparable to other state-of-the-art research using the same Malaria Cell Images dataset from Kaggle. Previous research papers, such as Jane Hung et al. [2] and Gautham Shekar et al. [4] have applied CNN models on this dataset and have also obtained resulting accuracies of around 96-97%.

However, we also uncover a relatively surprising finding: towards the end of training the 40 epochs, the accuracy for 100 labeled data is slightly higher than that of the 200 labeled data. One possible explanation for this inconsistency may be that sample selection for the 100 labeled data was "luckier" than usual. This prompts the question of whether or not sample selection for the 100 labels could be arranged in such a way that accuracy is optimized. After comparing the test results of 100 labeled data selected using various criteria, we have three key observations. First, selecting 100 labeled samples that are similar to one another (green curve) performs much worse than randomly selecting the 100 samples (e.g. Fig. 5), which shows that it is highly important to select samples in an informed way. Second, when selecting 100 labeled samples that are dissimilar to one another but not necessarily of low confidence (red curve), the model has an accuracy considerably better than the previous result, but still underperforms compared to random selection, as shown in Fig. 5. This result

demonstrates the importance of selecting labeled samples for which the model has low confidence. Finally, we observe that the results with 80% randomly selected data with 20% data being both dissimilar and difficult to classify (purple curve) had an accuracy exceeding that of the randomly selected labels. We also observe from Fig. 5 that the most dissimilar labels were also close to random selection, which supports the idea that selecting labels that are dissimilar and difficult to classify can potentially yield higher accuracies.

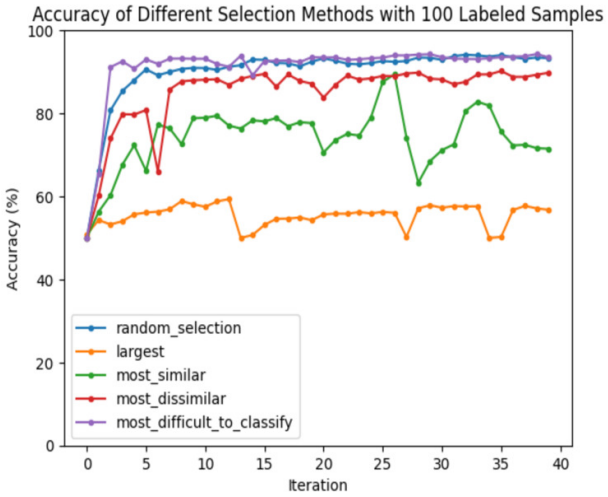

Fig. 5. Graph of accuracy as a function of number of epochs for sample selection with 100 labeled data.

D. Results for Grad-CAM and SHAP

We implemented Grad-CAM and SHAP only on the supervised model, due to its higher performance based on our metrics. We implemented Grad-Cam on two images, one for an uninfected image and one for a parasitized or infected image, both which the model classified correctly.

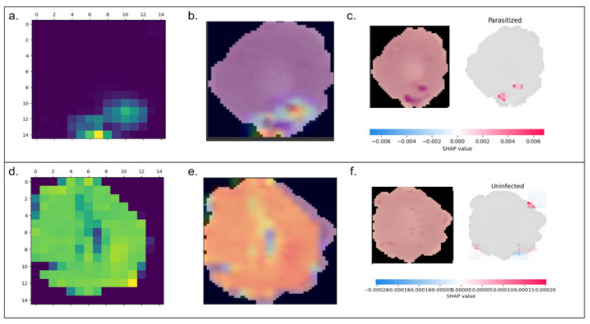

Fig. 6 XAI methods for infected(a, b, c) and uninfected(d, e, f) classes where model correctly predicted class. Images (a, d) are Grad-CAM heatmaps, images (b,c) are the superimposed heatmap on original image, and images (c, f) are SHAP values.

For the correctly classified images, we see that the Grad-CAM creates a heatmap on the area of the image that contributed most to the model's prediction. For the parasitized image (**FIG. 6a**), we clearly see that the targeted area of infection was towards the bottom right of the blood cell. This helps us validate that our model is indeed looking at the correct place, as the bottom right was where the malaria infection was occuring. Furthermore, the SHAP values help to further validate the Grad-CAM's findings, as SHAP indicates for the infected image that the bottom right had positive SHAP values, which meant those pixels helped push the model positively towards predicting this class (**FIG. 6c**). For the uninfected images, the images of the blood cells to the naked eye are harder due to the absence of a specific area that exhibits protruding colors unlike (**FIG. 6c).** As seen in the uninfected blood cell image (**FIG. 6f)**, despite having any area that stands out, our Grad-CAM model indicated that roughly the whole area of the blood cell was used to arrive at the conclusion of uninfected (**FIG. 6e).** However, SHAP indicates that the top right corner and the bottom left corner of the blood cell helped push the model towards classifying the uninfected class (**FIG. 6f)**; on the other hand, we see negative SHAP values or blue SHAP values towards the bottom right corner, indicating that this area had the possibility of pushing the model towards classifying the model towards predicting the wrong class.

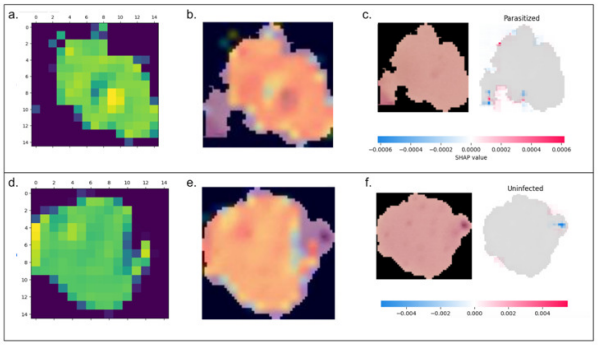

Fig. 7. XAI methods for infected(a, b, c) and uninfected(d, e, f) classes where model incorrectly predicted class. Images (a, d) are Grad-CAM heatmaps,

images (b,c) are the superimposed heatmap on original image, and images (c, f) are SHAP values.

For the incorrectly classified images, which happened to be less than 2.5% of all the images in the test set, we see that this occurs when there isn't a clear spot of infection in the original blood cell image (**Fig. 7f).** The incorrectly classified image is also indicated by a heatmap that spanned the entire image, with no clear distinction of colors (**Fig 7b and 7e**). Compared to the SHAP values displayed in **(FIG. 6c and 6f)**, the SHAP values for instances where the model predicted incorrectly contain a greater mix of red(positive) and blue(negative) SHAP values (**FIG. 7c and 7f).** This discrepancy between SHAP values indicates how confused the model is between predicting the parasitized class and the uninfected class, constantly being thrown in between positive and negative SHAP values. However, (**Fig 7f.)** shows that the malaria pigment in the upper right corner had negatively contributed to the model predicting the uninfected class, which means our model is doing the correct job of identifying where the malaria infection is; however, it seems like our model felt the positive SHAP values around the outer edges pushed the model into identifying the image as uninfected, which happened to be the wrong class. The same concept occurs in (**FIG. 7c**), except for the fact that the bottom left corner pigmentation had tricked the model into thinking the image was parasitized, when in reality it was uninfected. While the Grad-CAM thought the middle area of the blood cell was the most important in deciding that this image was parasitized (**FIG 7b),** SHAP seemed to think the bottom left corner was the area that influenced the model's prediction the most (**Fig 7c).** By using SHAP and Grad-CAM in conjunction, we see that we can better identify exactly where the model is looking at and gain insight into what pushed the model into making correct predictions and incorrect predictions.

IV. CONCLUSIONS

From its remarkable results using the FixMatch pipeline, we conclude that semi-supervised learning shows promise in the application of malaria detection and has the potential to solve the problem of the difficulty involved in collecting labeled data. It can compromise lots of labeled data with only a small decrease in accuracy and achieved results comparable to state-of-the-art research on malaria classification on the same dataset. Sample selection also shows great promise for improving semi-supervised learning. Additionally, we find that semi-supervised learning is slightly less effective compared to supervised learning in the context of malaria detection, however XAI methods such as Grad-CAM and SHAP give us deeper insight into our models and helps mitigate the "black box problem" with neural networks. Due to time constraints, we were unable to implement Grad-CAM and SHAP to the semi-supervised model, thus being unable to thoroughly explain why the semi-supervised model does worse than the supervised. In terms of future work, implementation of XAI into our semi-supervised model would be most obvious once its performance becomes on par with the fully supervised model. In addition, our model can be improved through data augmentation and helping the model focus on specific characteristics. Furthermore, the Grad-CAM and SHAP tools can be used to further advance our model and understanding of malaria classification by comparing these tools to relevant biological data on malaria, including professional doctor diagnosis of blood cells.

ACKNOWLEDGMENT

We would like to thank S. Shailja for her insightful lectures and ideas, as well as Arthur Caetano and Satish Kumar for all their support and immensely helpful feedback that made this research possible. We would also like to thank the Summer Research Academies in providing us the opportunity for conducting this collaborative research project for a college course.

REFERENCES

[1] "CDC Malaria Program." *Centers for Disease Control and Prevention*, 6 Apr. 2023, www.cdc.gov/malaria/resources/cdc_malaria_program_2023.html.

[2] Hung, Jane, et al. "Applying Faster R-CNN for Object Detection on Malaria Images." *Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Workshops*, U.S. National Library of Medicine, July 2017, www.ncbi.nlm.nih.gov/pmc/articles/PMC8691760/.

[3] Liang, Zhaohui, et al. "CNN-Based Image Analysis for Malaria Diagnosis." *IEEE*, 2016, ieeexplore.ieee.org/document/7822567.

[4] Shekar, Gautham, Revarthy, S, and Goud, Ediga Karthick. "Malaria Detection using Deep Learning." *IEEE*, 2020, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9143023.

[5] Zhu, Xiaojin. "Semi Supervised Learning Literature Survey." 2005 September 7. https://minds.wisconsin.edu/bitstream/handle/1793/60444/TR1530.pdf?sequence=1&isAllowed=y.

[6] Berthelot, David, et al. "MixMatch: A Holistic Approach to Semi-Supervised Learning." https://proceedings.neurips.cc/paper_files/paper/2019/file/1cd138d0499a68f4bb72bee04bbec2d7-Paper.pdf.

[7] Sohn, Kihyuk, et al. "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence." https://proceedings.neurips.cc/paper_files/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf.

[8] Emadi, Mona, et al. "A Selection Metric for Semi-Supervised Learning Based on Neighborhood Construction." *Information Processing & Management*, Pergamon, Mar. 2021, www.sciencedirect.com/science/article/pii/S0306457320309365.

[9] Mallapragada, P et al. "SemiBoost: Boosting for Semi-Supervised Learning." https://ieeexplore.ieee.org/abstract/document/4633363.

[10] "Malaria." *World Health Organization*, www.who.int/news-room/fact-sheets/detail/malaria.

[11] Kumar, Sumit, Priya, Sneha and Kumar, Ayush. "Malaria detection using Deep Convolution Neural Network." 4 Mar 2023, https://arxiv.org/pdf/2303.03397.pdf.

[12] Shal, Ayushi, and Richa Gupta. "A Comparative Study on Malaria Cell Detection Using Computer Vision." IEEE,/.

[13] "What Is Explainable AI (XAI)?" *IBM*, www.ibm.com/watson/explainable-ai#:~:text=Explainable%20artificial%20intelligence%20(XAI)%20is,expected%20impact%20and%20potential%20biases.

[14] Amann, Julia, et al. "Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective - BMC Medical Informatics and Decision Making." *BioMed Central*, 30 Nov. 2020, bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01332-6.

[15] Yadav, Samir S., and Shivajirao M. Jadhav. "Deep Convolutional Neural Network Based Medical Image Classification for Disease Diagnosis - Journal of Big Data." *SpringerOpen*, Springer International Publishing, 17 Dec. 2019, journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0276-2.

[16] Soberon, Arturo. "How to Interpret Machine Learning Models with Lime and Shap." *How to Interpret Machine Learning Models with LIME and SHAP*, svitla.com/blog/interpreting-machine-learning-models-lime-and-shap.

[17] Reiff, Daniel. "Understand Your Algorithm with Grad-CAM." *Medium*, Towards Data Science, 21 July 2021, towardsdatascience.com/understand-your-algorithm-with-grad-cam-d3b62fce353.

[18] Borys, Katarzyna, et al. "Explainable AI in Medical Imaging: An Overview for Clinical Practitioners – Saliency-Based XAI Approaches." *European Journal of Radiology*, Elsevier, www.sciencedirect.com/science/article/pii/S0720048X23001018.

