ISSN 1870-4069

Random Forest and Deep Learning Performance on the Malaria DREAM Sub Challenge One

Didier Barradas-Bautista

King Abdullah University of Science and Technology, Catalysis Center, Saudi Arabia

didier.barradasbautista@kaust.edu.sa

Abstract. In several countries, vector borne diseases play a significant role in the burden of public and individual health. Mosquitoes transmit diseases such as dengue and malaria. Malaria is a disease caused by *Plasmodium* parasites and transmitted by *Anopheline* mosquitoes species. In 2017, WHO estimates 200 million cases, affecting mainly children under five years old. On 2018 in Mexico, reported 799 Malaria cases showing a more similar trend for 2019 so far. Among the effort to eradicate malaria, a crowd-sourced event called DREAM challenge, where participants have to produce machine learning models and strategies on biological data. Also, the baseline truth is not released during the challenge. On this DREAM challenge, the participants can use any approach any design to build a predictor to predict the clearance of and concentration of artemisin (the standard antimalarial drug). On this work, I discuss differences in my strategy over sub challenge one in which my predictions ranked among the top three performers.

Keywords: DREAM challenge, ensemble methods, malaria.

1 Introduction

1.1 Malaria a World Wide Problem Almost Neglected

Malaria of "several diseases transmitted by insects, also known as Vector born diseases, with a considerable impact on human health [13]. It is caused by protozoans from the genus *Plasmodium* and transmitted by *Anopheles* mosquitoes. It has been more than 100 years since Ronald Ross discovered the *Plasmodium* parasites on Anopheles mosquitoes guts and understood the spreading mechanism of this disease [22]. But in Malaria is still an ongoing problem, in 2017, the World Health Organization (WHO) estimates 200 million cases, affecting mainly children under five years old [20]. In 2018 in Mexico reported 799 Malaria cases showing a similar trend for 2019 so far [23]. *P. falciparum* and *P. vivax* are the most common cause of Malaria in Africa and America, respectively.

Over the years, the global number of Malaria cases has dropped thanks to initiatives like "roll back malaria" [9, 18] or "The Malaria day in the Americas" [1].

Plasmodium protozoans have a very complicated life cycle where it changes several times the proteins in its surface to avoid recognition from the mosquito and human immune system [15]. The continue adaptation makes it difficult for the development of a highly effective vaccine or drug [16]. Moreover, the problem can become more complicated with the rise of resistant mosquitoes to insecticides and *Plasmodium* parasites resistant to drugs [25].

As part of the effort to fight Malaria, crowd-sourcing projects can help to develop new antimalarial drugs or to predict their activity [19]. Also, crowd-sourced challenges can help understand the mechanism of the underlying biology of Malaria, like the Malaria DREAM Challenge [11, 7].

1.2 Palliative Treatment of Malaria

The WHO Patients diagnosed with Malaria are preferably treated with artemisinin combination therapies [2]. Artemisinin is a drug that eliminates the majority of *Plasmodium* parasites, while the remaining parasites are eliminated by other partner drugs [8]. On the other hand, drug-resistant *Plasmodium* is a recurrent problem [5].

1.3 The Malaria DREAM Sub Challenge One Design

With the advent of the so-called OMIC techniques, a high throughput machine can process hundreds or thousands of samples producing big data over complex biological scenarios. In this situation, computational models and methods are the best tools to understand the underlying biological process on these. The DREAM challenge is an open science effort inspired by the crowd-sourced creation of such computational tools and analysis from the scientific community.

The general objective of the challenge was to predict the artemisinin resistance level. Thus the DREAM Malaria challenge was divided into two parts. The first challenge consisted of the concentration values for half of the inhibitory concentration of artemisinin (IC50). The second part was dedicated to predict the patient status of clearance from parasites post-treatment. In both cases, the data consisted of expression changes of all the genes of the *Plasmodium* parasite. Still, the first part used *in-vitro* to predict *in-vitro* expression; on the second part, the *in-vitro* data was used to predict *in vivo* response.

2 Methods

The organizer supplied the data, that consisted of expression data form the 5542 genes for the *Plasmodium* parasite and four additional descriptors. These four descriptors provided information about the time of recollection ("Timepoint"), type of treatment ("Treatment," describing the use of artemisinin), bio-repeats ("Biorep"), and the id of origin.

To preprocess the data, I used the pandas library [17]. I used two different approaches to build the models, first I used sci-kit learn [21, 24] for the random

forest model and hyperparametric search and second, I used Tensorflow [3] and Keras [6] to build the neural network model.

2.1 Data Cleaning

The data for this sub-challenge was complete since it does not have any empty values. I converted the categorical columns "Treatment," and "Timepoint" to dummy variables to manage this relevant information. I dropped the isolate, and Biorep features the training and testing data set. Isolate feature is an identification label that not informative for training since a different set of isolates constitute the testing data. The Biorep feature is another id label that is not included on the independent testing set; thus is not informative. The training data was used to calculate the mean and standard deviation to use for later scaling of itself and the testing data. I trained the models from this standardized data.

3 Result

On the DREAM challenges, participants should provide the source code for the final models used for predictions. This setup is a two-fold advantage for all the community since the idea is to share the ideas and secure the reproducibility of the code. Thus my code is available on the challenge page[1].

3.1 Random Forest Model

To obtain the best random forest model, I performed a random search of hyperparameters with five-fold cross-validation. A random search of parameters allows testing several combinations of decision trees with a minimal number of samples and splits. According to the search the best parameters for a random forest models are 6333 trees taking all the data to build each tree, with a maximum number of depth features set to 30, a minimum of ten samples to node split and a minimum of 8 samples to be considered a leaf. The produced model achieved 69.59% accuracy with 0.3899 degrees of Mean average error (MAE), and 0.2746 degrees of mean squared error (MSE), see Fig. 1A. The detailed script of the search is available on GitHub[2].

3.2 Neural Network Model

A fully connected neural network, with five layers with 640 neurons to deal with the several combinations of input, four layers with 64 neurons to condense the previous information and a final output layer.

I used the Rectified Linear Unit as an activation function and RMSprop with the default learning rate as the optimizer. This architecture had 5,243,329

[1] https://www.synapse.org/#!Synapse:syn20609261
[2] https://github.com/D-Barradas/random_Hyperparameter_Search-

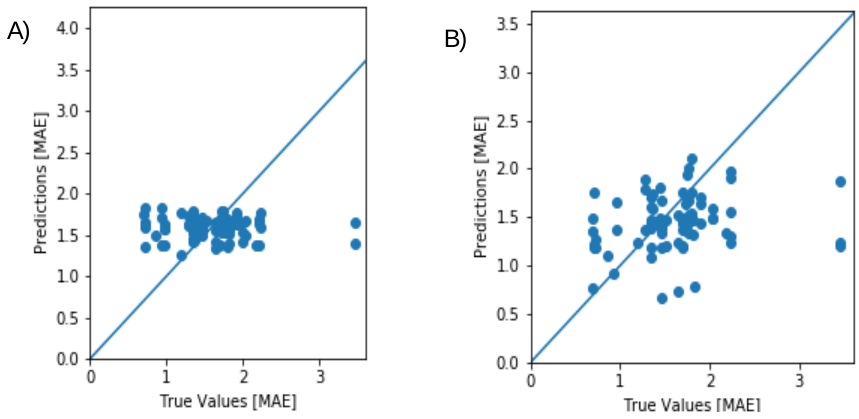

Fig. 1. Scatter plot of predicted vs true values. A) Resulting prediction distribution from the random forest algorithm. B) Resulting prediction distribution from the deep neural network.

of trainable parameters. The batch size set to 2 samples. I trained the model on the scaled data, and the model stopped around 20 epochs because there was an early stop clause for when the learning rate did not improve. The resulting model had an accuracy of 72.20% but with bigger MAE and MSE, 0.4143 and 0.349 respectively (Fig. 1B).

4 Discussion

The typical biological data set contains measurements of individual genes of organisms; in the case of the DREAM malaria Challenge, the data is the whole expressed genome of the *Plasmodium* parasite. This means the data reflected the actual genetic response of the parasite in the presence of a drug. Hence, I considered this problem as a multivariate data set. The underlying problem is, in fact, an interaction network, and searching for correlations among the 5542 genes is computationally a challenge. Understanding that is a network of interactions discards some approaches to predict the actual response, such as lineal regression or support vector machines. To the best of my knowledge, ensemble methods such as Decision trees and densely connected neural networks are the best algorithms to build predictions from complex feature relationships.

On that line, I searched for the best random forest predictive model. The resulting model had roughly one tree per gene where at least eight genes are needed to set the split for a new leaf of the decision tree, with a maximum depth of 30 related genes. This kind of arrangement of the decision trees is a consequence of the underlying network previously mentioned.

Thus, we could say that up to a combination of 30 genes are involved in some way to the response to artemisinin. Moreover, overlooking the feature

Table 1. Top 10 features ranked according the importance from the random forest model.

Description	Genes ID	Importance
Unknown function	PF3D7_1326200	0.004056
Unknown function	PF3D7_1217000	0.003661
Unknown function	PF3D7_0108200	0.003535
Sec1 family protein	PF3D7_1034000	0.003250
Unknown function	PF3D7_1307700	0.003152
O-fucosyltransferase 2	PF3D7_0909200	0.002717
Unknown function	PF3D7_0626200	0.002685
MORN repeat protein	PF3D7_1426400	0.002528
Unknown function	PF3D7_0502500	0.002503
Liver merozoite formation protein	PF3D7_0602300	0.002480

importance (Table 1), we can obtain a list of the genes and their functions. It is known that about half of the *Plasmodium falciparum* genome encodes conserved proteins of unknown function [4], so the most of the features used have a biological unknown function. However on Table 1 a few features have a function assigned that are related to membrane dynamics on a particular moment in the parasite life cicle. Sec1 proteins plays a role on membrane exocitosis [12], MORN repeat proteins are associated with cell budding and nuclear division [10], and O-fucosyltransferase 2 is required to the proper assembly of trafficking of proteins [14].

All these functions together tied together are indicators of the schizogony stage (asexual reproduction by fission) which is previous to the release of merozoites hence among the features we can find the "Liver merozoite formation protein" among the top ten features used on the random forest model. These genes could become new markers or drug targets.

Spite this apparent predictive power, the predictions with ensemble methods become constrained and somewhat limited by the number of genes that can be combined on each of the decision trees. A highly connected neural network could, in theory, overcome the problem. I trained a deep neural network with a funnel-like architecture, diminishing the number of neurons towards the output layer. The resulting architecture also had batches of two samples, generating a framework that searches for the best correlation on binary interactions.

Using deep learning comes at the cost of losing the precise descriptive power of decision trees. But on the other hand, we can monitor the training of the model through the use of quality metrics like MAE and MSE. The curves for training and validation showed a very early stoppage (before 20 epochs). For MAE (Fig. 2A) as for MSE (Fig. 2B), the validation error was slightly above the training error, indicating the possibility of over-fitting. Probably a similar training occurred for the best random forest model.

The random forest used the whole data set to build each tree, while the deep neural network was driven by all the combinations of the entire data-set on

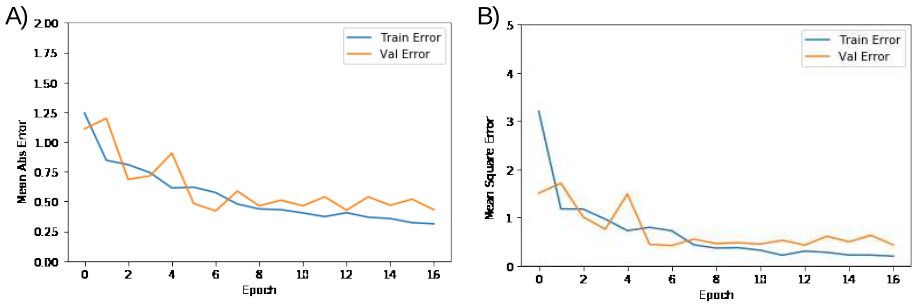

Fig. 2. History of the neural network training on the data. A)MAE values during the neural network training. B)MSE values during the neural network training.

the first layers. However, the prediction from the neural network has a broader range, reflected in the difference of MSE with the random forest. This wider range is also the reason for the 3.39% increase on the accuracy compared to the random forest.

5 Conclusion

Multivariate data requires an algorithm with complex decision-making schemes. The kind of model that can present this is ensemble methods and neural networks. I based my choice to use deep neural networks over the random forest model on the accuracy and the visual inspection of the prediction range. While using random forest explains the decision making, it requires thousands of trees for the prediction. Then deep learning is an alternative option that can fit multivariate data with the clear drawback of becoming unable to know the detail of the combinations made on each neuron. Overall the deep learning model made predictions close enough to be the third place on the sub-challenge. Further feature engineering and selection could improve both models. Reviewing the feature importance, we can discover new drug targets or assign functions that remained unknown for some genes so far.

Acknowledgment. The work is supported by KAUST Catalysis Center. I want to thank the KAUST Supercomputing Laboratory (KSL) for allowing me to use the resources available, especially the Shaheen and Ibex supercomputers.

References

1. Campeones de la malaria día del paludismo en las Américas

2. Guidelines for the treatment of malaria. Geneva (2006)
3. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., a nd Sherry Moore, R.M., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (2015)
4. Aurrecoechea, C., Barreto, A., Basenko, E.Y., Brestelli, J., Brunk, B.P., Cade, S., Crouch, K., Doherty, R., Falke, D., Fischer, S., Gajria, B., Harb, O.S., Heiges, M., Hertz-Fowler, C., Hu, S., Iodice, J., Kissinger, J.C., Lawrence, C., Li, W., Pinney, D.F., Pulman, J.A., Roos, D.S., Shanmugasundram, A., Silva-Franco, F., Steinbiss, S., Stoeckert, C.J., Spruill, D., Wang, H., Warrenfeltz, S., Zheng, J.: EuPathDB: the eukaryotic pathogen genomics database resource. Nucleic Acids Research 45(D1), D581–D591 (2017)
5. Barnes, K.I., White, N.J.: Population biology and antimalarial resistance: The transmission of antimalarial drug resistance in Plasmodium falciparum, vol. 94 (2005)
6. Chollet, F., others: Keras (2015), https://keras.io
7. Davis, S., Button-Simons, K., Bensellak, T., Ahsen, E.M., Checkley, L., Foster, G.J., Su, X., Moussa, A., Mapiye, D., Khoo, S.K., Nosten, F., Anderson, T.J.C., Vendrely, K., Bletz, J., Yu, T., Panji, S., Ghouila, A., Mulder, N., Norman, T., Kern, S., Meyer, P., Stolovitzky, G., Ferdig, M.T., Siwo, G.H.: Leveraging crowdsourcing to accelerate global health solutions. Nat. Biotechnol. 37(8), 848–850 (2019)
8. Dondorp, A.M., Nosten, F., Yi, P., Das, D., Phyo, A.P., Tarning, J., Lwin, K.M., Ariey, F., Hanpithakpong, W., Lee, S.J., Ringwald, P., Silamut, K., Imwong, M., Chotivanich, K., Lim, P., Herdman, T., An, S.S., Yeung, S., Singhasivanon, P., Day, N.P.J., Lindegardh, N., Socheat, D., White, N.J.: Artemisinin resistance in Plasmodium falciparum malaria. N. Engl. J. Med. 361(5), 455–467 (2009)
9. Feachem, S.R.: Roll Back Malaria: an historical footnote. Malaria Journal 17(1), 433 (2018)
10. Ferguson, D.J.P., Sahoo, N., Pinches, R.A., Bumstead, J.M., Tomley, F.M., Gubbels, M.J.: MORN1 Has a Conserved Role in Asexual and Sexual Development across the Apicomplexa. Eukaryotic Cell 7(4), 698–711 (2008)
11. Ghouila, A., Siwo, G.H., Entfellner, J.B.D., Panji, S., Button-Simons, K.A., Davis, S.Z., Fadlelmola, F.M., The DREAM of Malaria Hackathon Participants, Ferdig, M.T., Mulder, N.: Hackathons as a means of accelerating scientific discoveries and knowledge transfer. Genome Research 28(5), 759–765 (2018)
12. Halachmi, N., Lev, Z.: The Sec1 Family: A Novel Family of Proteins Involved in Synaptic Transmission and General Secretion. Journal of Neurochemistry 66(3), 889–897 (2002)
13. Institute of Medicine (US) Forum on Microbial Threats: Vector-Borne Diseases: Understanding the Environmental, Human Health, and Ecological Connections, Workshop Summary. The National Academies Collection: Reports funded by National Institutes of Health, National Academies Press (US), Washington (DC) (2008), http://www.ncbi.nlm.nih.gov/books/NBK52941/
14. Lopaticki, S., Yang, A.S.P., John, A., Scott, N.E., Lingford, J.P., O'Neill, M.T., Erickson, S.M., McKenzie, N.C., Jennison, C., Whitehead, L.W., Douglas, D.N., Kneteman, N.M., Goddard-Borger, E.D., Boddey, J.A.: Protein O-fucosylation

in Plasmodium falciparum ensures efficient infection of mosquito and vertebrate hosts. Nature Communications 8(1), 561 (2017)
15. Matuschewski, K.: Getting infectious: formation and maturation of Plasmodium sporozoites in the Anopheles vector. Cellular Microbiology 8(10), 1547–1556 (2006)
16. Matuschewski, K.: Vaccines against malaria-still a long way to go. The FEBS Journal 284(16), 2560–2568 (2017)
17. McKinney, W.: pandas: a Foundational Python Library for Data Analysis and Statistics
18. Nabarro, D.: Roll back malaria. Parassitologia 41(1-3), 501–504 (1999)
19. OpenWetWare: Opensourcemalaria:faq (2016), https://openwetware.org/
20. Organisation Mondiale de la Sante: World malaria report 2018 (2018)
21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python, vol. 12 (2011)
22. Rajakumar, K., Weisse, M.: The Centennial Year of Ronald Ross' Epic Discovery of Malaria Transmission: An Essay and Tribute, vol. 43 (1998)
23. Secretaria de Salud: Boletin epidemiologico
24. Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., Mueller, A.: Scikit-learn, vol. 19 (2015)
25. White, N.J.: Antimalarial drug resistance. Journal of Clinical Investigation 113(8), 1084–1092 (2004)