Out of Distribution Detection and Adversarial Attacks on Deep Neural Networks for Robust Medical Image Analysis

Anisie Uwimana [1] **Ransalu Senanayake** [2]

Abstract

Deep learning models have become a popular choice for medical image analysis. However, the poor generalization performance of deep learning models limits them from being deployed in the real world as robustness is critical for medical applications. For instance, the state-of-the-art Convolutional Neural Networks (CNNs) fail to detect adversarial samples or samples drawn statistically far away from the training distribution. In this work, we experimentally evaluate the robustness of a Mahalanobis distance-based confidence score, a simple yet effective method for detecting abnormal input samples, in classifying malaria parasitized cells and uninfected cells. Results indicated that the Mahalanobis confidence score detector exhibits improved performance and robustness of deep learning models, and achieves state-of-the-art performance on both *out-of-distribution (OOD)* and *adversarial* samples.

1. Introduction

Deep learning is increasingly making its way into groundbreaking technologies that have high-value applications in the real-world clinical environment. Innovative medical imaging applications and diagnostics are among the most exciting use cases. One such application is developing microscopy-based malaria diagnosis procedures (Ravendran et al., 2015; Silva et al., 2013; Yang et al., 2019). Malaria is a deadly mosquito-borne disease infecting around 300 million people annually (World Health Organization). Since it is mostly prevalent in low-income countries, developing semi-automated microscopy techniques, as alternatives to polymerase chain reaction (PCR) tests and rapid diagnostic tests (RDT), is a low-cost and reliable solution (Wongsrichanalai et al., 2007).

1.1. Applications of deep learning in medical diagnosis

Esteva et al. (2017) developed a convolutional neural network (CNN) model that was trained on $130,000$ clinical images of skin pathologies to detect cancer. The proposed model achieves performance on par with all tested experts, demonstrating an artificial intelligence model capable of classifying skin cancer with a level of competence comparable to dermatologists. In 2018, another research study showed that a convolutional neural network trained to analyze dermatology images identified melanoma with ten percent more specificity than human clinicians (Haenssle et al., 2018). Another algorithm trained on $42,000$ chest CT scans outperformed expert radiologists in detecting lung cancers (Ardila et al., 2019). It was able to find malignant lung modes 5-9.5% more often than human specialists. Recently, a CNN model designed to predict malignancy and identify 134 skin disorders (Cho et al., 2020). The proposed algorithm is capable of distinguishing, at the human expert level, melanoma from birthmarks. There have also been various studies on assessing the uncertainty and robustness in medical data (Senanayake et al., 2016; Laves et al., 2020; Asgharnezhad et al., 2020).

1.2. Applications of deep learning for malaria diagnosis

Various computer vision algorithms have been used for malaria diagnosis (Ravendran et al., 2015). Deep learning algorithms are recently being used increasingly by researchers especially for malaria detection because of its applicability in building automated diagnostic system. Liang et al. (2016) presented a 16-layer CNN towards classifying uninfected and parasitized cells. The study reported that the custom model was more accurate, sensitive, and specific than the pre-trained model. Dong et al. (2017) evaluated three well-known CNNs (LeNet, AlexNet and GoogLeNet) on classifying parasite/not parasite slide images of thin blood stains. Simulation results showed that all three CNNs achieved classification accuracy scores of over 95%. Rajaraman et al. (2018b) introduced a pretrained CNN as a feature extractor towards improved malaria parasite detection in thin blood

[1] African Institute for Mathematical Sciences (AIMS), Rwanda, Kigali [2] Durand Building, 496 Lomita Mall, Stanford University, Stanford, CA 94305. Correspondence to: Anisie Uwimana <auwimana@aimsammi.org>.

Accepted by the ICML 2021 workshop on A Blessing in Disguise: The Prospects and Perils of Adversarial Machine Learning. Copyright 2021 by the author(s).

smear images, and the results present the use of pretrained CNNs as a promising tool in malaria detection.

Rajaraman et al. (2018a) has demonstrated that deep neural networks can be used to detect malaria from microscopic images. However, in order to deploy such systems in medical facilities, it is vital to ensure that the automated detection systems are indeed robust. Nonetheless, deep learning algorithms work on the premise that both training and test data are drawn from the same application-specific distribution. However, in real-world applications, they need to be able to robustly handle anomalous inputs including, 1) adversarial samples arising from image distortion and 2) samples drawn from a different distribution but belong to the same input space.

In this work, we propose using a Mahalanobis distance-based confidence score method (Lee et al., 2018; Kamoi & Kobayashi, 2020; Nitsch et al., 2021) for detecting abnormal (both OOD and adversarial) samples to improve the performance and robustness of pre-trained convolutional neural network models to improve the robustness of malaria detection (Figure 1). The suggested method gives better results compared to the current state-of-the-art method ODIN (Liang et al., 2018) in detecting OOD malaria samples. We also demonstrate that Mahalanobis distance-based confidence score outperforms the state-of-the-art detection, LID, in all test cases, in detecting adversarial samples generated by four adversarial attacking methods: FGSM (Goodfellow et al., 2015), BIM (Kurakin et al., 2016), DeepFool (Moosavi-Dezfooli et al., 2016), and CW (Carlini & Wagner, 2017).

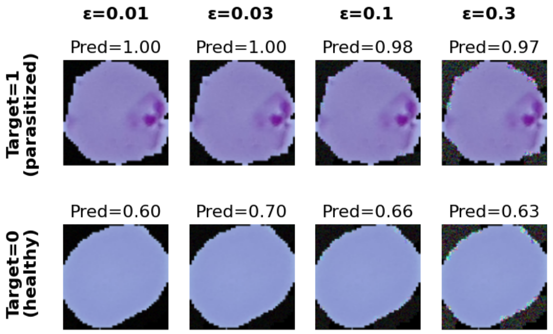

Figure 1. FGSM adversarial attack on parasitized and healthy cells for varying levels of noise. Noise can be visually inspected on the black areas and cell boundary, especially for $\epsilon = 0.3$). On the top row, parasites can be seen in dark color inside the cell. The prediction probabilities are indicated above each image.

2. Robustness of deep learning models

The robustness of deep learning algorithms needs to be evaluated before deploying them in real-wold. Therefore, it is crucial to ensure the neural networks can detect abnormal inputs in safety- and security-sensitive applications such as medical diagnosis, biometric authentication, intrusion detection, and autonomous driving (Emmott et al., 2016; Nitsch et al., 2021).

2.1. Robust out-of-distribution detection for neural networks

Out-of-Distribution (OOD) samples, the test samples that are not well covered by training data, is a major cause of poor performance in deep learning models. OOD samples are able to both evade the deep learning algorithms as well as achieve targeted misclassification with high confidence. There are currently many approaches that can detect OOD examples. They work well when tested on natural samples from a distribution that is sufficiently different from the distribution of the training data (Chen et al., 2020).

Hendrycks & Gimpel (2016) recently proposed a baseline for detecting misclassified and OOD examples in deep neural networks (DNNs), and Liang et al. (2017) improved it by processing the input and output of the DNNs. The Softmax Baseline Mode computes softmax probabilities with the fast-growing exponential function. Thus minor changes to the softmax inputs, can lead to major changes in the output distribution. A softmax baseline method uses probabilities from softmax distributions to predict whether a test example is from a different distribution from the training data or from within the same distribution. Liang et al. (2018) proposed ODIN (Out-of-DIstribution detector for Neural networks) which is a simple and effective method for detecting OOD images in neural networks. ODIN does not require re-training the neural network and is compatible with diverse network architectures and datasets.

2.2. Robust adversarial detection for neural networks

Recent studies have concentrated on identifying adversarial examples despite the inefficiency of adversarial defense (Fawzi et al., 2016). Goodfellow et al. (2014) suggested a framework for estimating abnormal samples via adversarial networks. Local Intrinsic Dimensionality (LID) is one of the successful adversarial detection techniques proposed by Ma et al. (2018). With the assumption that adversarial subspaces are low probability regions that are densely scattered in the high dimensional representation space of DNNs. The properties of adversarial regions is considered as a key requirement for adversarial defense (Ma et al., 2018).

3. Mahalanobis confidence score

One of the limitations of both ODIN and LID is that they are designed for either OOD or adversarial corruption but not for both. More recently, Lee et al. (2018) proposed a simple yet effective method for detecting both OOD samples and adversarial samples.

Mahalanobis distance-based confidence score is a class-conditional anomaly detection method, motivated by classification prediction confidence (Kamoi & Kobayashi, 2020).

Let us consider a dataset $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ with input-label pairs. The labels belong to one of the classes $\{1, \ldots, C\}$. For malara parasite detection, labels are either parasitized or healthy. For a deep neural network, f_ϕ, with parameters ϕ, we consider a pre-trained softmax classifier,

$$p_\theta(y = c|\mathbf{x}) = \frac{\exp(\mathbf{w}_c^\top f_\phi(\mathbf{x}))}{\sum_{c'} \exp(\mathbf{w}_{c'}^\top f_\phi(\mathbf{x}))}. \tag{1}$$

For each class, we define a multivariate Gaussian distribution, $p(f_\phi(\mathbf{x})|y = c) = \mathcal{N}(f_\phi(\mathbf{x})|\boldsymbol{\mu}_c, \Sigma)$ with class mean $\boldsymbol{\mu}_c$ and pooled-covariance Σ. This way, we compute the empirical statistics $(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \ldots, \hat{\boldsymbol{\mu}}_C, \hat{\Sigma})$ from the training dataset $\mathcal{D}$. With these statistics, the closest class $\tilde{c}$ to a query input $\mathbf{x}_*$ can be computed using the Mahalanobis distance,

$$\tilde{c} = \min_{c \in \{1,2,\ldots,C\}} \sqrt{(f(\mathbf{x}_*) - \hat{\boldsymbol{\mu}}_c)^\top \hat{\Sigma}^{-1} (f(\mathbf{x}_*) - \hat{\boldsymbol{\mu}}_c)}. \tag{2}$$

Following Liang et al. (2017), a controlled noise ϵ is added to the input,

$$\tilde{\mathbf{x}}_* = \mathbf{x}_* - \epsilon \cdot \text{sign}\big(\nabla_{\mathbf{x}} (f(\mathbf{x}_*) - \boldsymbol{\mu}_{\tilde{c}})^\top \hat{\Sigma}^{-1} (f(\mathbf{x}_*) - \hat{\boldsymbol{\mu}}_{\tilde{c}})\big), \tag{3}$$

for better calibration. Then, we can compute the confidence score,

$$M(\tilde{\mathbf{x}}_*) = \max_{c \in \{1,2,\ldots,C\}} -(f(\tilde{\mathbf{x}}_*) - \hat{\boldsymbol{\mu}}_c)^\top \hat{\Sigma}^{-1} (f(\tilde{\mathbf{x}}_*) - \hat{\boldsymbol{\mu}}_c). \tag{4}$$

By doing this for all $l = \{1, \ldots, L\}$ layers of the neural network with weights α_l of the logistic regression classifier (separately trained for each layer on a validation dataset (Ma et al., 2018)), we can compute the overall score $M^*(\mathbf{x}_*) = \sum_{l=1}^{L} \alpha_l M_l(\mathbf{x}_*)$. For a given threshold ρ, the query samples, $\mathbf{x}_*$, is *in-distribution*, if $M^*(\mathbf{x}_*) \geq \rho$.

4. Experiments

In our experiments, we used a publicly accessible and annotated malaria dataset of healthy and infected blood smear images [1]. It contains 13,779 parasitized and 13,779 uninfected cell images. We split the dataset into 60 : 10 : 30 for train, validation, and test datasets, respectively. We resized the images to 125 × 125 pixels and normalized them to assist in faster convergence. To prevent over-fitting and to account for possible variations in photomicroscopy, we have applied data augmentation techniques such as rotation, shearing, translation, and zooming. For OOD, another malaria dataset consisting of 22,046 was used[2].

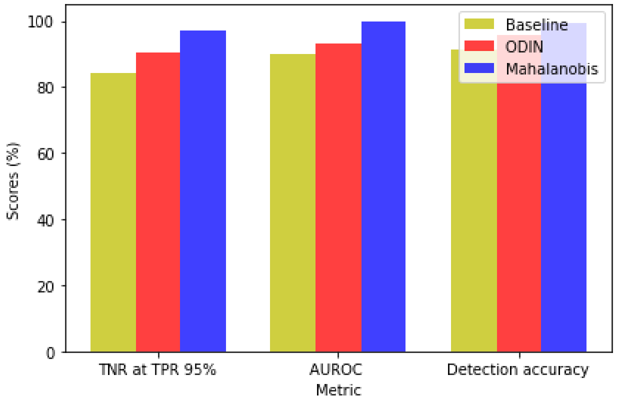

Figure 2. Robustness against *out-of-distribution* samples: ResNet-18

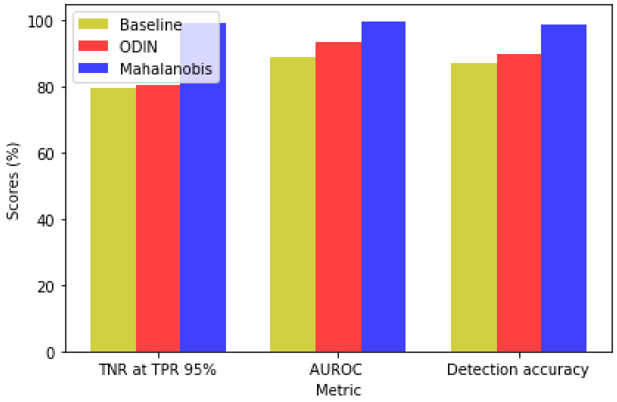

Figure 3. Robustness against *out-of-distribution* samples: VGG-19

For OOD and adversarial samples detection, the suggested method to improve the robustness of DL models was evaluated on both VGG-19 and ResNet-18 using a threshold-based detector. We evaluate the models with the following metrics: the true negative rate (TNR) at 95% true positive rate (TPR), the area under the receiver operating characteristic (AUROC) curve, the area under the precision-recall (AUPR) curve, and the detection accuracy. The Mahalanobis

[1] https://lhncbc.nlm.nih.gov/publication/pub9932

[2] https://github.com/shriyakabra97/malaria-parasite-detection

Table 1. Robustness against *adversarial samples*: a comparison of the performance of LID and Mahalanobis (proposed) towards detecting adversarial test samples generated from malaria image datasets.

Model	**Metric**	**LID**				**Mahalanobis**			
		FGSM	BIM	DeepFool	CW	FGSM	BIM	DeepFool	CW
VGG-19	TNR at TPR 95%	99.96	96.87	74.48	75.93	100.00	100.00	75.96	98.11
	AUROC	97.30	96.68	78.01	89.64	99.98	99.68	83.56	99.35
	Detection accuracy	99.41	90.46	46.05	72.96	99.98	99.99	61.95	97.51
ResNet-18	TNR at TPR 95%	96.84	95.61	63.59	73.09	99.99	97.98	76.22	98.90
	AUROC	97.30	96.68	78.01	89.64	99.98	99.68	83.56	99.35
	Detection accuracy	97.02	97.65	49.56	84.66	99.75	99.95	64.95	98.05

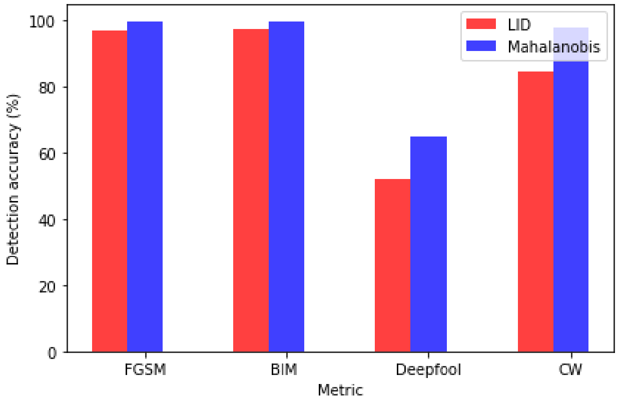

Figure 4. A comparison of performance between LID and Mahalanobis distance-based confidence score for ResNet-18 pre-trained model.

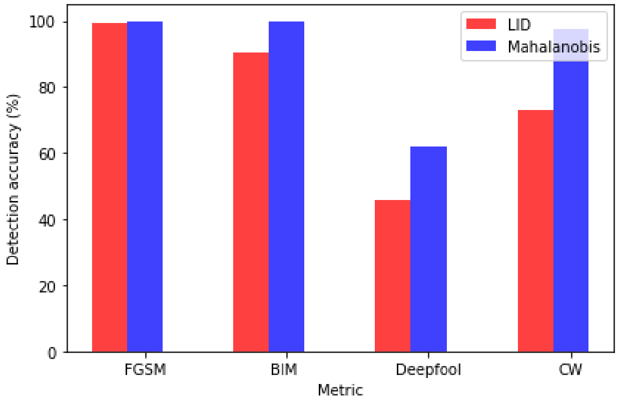

Figure 5. A comparison of detection performance between LID and Mahalanobis distance-based confidence score for VGG-19 pre-trained model

confidence score was compared with the baseline method and state-of-the-art ODIN for OOD samples. It was also compared with the state-of-the-art LID toward adversarial samples detection. Comparing with the baseline method, ODIN and LID, as shown in Table 1 and Figures 2, 3,4,and 5, the proposed approach outperforms on detecting abnormal samples.

As the Mahalanobis distance-based score method outperforms for the tasks of detecting OOD samples and adversarial samples, it can serve as a diagnostic framework for evaluating deep neural networks, as it is able to reveal their potentially non-obvious vulnerabilities and reliability. Such frameworks help to ensure that deep neural networks are effective, secure, and easy to deploy in a broad range of medical imaging applications beyond malaria detection. Our future work will extend this framework to test medical images under various lighting and other possible sources of corruption. We envision this, in the long-term, will enable low-cost, yet reliable, imaging procedures.

Broader impact statement

Broadly, our research is a step towards developing robust semi-automated medical image analysis techniques. Specifically, we focus on ensuring that malaria detection procedures are reliable enough before deployment. It, in the long-term, will help low-income countries to eradicate malaria. These systems, however, need to be rigorously validated before deploying in medical facilities.

References

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019.

Asgharnezhad, H., Shamsi, A., Alizadehsani, R., Khosravi, A., Nahavandi, S., Sani, Z. A., and Srinivasan, D. Objective evaluation of deep uncertainty predictions for covid-19 detection. *arXiv preprint arXiv:2012.11840*, 2020.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14, 2017.

Chen, J., Wu, X., Liang, Y., Jha, S., et al. Robust out-of-distribution detection in neural networks. *arXiv preprint arXiv:2003.09711*, 2020.

Cho, S. I., Sun, S., Mun, J.-H., Kim, C., Kim, S., Cho, S., Youn, S., Kim, H. C., and Chung, J. Dermatologist-level classification of malignant lip diseases using a deep convolutional neural network. *British Journal of Dermatology*, 182(6):1388–1394, 2020.

Dong, Y., Jiang, Z., Shen, H., David Pan, W., Williams, L. A., Reddy, V. V. B., Benjamin, W. H., and Bryan, A. W. Evaluations of deep convolutional neural networks for automatic identification of malaria infected cells. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pp. 101–104, 2017.

Emmott, A., Das, S., Dietterich, T., Fern, A., and Wong, W.-K. A meta-analysis of the anomaly detection problem, 2016.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Fawzi, A., Moosavi-Dezfooli, S.-M., and Frossard, P. Robustness of classifiers: from adversarial to random noise. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1632–1640. Curran Associates, Inc., 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *The International Conference on Learning Representations*, 2015.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Kamoi, R. and Kobayashi, K. Why is the mahalanobis distance effective for anomaly detection? *arXiv preprint arXiv:2003.00402*, 2020.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Laves, M.-H., Ihler, S., Fast, J. F., Kahrs, L. A., and Ortmaier, T. Well-calibrated regression uncertainty in medical imaging with deep learning. In *Medical Imaging with Deep Learning*, pp. 393–412. PMLR, 2020.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pp. 7167–7177, 2018.

Liang, S., Li, Y., and Srikant, R. Principled detection of out-of-distribution examples in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *The International Conference on Learning Representations*, 2018.

Liang, Z., Powell, A., Ersoy, I., Poostchi, M., Silamut, K., Palaniappan, K., Guo, P., Hossain, M. A., Antani, S. K., Maude, R. J., Huang, X., Jaeger, S., and Thoma, G. R. Cnn-based image analysis for malaria diagnosis. *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 493–496, 2016.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. In *The International Conference on Learning Representations*, 2018.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Nitsch, J., Itkina, M., Senanayake, R., Nieto, J., Schmidt, M., Siegwart, R., Kochenderfer, M. J., and Cadena, C. Out-of-distribution detection for automotive perception. In *24th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2021. doi: arXiv:2011.01413.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., and Thoma, G. R. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6:e4568, 2018a.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., and Thoma, G. R. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6:e4568, 2018b.

Ravendran, A., de Silva, K. R. T., and Senanayake, R. Moment invariant features for automatic identification of critical malaria parasites. In *2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS)*, pp. 474–479. IEEE, 2015.

Senanayake, R., O'Callaghan, S., and Ramos, F. Predicting spatio-temporal propagation of seasonal influenza using variational gaussian process regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Silva, A. D. N., Wijesundara, M. N., and Senanayake, R. Computer controlled digital microscope with photomicrograph enhancement. In *2013 International Conference of Information and Communication Technology (ICoICT)*, pp. 44–47. IEEE, 2013.

Wongsrichanalai, C., Barcus, M. J., Muth, S., Sutamihardja, A., and Wernsdorfer, W. H. A review of malaria diagnostic tools: microscopy and rapid diagnostic test (rdt). *The American journal of tropical medicine and hygiene*, 77 (6_Suppl):119–127, 2007.

World Health Organization. World malaria report 2019. https://www.who.int/news-room/fact-sheets/detail/malaria. Accessed: 2020-05-25.

Yang, F., Poostchi, M., Yu, H., Zhou, Z., Silamut, K., Yu, J., Maude, R. J., Jaeger, S., and Antani, S. Deep learning for smartphone-based malaria parasite detection in thick blood smears. *IEEE journal of biomedical and health informatics*, 24(5):1427–1438, 2019.