Machine learning model for predicting malaria using clinical information

You Won Lee, Jae Woo Choi, Eun-Hee Shin

PII: S0010-4825(20)30482-0

DOI: https://doi.org/10.1016/j.compbiomed.2020.104151

Reference: CBM 104151

To appear in: *Computers in Biology and Medicine*

Received Date: 22 September 2020

Revised Date: 9 November 2020

Accepted Date: 24 November 2020

Please cite this article as: Y.W. Lee, J.W. Choi, E.-H. Shin, Machine learning model for predicting malaria using clinical information, *Computers in Biology and Medicine*, https://doi.org/10.1016/j.compbiomed.2020.104151.

Machine learning model for predicting malaria using clinical information

You Won Lee[1] , Jae Woo Choi[2,3], Eun-Hee Shin[1,4,*]

[1]Department of Tropical Medicine and Parasitology, Seoul National University College of Medicine and Institute of Endemic Diseases, Seoul 03080, Republic of Korea

[2]Department of Pharmacology, Yonsei University College of Medicine, Seoul 03722, Republic of Korea

[3]Severance Biomedical Science Institute, Yonsei University College of Medicine, Seoul 03722, Republic of Korea

[4]Seoul National University Bundang Hospital, Seongnam 13620, Republic of Korea

***Corresponding author:** Eun-Hee Shin, PhD

Associate Professor, Department of Tropical Medicine and Parasitology, Seoul National University College of Medicine

103 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea

Tel.: +82-2-740-8344, E-mail: ehshin@snu.ac.kr

Abstract

Background: Rapid diagnosing is crucial for controlling malaria. Various studies have aimed at developing machine learning models to diagnose malaria using blood smear images; however, this approach has many limitations. This study developed a machine learning model for malaria diagnosis using patient information.

Methods: To construct datasets, we extracted patient information from the PubMed abstracts from 1956 to 2019. We used two datasets: a solely parasitic disease dataset and total dataset by adding information about other diseases. We compared six machine learning models: support vector machine, random forest (RF), multilayered perceptron, AdaBoost, gradient boosting (GB), and CatBoost. In addition, a synthetic minority oversampling technique (SMOTE) was employed to address the data imbalance problem.

Results: Concerning the solely parasitic disease dataset, RF was found to be the best model regardless of using SMOTE. Concerning the total dataset, GB was found to be the best. However, after applying SMOTE, RF performed the best. Considering the imbalanced data, nationality was found to be the most important feature in malaria prediction. In case of the balanced data with SMOTE, the most important feature was symptom.

Conclusions: The results demonstrated that machine learning techniques can be successfully applied to predict malaria using patient information.

Keywords: Machine learning, Malaria, Diagnosis, Case reports, Patient information

Introduction

Malaria is a dangerous infection disease caused by various species of *Plasmodium* worldwide, which can be cured using drugs [1]. The World Health Organization (WHO)'s World Malaria report 2019 indicated 228 million cases of malaria, with 40,500 deaths, in more than 90 countries in 2018 [2, 3]. Early diagnosis of malaria is very important as it allows performing appropriate disease management and treatment [4-6]. Therefore, various diagnosis methods of malaria have been proposed so far, such as polymerase chain reaction (PCR), rapid diagnostic tests (RDTs), and microscopy [7-9]. Frickmann et al. evaluated a PCR assay corresponding to the differentiation of plasmodium [10]. Amaral established ribosomal- and non-ribosomal-targeting PCR assays for detecting low-density and mixed malaria [11]. Makuuchi evaluated RDTs by comparing their results with those of microscopy analysis [12]. However, these methods are generally expensive in terms of time and expert labor. Recently, machine learning-based diagnoses have been investigated to increase the diagnosis speed [5, 6, 13].

Various studies have been conducted for diagnosing malaria using machine learning [6, 14], most of which focused on the blood smear image approach [14]. Evidently, blood smear microscopic examination is the most reliable clue in parasitic disease diagnoses [15]; moreover, machine learning-based diagnosis reduces the required costs and professional labor while increasing the diagnosis accuracy [1]. However, supervised learning requires establishing appropriate labeling of images by experts to construct trained datasets, performing the so-called annotation [13, 16]. Moreover, diagnoses using microscopy methods considerably depend on the skills and experience of experts [1, 15]. Therefore, these methods require greater specificity and sensitivity of an expert [13, 15].

Meanwhile, another important indicator that needs to be considered in malaria diagnosis is patient information, including symptomatology, nationality, age, gender, and travel history [17, 18]. However, it is difficult to discriminate malaria infection from other parasitic diseases [18], as usually, patients exhibit similar symptoms of malaria. Various effective methods for machine learning diagnosis have been developed using patient information. Spathis et al. considered age, gender, and symptomatology of a patient as variables for diagnosing chronic obstructive pulmonary disease [19]. Terrada et al. classified and predicted atherosclerosis using a machine learning approach, which trained the model on data including age, gender, and symptoms [20]. Mello-Roman et al. predicted dengue using data on age, gender, region, and symptomatology of a patient [21]. However, no study, so far, has attempted to diagnose malaria using machine learning models trained on patient information.

Therefore, this paper proposes a machine learning model to predict malaria by using patient information obtained from parasite case reports. We extract the data on nationality, disease, gender, age, symptoms and body region of patients with symptoms. Then, we train six machine learning models on these data.

Methods

Dataset

Using BioPython, we obtained the data corresponding to 56 parasitic disease reports provided by the Center for Disease Control and Prevention (CDC) [22] and abstracts of case reports of non-parasitic diseases (cancer, Alzheimer, rheumatoid disease, and diabetes) published from 1956 to 2019 by PubMed [23]. Based on CDC parasitic disease list, we classified 56 diseases based on causative parasite genus or if the disease name was the same. For example, Hydatid disease, Alveolar Echinococcosis, and Echinococosis were caused by the same genus, i.e., *Echinocococcus*, and categorized as the same parasitic disease. Filaria, Filariasis, Elephantiasis, and the infection of *Wuchereria bancrofti*, Brugia genus are regarded as the same parasitic disease. Nonpathogenic intestinal protozoa (*Enteromonas hominis, Retortamonas intestinalis*, and *Pentatrichomonas hominis*) can be present in feces but they are not harmful and nonpathogenic. Therefore, we combined them into one disease. Using this method, we were able to reorganize 56 parasitic diseases.

We selected non-parasitic diseases using the following standard. i) We chose diseases that constitute the top ten causes of deaths worldwide [24] or diseases with more than 10000 cases from 1956 to 2019. This approach was used because we wanted to have a diverse sample of patients. ii) We chose a disease with a name without an organ name when collecting case reports. For heart or brain diseases, if the name of the organ was included, it could overlap with the case report of parasites. iii) We excluded infections because symptoms of infections were similar to those of parasite infection diseases. Thus, we wanted to confirm the applicability of our model in non-parasitic disease patients who had other symptoms. The diseases that we used met these standards.

To extract relevant data, we ran queries using logical combinations such as operators “AND” and “OR.” We added our queries in Appendix A Supplementary Table 1. Parasitic diseases have many related names. Therefore, we use “OR” and “AND”. For example, another name for sleeping sickness is African trypanosomiasis, and the causative parasite is Trypanosoma genus. In addition, there are two types of trypanosomiasis: American trypanosomiasis (Chagas disease) and African trypanosomiasis. We are only interested in African trypanosomiasis and not American trypanosomiasis. Then, we used the following query: *((Trypanosoma OR Sleeping Sickness OR trypanosomiasis) AND Africa) AND case report.*

If we are interested in American trypanosomiasis, then we used the following query: *((Trypanosoma OR Chagas Disease OR trypanosomiasis) AND America) AND case report.*

We derived information regarding nation (meaning nationality or travel region of a patient), disease, gender, age, symptom and body region of patients with symptoms using Python scripts. The lists of body regions and symptoms were prepared by referring to the 10th edition of International Classification of Disease.

Dataset preprocessing

Fig. 1 shows the data processing scheme. We removed missing variables or values from the dataset, except for symptoms and body regions. If the symptoms or body regions had at least one value, we did not remove these data. First, we constructed a dataset comprising only parasitic disease patient information, and then, prepared the total dataset by adding information about other diseases (Alzheimer, rheumatoid, cancer, and diabetes). All the values were categorized using integers. Note that the prepared datasets incurred the data imbalance problem. To address this problem, we applied the synthetic minority oversampling technique (SMOTE) [25] provided by Scikit-learn [26].

Model development

We used various machine learning techniques to develop six models to diagnose malaria: support vector machine (SVM) [27], random forest (RF) [28], multilayered perceptron (MLP)

[29], AdaBoost (Ada) [30], gradient boosting (GB) [31], and CatBoost (CB) [32].

SVM

SVM is a widely used supervised learning approach for classification or regression analysis. It can be applied to transform training data into a high-dimensional feature space and determine a linear optimal solution by separating a hyperplane that provides the smallest distance between the hyperplane points and the largest margin between the classes [27, 33-35].

RF

RF is an ensemble supervised learning method composed of multiple decision trees corresponding to various subdatasets. Each tree calculates the results and obtains the average of the prediction outcomes. This approach allows reducing variance in decision trees [28, 36, 37].

MLP

MLP is a supervised machine learning algorithm used for data classification tasks. It is composed of three layers: an input layer, which includes input data; a hidden layer, which computes complicated associations across the network; and an output layer, which generates the final result. This process can be terminated when the error rate becomes sufficiently small. We optimized the log-loss function using the stochastic gradient descent [29, 38].

Ada

Ada is an ensemble learning algorithm used to elevate a weak classifier to a strong one. First, it trains a base classifier and assigns higher weights to the misclassified samples; thereafter, it is applied to the next process. This iterative process continues until a stop condition is reached or the error rate becomes sufficiently small [30, 39, 40].

GB

GB is an ensemble model based on decision trees. It minimizes the residual (negative gradient) using gradient descents to classify data [31, 36, 41].

CB

CB is a modification of GB, and yields high performance in case of categorical features [32, 42].

These models were trained to execute effective malaria diagnosis.

Model evaluation

A 10-fold cross-validation (CV) was applied to avoid overfitting. In 10-fold CV, the data are first randomly split into ten parts. Then, each subset is considered as the testing one and the remaining subsets are used for training. The results of 10-fold CV are the averaged values of accuracy obtained from the ten tests. We evaluated each model in terms of accuracy, precision, recall, and F1-score. These parameters are defined as follows:

$$Accuracy = \frac{True\ positives + True\ neagtives}{True\ positives + True\ negatives + False\ positives + False\ negatives}$$

$$Precision = \frac{True\ positives}{True\ positives\ +\ False\ positives}$$

$$Recall = \frac{True\ positives}{True\ positives + False\ negatives}$$

$$F1-score=\frac{2*Precision*Recall}{Precision+Recall}$$

In addition, an AUC curve was drawn to compare the performance of each model.

Feature importance

We analyzed the feature importance using RF with optimized hyperparameters to evaluate how each model works using yellowbrick [43]. The result is calculated based on the average of the feature importance associated with each feature in terms of achieving high performance for the model.

Results

Data statistics

In this study, information about 1,846 patients, obtained from case reports, was considered (Table 1). Overall, 1,698 patients had parasitic diseases, 135 patients had malaria, and 148 patients had non-parasitic diseases. In the total dataset, the portion of malaria patients was 7.31% and that of the solely parasitic disease dataset was 8%. Table 1 provides information about the patient demographics.

Model performance

Tables 2 and 3 and Fig. 2 describe the performance of the considered predictive models.

Concerning the solely parasitic disease dataset, the RF model achieved the best performance with AUC of 73.2%. The worst model was Ada, with AUC of 59.6%. After applying SMOTE, the AUC values of almost all models increased, except GB. In this case, RF achieved the best performance (with AUC of 73.5%), while Ada demonstrated the worst performance (with AUC of 68.8%). Within the total dataset, GB achieved the highest AUC (85.6%). The performance of the models trained on the total dataset was higher compared to those trained on the solely parasitic disease dataset. The values of accuracy, precision, recall, and F1-score were also higher in the case of training on the total dataset. The model that showed the worst results was SVM, with AUC of 77.6%. The AUC values of all models with SMOTE were decreased; however, the values of accuracy, precision, recall, and F1-score were higher. RF achieved the best performance among all classifiers (with AUC of 80.5%), while the worst performing model was MLP (with AUC of 67.4%).

Feature importance

We calculated the feature importance using RF, which achieved the highest performance overall (Fig. 3). Both in the total dataset and the solely parasitic disease dataset, the most important feature was “nation,” followed by “age” (Fig. 3A and C). However, in case of the dataset with SMOTE, “symptom” was the most important feature, followed by “nation” (Fig. 3B and D).

Discussion

Recently, an increasing number of studies have been conducted on malaria diagnosis using artificial intelligence (AI). Kim et al. and Wang et al. predicted malaria incidence by using a seasonal climate dataset [44, 45]. Moreover, the methods based on AI for diagnosis using blood smear images have been extensively investigated [1, 4, 6, 14]. Rajaraman et al. used thin-blood smear images to construct deep neural ensemble models [4]. Molina et al. introduced a machine learning model to discriminate infected blood cells from normal ones [6].

In the present study, we used the parasitic disease patient information derived from the abstracts of case reports provided by PubMed to train the models. Evidently, it is possible to consider various databases for obtaining the epidemiology or symptom data on parasitic diseases, such as Gideon [46] and CDC [22]. Moreover, the National Health and Nutrition Examination Survey was used as a source for obtaining health and nutrition information about patients. However, no database on parasitic disease patients provides information about patients' nationality, age, symptoms, and gender. Even if such information was available, it would not exhibit diversity in terms of regions or conditions of patients [47-49]. Therefore, we constructed datasets based on the information obtained from the abstracts of all parasitic disease case reports available in PubMed that were published from 1956 to 2019. Note that the abstracts did not provide detailed information about patients; however, the available data were sufficient to perform analysis to diagnose malaria using the methods considered in the present study. Moreover, these data reflected the trend of overall parasitic disease patient information with sufficient accuracy.

The performance estimates of almost all models trained on the solely parasitic disease dataset were lower than those of the models trained on the total dataset and those trained on the data with SMOTE. The observed results could be explained not only by a smaller dataset but also by the characteristics of parasitic diseases. In clinical cases, the symptoms of a parasitic disease are similar to those of malaria, and therefore, it was more difficult to discriminate malaria using the solely parasitic disease dataset. For example, fever, which is a standard symptom of malaria, can also indicate conditions such as toxoplasmosis [50] and pulmonary eosinophilia [51]. Similarly, abdominal pain is a generic symptom for conditions such as amoebic liver abscesses [52] and trichinellosis [53]. We hypothesized that SMOTE can address this problem through oversampling; however, the AUC values were still lower than those of the models trained on the total dataset.

According to the obtained results, RF achieved the best performance, except for the total dataset without SMOTE. The remarkable performance of an RF model has also been reported by other studies concerning various diseases [54-56]. An RF model has also been applied to neuroimaging classification [54], the prediction of in-hospital cardiac arrest [56], and biomarker prediction based on gene expression data [55]; in all these applications, the RF model demonstrated great performance.

Meanwhile, the results of feature importance analysis indicated that nationality and age are important factors to consider in diagnosing malaria using imbalanced data. Many previous studies have reported that parasitic diseases, such as malaria, depend on the places in which patients live or travel to [2, 57]. The obtained results suggested that nationality and the region of traveling are crucial factors in the diagnosis of parasitic diseases.

Limitations

The limitations of our study are related to the small size of the datasets used and the limited number of features without the process of feature selection. Moreover, the observed values of precision, recall, and F1-score were lower than those reported previously. Specifically, the dataset has high imbalance between parasitic and non-parasitic cases. We collected data from more than 35519 non-parasitic patients, expecting to be able to obtain more than 1698 parasitic patients. However, the number of samples inevitably decreased when extracting only patients whose information, such as the country, age, gender, symptoms, and body region of patients with symptoms, was available. In particular, cancer, rheumatism, diabetes, and Alzheimer’s patients were able to provide less nationality information in case reports. Therefore, we had no choice but to create a dataset with a small number of patients having non-parasitic diseases. We considered that they could be improved by applying SMOTE. However, even after the application of SMOTE, the value of precision, recall and F1-score did not exceed 0.5. This can be addressed by increasing the number of patients and features in the datasets.

Conclusions

This is the first study that aims to diagnose malaria using patient information. The novelty of the utilized datasets lies in the fact that the data were obtained for parasitic disease patients spread globally. We compared several machine learning models applied to malaria prediction trained on parasitic disease patient data. The results showed that RF was the best model for the diagnosis, indicating the possibility of diagnosing using only patient information with AI.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

EHS provided the research idea. YWL and EHS conceived and designed the study. YWL and JWC collected and analyzed data. YWL and JWC contributed materials and analysis tools. YWL and EHS wrote the paper. EHS was responsible for the overall project administration and acquiring of financial support.

References

[1] M. Poostchi, K. Silamut, R.J. Maude, S. Jaeger, G. Thoma, Image analysis and machine learning for detecting malaria, Transl Res, 194 (2018) 36-55 https://doi.org/10.1016/j.trsl.2017.12.004.

[2] L. Zekar, T. Sharman, Malaria (Plasmodium Falciparum), StatPearls, Treasure Island (FL), 2020.

[3] W.H. Organization, World malaria report 2019, (2019).

[4] S. Rajaraman, S. Jaeger, S.K. Antani, Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images, PeerJ, 7 (2019) e6977 https://doi.org/10.7717/peerj.6977.

[5] K. Torres, C.M. Bachman, C.B. Delahunt, J. Alarcon Baldeon, F. Alava, D. Gamboa Vilela, S. Proux, C. Mehanian, S.K. McGuire, C.M. Thompson, T. Ostbye, L. Hu, M.S. Jaiswal, V.M. Hunt, D. Bell, Automated microscopy for routine malaria diagnosis: a field comparison on Giemsa-stained blood films in Peru, Malar J, 17 (2018) 339 https://doi.org/10.1186/s12936-018-2493-0.

[6] A. Molina, S. Alferez, L. Boldu, A. Acevedo, J. Rodellar, A. Merino, Sequential classification system for recognition of malaria infection using peripheral blood cell images, J Clin Pathol, (2020) https://doi.org/10.1136/jclinpath-2019-206419.

[7] Z. Zheng, Z. Cheng, Advances in Molecular Diagnosis of Malaria, Adv Clin Chem, 80 (2017) 155-192 https://doi.org/10.1016/bs.acc.2016.11.006.

[8] P. Berzosa, A. de Lucio, M. Romay-Barja, Z. Herrador, V. Gonzalez, L. Garcia, A. Fernandez-Martinez, M. Santana-Morales, P. Ncogo, B. Valladares, M. Riloha, A. Benito, Comparison of three diagnostic methods (microscopy, RDT, and PCR) for the detection of malaria parasites in representative samples from Equatorial Guinea, Malar J, 17 (2018) 333 https://doi.org/10.1186/s12936-018-2481-4.

[9] K.O. Mfuh, O.A. Achonduh-Atijegbe, O.N. Bekindaka, L.F. Esemu, C.D. Mbakop, K. Gandhi, R.G.F. Leke, D.W. Taylor, V.R. Nerurkar, A comparison of thick-film microscopy, rapid diagnostic test, and polymerase chain reaction for accurate diagnosis of Plasmodium falciparum malaria, Malar J, 18 (2019) 73 https://doi.org/10.1186/s12936-019-2711-4.

[10] H. Frickmann, C. Wegner, S. Ruben, C. Behrens, H. Kollenda, R. Hinz, S. Rojak, N.G. Schwarz, R.M. Hagen, E. Tannich, Evaluation of the multiplex real-time PCR assays RealStar malaria S&T PCR kit 1.0 and FTD malaria differentiation for the differentiation of

Plasmodium species in clinical samples, Travel Med Infect Dis, 31 (2019) 101442 https://doi.org/10.1016/j.tmaid.2019.06.013.

[11] L.C. Amaral, D.R. Robortella, L.F.F. Guimaraes, J.E. Limongi, C.J.F. Fontes, D.B. Pereira, C.F.A. de Brito, F.S. Kano, T.N. de Sousa, L.H. Carvalho, Ribosomal and non-ribosomal PCR targets for the detection of low-density and mixed malaria infections, Malar J, 18 (2019) 154 https://doi.org/10.1186/s12936-019-2781-3.

[12] R. Makuuchi, S. Jere, N. Hasejima, T. Chigeda, J. Gausi, The correlation between malaria RDT (Paracheck pf.(R)) faint test bands and microscopy in the diagnosis of malaria in Malawi, BMC Infect Dis, 17 (2017) 317 https://doi.org/10.1186/s12879-017-2413-x.

[13] A. Rehman, N. Abbas, T. Saba, Z. Mehmood, T. Mahmood, K.T. Ahmed, Microscopic malaria parasitemia diagnosis and grading on benchmark datasets, Microsc Res Tech, 81 (2018) 1042-1058 https://doi.org/10.1002/jemt.23071.

[14] S. Rajaraman, S.K. Antani, M. Poostchi, K. Silamut, M.A. Hossain, R.J. Maude, S. Jaeger, G.R. Thoma, Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images, PeerJ, 6 (2018) e4568 https://doi.org/10.7717/peerj.4568.

[15] A. Mbanefo, N. Kumar, Evaluation of Malaria Diagnostic Methods as a Key for Successful Control and Elimination Programs, Trop Med Infect Dis, 5 (2020) https://doi.org/10.3390/tropicalmed5020102.

[16] K. Smith, F. Piccinini, T. Balassa, K. Koos, T. Danka, H. Azizpour, P. Horvath, Phenotypic Image Analysis Software Tools for Exploring and Understanding Big Image Data from Cell-Based Assays, Cell Syst, 6 (2018) 636-653

https://doi.org/10.1016/j.cels.2018.06.001.

[17] F. Jimenez-Morillas, M. Gil-Mosquera, E.J. Garcia-Lamberechts, I.-S. en representacion de la seccion de enfermedades tropicales de, I.-S. Seccion de enfermedades tropicales de, Fever in travellers returning from the tropics, Med Clin (Barc), 153 (2019) 205-212 https://doi.org/10.1016/j.medcli.2019.03.017.

[18] C. JY, Seo and Lee’s Clinical Parasitology, Seoul National University Publishing Council2011.

[19] D. Spathis, P. Vlamos, Diagnosing asthma and chronic obstructive pulmonary disease with machine learning, Health Informatics J, 25 (2019) 811-827 https://doi.org/10.1177/1460458217723169.

[20] O. Terrada, B. Cherradi, A. Raihani, O. Bouattane, Classification and Prediction of atherosclerosis diseases using machine learning algorithms, 2019 5th International Conference on Optimization and Applications (ICOA), IEEE, 2019, pp. 1-5.

[21] J.D. Mello-Roman, J.C. Mello-Roman, S. Gomez-Guerrero, M. Garcia-Torres, Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay, Comput Math Methods Med, 2019 (2019) 7307803 https://doi.org/10.1155/2019/7307803.

[22] C.f.D.C.a. Prevention, DPDx - Laboratory Identification of Parasites of Public Health Concern 2020.

[23] P.J. Cock, T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M.J. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics, Bioinformatics, 25 (2009) 1422-1423 https://doi.org/10.1093/bioinformatics/btp163.

[24] J.S. Rana, S.S. Khan, D.M. Lloyd-Jones, S. Sidney, Changes in Mortality in Top 10 Causes of Death from 2011 to 2018, J Gen Intern Med, (2020) https://doi.org/10.1007/s11606-020-06070-z.

[25] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P.J.J.o.a.i.r. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, 16 (2002) 321-357.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V.J.t.J.o.m.L.r. Dubourg, Scikit-learn: Machine learning in Python, 12 (2011) 2825-2830.

[27] C. Cortes, V.J.M.l. Vapnik, Support-vector networks, 20 (1995) 273-297.

[28] L.J.M.l. Breiman, Random forests, 45 (2001) 5-32.

[29] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010, pp. 249-256.

[30] T. Hastie, S. Rosset, J. Zhu, H.J.S. Zou, i. Interface, Multi-class adaboost, 2 (2009) 349-360.

[31] J.H.J.A.o.s. Friedman, Greedy function approximation: a gradient boosting machine, (2001) 1189-1232.

[32] A.V. Dorogush, V. Ershov, A.J.a.p.a. Gulin, CatBoost: gradient boosting with categorical features support, (2018).

[33] A. Gupta, B. Kahali, Machine learning-based cognitive impairment classification with optimal combination of neuropsychological tests, Alzheimers Dement (N Y), 6 (2020)

e12049 https://doi.org/10.1002/trc2.12049.

[34] N. Liu, R. Zhao, L. Qiao, Y. Zhang, M. Li, H. Sun, Z. Xing, X. Wang, Growth Stages Classification of Potato Crop Based on Analysis of Spectral Response and Variables Optimization, Sensors (Basel), 20 (2020) https://doi.org/10.3390/s20143995.

[35] A. Gupta, R. Katarya, Social media based surveillance systems for healthcare using machine learning: A systematic review, J Biomed Inform, 108 (2020) 103500 https://doi.org/10.1016/j.jbi.2020.103500.

[36] A. Dinh, S. Miertschin, A. Young, S.D. Mohanty, A data-driven approach to predicting diabetes and cardiovascular disease with machine learning, BMC Med Inform Decis Mak, 19 (2019) 211 https://doi.org/10.1186/s12911-019-0918-5.

[37] C. Wang, X. Chen, L. Du, Q. Zhan, T. Yang, Z. Fang, Comparison of machine learning algorithms for the identification of acute exacerbations in chronic obstructive pulmonary disease, Comput Methods Programs Biomed, 188 (2020) 105267 https://doi.org/10.1016/j.cmpb.2019.105267.

[38] A.M. Ahmed, S.F. Aly, Egyptian License Plates Recognition System Using Morphologial Operations and Multi Layered Perceptron, ICT in our lives-2019, 2019.

[39] B.X. Tran, G.H. Ha, L.H. Nguyen, G.T. Vu, M.T. Hoang, H.T. Le, C.A. Latkin, C.S.H. Ho, R.C.M. Ho, Studies of Novel Coronavirus Disease 19 (COVID-19) Pandemic: A Global Analysis of Literature, Int J Environ Res Public Health, 17 (2020) https://doi.org/10.3390/ijerph17114095.

[40] L. Liu, C. Zhang, G. Zhang, Y. Gao, J. Luo, W. Zhang, Y. Li, Y. Mu, A study of aortic dissection screening method based on multiple machine learning models, J Thorac Dis, 12

(2020) 605-614 https://doi.org/10.21037/jtd.2019.12.119.

[41] Y. Ye, Y. Xiong, Q. Zhou, J. Wu, X. Li, X. Xiao, Comparison of Machine Learning Methods and Conventional Logistic Regressions for Predicting Gestational Diabetes Using Routine Clinical Data: A Retrospective Cohort Study, J Diabetes Res, 2020 (2020) 4168340 https://doi.org/10.1155/2020/4168340.

[42] A. Gupta, A.S.R. Potty, D. Ganta, R.J. Mistovich, S. Penna, C. Cady, A.G. Potty, Streamlining the KOOS Activities of Daily Living Subscale Using Machine Learning, Orthop J Sports Med, 8 (2020) 2325967120910447 https://doi.org/10.1177/2325967120910447.

[43] B. Bengfort, R.J.J.o.O.S.S. Bilbro, Yellowbrick: Visualizing the scikit-learn model selection process, 4 (2019) 1075.

[44] Y. Kim, J.V. Ratnam, T. Doi, Y. Morioka, S. Behera, A. Tsuzuki, N. Minakawa, N. Sweijd, P. Kruger, R. Maharaj, C.C. Imai, C.F.S. Ng, Y. Chung, M. Hashizume, Malaria predictions based on seasonal climate forecasts in South Africa: A time series distributed lag nonlinear model, Sci Rep, 9 (2019) 17882 https://doi.org/10.1038/s41598-019-53838-3.

[45] M. Wang, H. Wang, J. Wang, H. Liu, R. Lu, T. Duan, X. Gong, S. Feng, Y. Liu, Z. Cui, C. Li, J. Ma, A novel model for malaria prediction based on ensemble algorithms, PLoS One, 14 (2019) e0226910 https://doi.org/10.1371/journal.pone.0226910.

[46] G. Informatics, GIDEON, 2020.

[47] S. Mahmoudi, S. Mamishi, M. Banar, B. Pourakbari, H. Keshavarz, Epidemiology of echinococcosis in Iran: a systematic review and meta-analysis, BMC Infect Dis, 19 (2019) 929 https://doi.org/10.1186/s12879-019-4458-5.

[48] M. Kotepui, K.U. Kotepui, Prevalence and laboratory analysis of malaria and dengue co-

infection: a systematic review and meta-analysis, BMC Public Health, 19 (2019) 1148 https://doi.org/10.1186/s12889-019-7488-4.

[49] D. Pierce, L. Merone, C. Lewis, T. Rahman, J. Croese, A. Loukas, M. McDonald, P. Giacomin, R. McDermott, Safety and tolerability of experimental hookworm infection in humans with metabolic disease: study protocol for a phase 1b randomised controlled clinical trial, BMC Endocr Disord, 19 (2019) 136 https://doi.org/10.1186/s12902-019-0461-5.

[50] A.S. Kota, N. Shabbir, Congenital Toxoplasmosis, StatPearls, Treasure Island (FL), 2020.

[51] S.K. Jha, B. Karna, K. Mahajan, Tropical Pulmonary Eosinophilia, StatPearls, Treasure Island (FL), 2020.

[52] T. Tharmaratnam, T. Kumanan, M.A. Iskandar, K. D'Urzo, P. Gopee-Ramanan, M. Loganathan, T. Tabobondung, T.A. Tabobondung, S. Sivagurunathan, M. Patel, I. Tobbia, Entamoeba histolytica and amoebic liver abscess in northern Sri Lanka: a public health problem, Trop Med Health, 48 (2020) 2 https://doi.org/10.1186/s41182-020-0193-2.

[53] P. Rawla, S. Sharma, Trichinella Spiralis (Trichnellosis), StatPearls, Treasure Island (FL), 2020.

[54] S.I. Dimitriadis, D. Liparas, I. Alzheimer's Disease Neuroimaging, How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: from Alzheimer's disease neuroimaging initiative (ADNI) database, Neural Regen Res, 13 (2018) 962-970 https://doi.org/10.4103/1673-5374.233433.

[55] L. Guo, Z. Wang, Y. Du, J. Mao, J. Zhang, Z. Yu, J. Guo, J. Zhao, H. Zhou, H. Wang, Y. Gu, Y. Li, Random-forest algorithm based biomarkers in predicting prognosis in the patients with hepatocellular carcinoma, Cancer Cell Int, 20 (2020) 251

https://doi.org/10.1186/s12935-020-01274-z.

[56] R. Ueno, L. Xu, W. Uegami, H. Matsui, J. Okui, H. Hayashi, T. Miyajima, Y. Hayashi, D. Pilcher, D. Jones, Value of laboratory results in addition to vital signs in a machine learning algorithm to predict in-hospital cardiac arrest: A single-center retrospective cohort study, PLoS One, 15 (2020) e0235835 https://doi.org/10.1371/journal.pone.0235835.

[57] F. Jimenez-Morillas, M. Gil-Mosquera, E.J. Garcia-Lamberechts, I.-S.t.d. department, Fever in travellers returning from the tropics, Med Clin (Engl Ed), 153 (2019) 205-212 https://doi.org/10.1016/j.medcle.2019.03.013.

Figure & table legends

Table 1. Dataset review

Table 2. Solely parasite dataset

Table 3. Total dataset

Fig 1. Data processing

Fig 2. AUC curve: A) the solely parasitic disease dataset; B) solely parasitic disease dataset with SMOTE; C) total dataset; D) total dataset with SMOTE.

Fig 3. Feature importance: A) the solely parasitic disease dataset; B) solely parasitic disease dataset with SMOTE; C) total dataset; D) total dataset with SMOTE.

Appendix A. Supplementary Table 1. Query list of parasitic diseases

Fig 1. Data processing

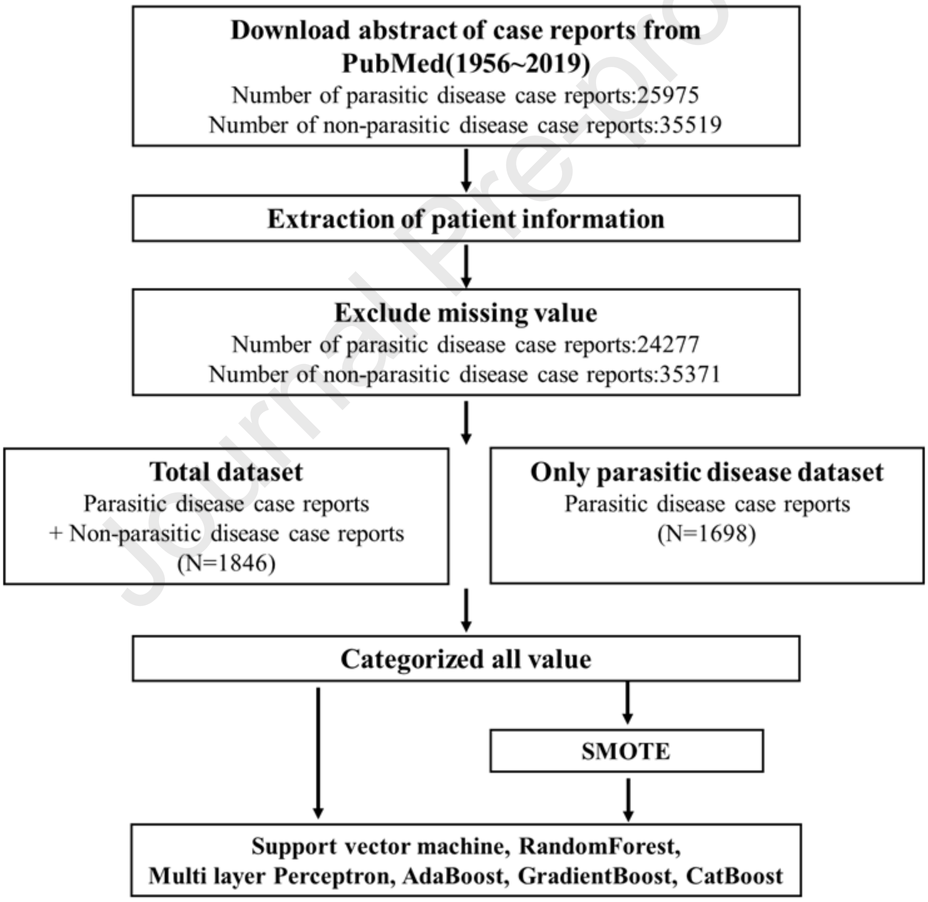
Download abstract of case reports from PubMed(1956~2019)
Number of parasitic disease case reports:25975
Number of non-parasitic disease case reports:35519
Extraction of patient information
Exclude missing value
Number of parasitic disease case reports:24277
Number of non-parasitic disease case reports:35371
Total dataset
Parasitic disease case reports
+ Non-parasitic disease case reports
(N=1846)
Only parasitic disease dataset
Parasitic disease case reports
(N=1698)
Categorized all value
SMOTE
Support vector machine, RandomForest,
Multi layer Perceptron, AdaBoost, GradientBoost, CatBoost

Fig 2. AUC curve: A) the solely parasitic disease dataset; B) solely parasitic disease dataset with SMOTE; C) total dataset; D) total dataset with SMOTE.

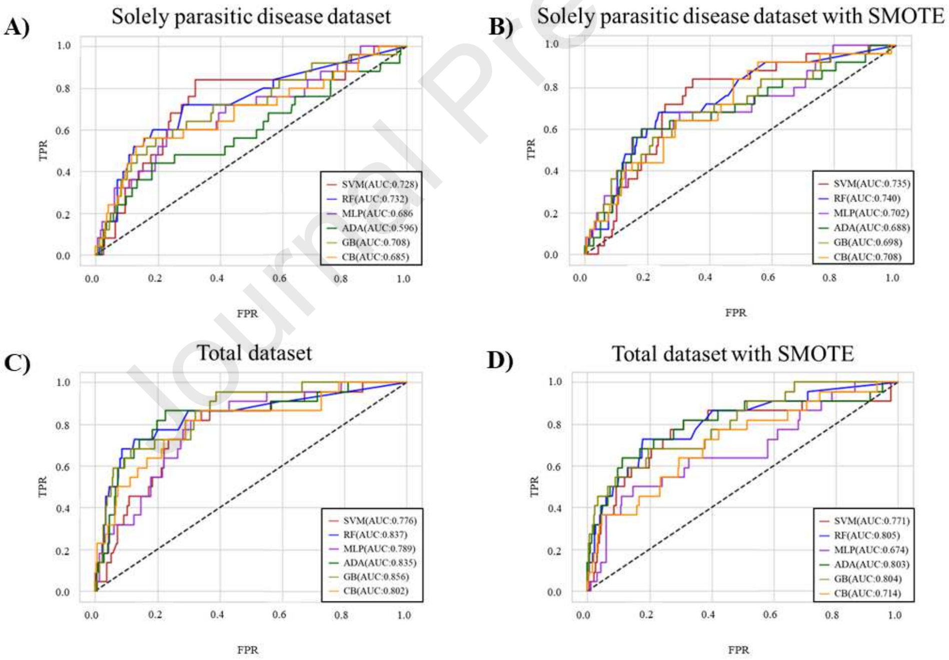

Fig 3. Feature importance: A) the solely parasitic disease dataset; B) solely parasitic disease dataset with SMOTE; C) total dataset; D) total dataset with SMOTE.

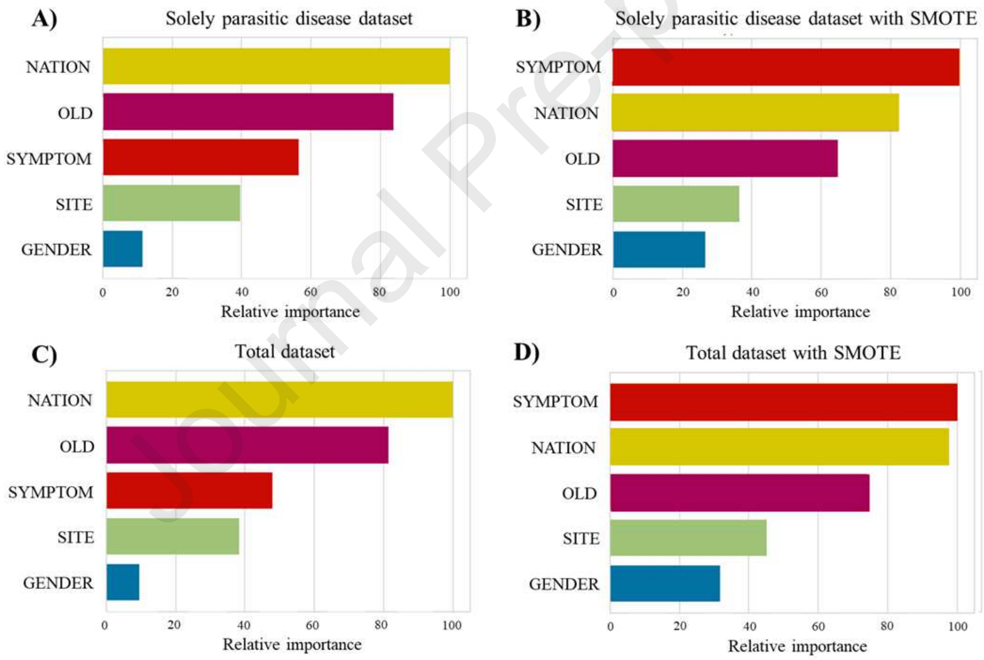

Table 1. Dataset review

Table 1. Dataset review

	Non - parasitic disease (N=148)	Only parasitic disease(N=1698)			Overall(N=1846)
		Non - malaria (N=1563)	Malaria (N=135)	Total	
Gender(n)					
Male	78	881	89	970	1048
Female	70	682	46	728	798
Age(n)					
1~20	6	334	17	351	357
21~40	24	562	56	618	642
41~60	54	421	50	471	525
61~80	56	233	12	245	301
81~	8	13	0	13	21
Nationality(n)					
Africa	16	251	69	320	336
America	16	309	14	323	339
Asia	87	591	37	628	715
Europe	27	364	13	377	404
Ocearnia & Caribbean	2	48	2	50	52
Symptomic body region(n, (%))	122(82.4)	1360(87)	78(57.8)	1438(84.7)	1560(83.7)
ABDOMEN	21	476	24	500	521
BACK	7	32	7	39	46
CHEST	15	67	4	71	86

EAR	1	9	2	11	12
EXTREMITIES	4	22	1	23	27
GASTROINTESTINAL	8	68	2	70	78
HAIR	0	9	0	9	9
HEAD	0	28	0	28	28
LYMPH NODE	8	35	0	35	43
MOUTH	0	12	1	13	13
NAIL	1	4	0	4	5
NECK	3	41	1	42	45
NEUROLOGICAL	5	78	9	87	92
OCULAR	3	93	1	94	97
PELVIS	0	5	0	5	5
PSYCHIATRIC	0	7	6	13	13
PULMONARY	26	152	13	165	191
RECTUM	1	8	0	8	9
SKIN	15	164	5	169	184
TOOTH	1	1	0	1	2
VAGINA	1	17	0	17	18
VISION	2	32	2	34	36
Symptom(n, (%))	46(31.1)	437(27.3)	86(63.7)	523(30.8)	575(31.1)
ALOPECIA	1	2	0	2	3
APATHY	0	4	0	4	4
APHASIA	1	6	0	6	7
APNEA	0	0	0	0	1
APRAXIA	0	2	0	2	2
ARRHYTHMIA	0	0	1	1	1
ARTHRALGIA	5	11	5	16	21
ASTHENIA	0	1	0	1	1
ATAXIA	1	12	3	15	16
BACK PAIN	7	18	0	18	25
BLEEDING	0	19	2	21	21
BLINDNESS	0	12	1	13	13
BLURRED VISION	0	6	0	6	6
CHILLS	1	8	7	15	16
CHRONIC PAIN	0	0	0	0	1
CONFUSION	2	7	2	9	11
DEFORMITY	0	4	0	4	4
DEPRESSION	0	3	1	4	4
DISCHARGE	2	15	1	16	18

DIZZINESS	0	5	0	5	5
DYSARTHRIA	0	0	0	0	2
FECAL INCONTINENCE	0	0	0	0	1
FEVER	14	131	43	174	188
HALLUCINATION	0	2	0	2	2
HEARING LOSS	0	2	2	4	4
HEARTBURN	0	3	0	3	3
HEMATEMESIS	0	2	0	2	2
INFERTILITY	0	3	0	3	3
IRRITABILITY	0	1	1	2	2
LACERATION	0	2	0	2	2
LHERMITTE'S SIGN	0	1	0	1	1
LOSS OF CONSCIOUSNESS	0	13	1	14	14
MALAISE	0	10	5	15	15
MYOCLONUS	0	0	0	0	1
NECK STIFFNESS	0	3	1	4	4
PARALYSIS	0	3	0	3	3
PARESIS	0	5	0	5	5
PELVIC PAIN	0	7	0	7	7
PETECHIA	0	3	0	3	3
PURPURA	0	3	1	4	4
RASH	0	1	0	1	1
SHIVERING	0	3	4	7	7
SHORT OF BREATH	0	4	0	4	4
SORE THROAT	0	2	0	2	2
SUICIDAL IDEATION	0	0	3	3	3
SWEATS	1	6	1	7	8
SWELLING	6	71	0	71	77
TINGLING	0	2	0	2	2
TREMOR	3	0	1	1	4
TRISMUS	0	1	0	1	1
URINARY RETENTION	1	10	0	10	11
VAGINAL DISCHARGE	1	4	0	4	5
VOMIT	0	4	0	4	4

Table 2. Solely parasite dataset

Table 2. Solely parasitic disease dataset

Model	Accuracy	Precision	Recall	F1-Score	CV-10	AUC
SVM	0.915	0.000	0.000	0.000	0.914	0.728
RF	0.903	0.250	0.160	0.195	0.906	0.732
MLP	0.909	0.286	0.160	0.205	0.916	0.686
Ada	0.894	0.211	0.160	0.182	0.905	0.596
GB	0.891	0.227	0.200	0.213	0.913	0.708
CB	0.909	0.250	0.120	0.162	0.919	0.685
SMOTE+SVM	0.874	0.091	0.080	0.085	0.914	0.735
SMOTE+RF	0.871	0.120	0.120	0.120	0.906	0.740
SMOTE+MLP	0.721	0.150	0.600	0.240	0.916	0.702
SMOTE+Ada	0.865	0.161	0.200	0.179	0.915	0.688
SMOTE+GB	0.885	0.208	0.200	0.204	0.912	0.698
SMOTE+CB	0.871	0.194	0.240	0.214	0.919	0.708

Table 3. Total dataset

Table 3. Total dataset

Model	Accuracy	Precision	Recall	F1-Score	CV-10	AUC
SVM	0.938	0.000	0.000	0.000	0.921	0.776
RF	0.930	0.300	0.136	0.187	0.917	0.837
MLP	0.914	0.222	0.182	0.200	0.917	0.789
Ada	0.932	0.333	0.136	0.194	0.910	0.835
GB	0.930	0.300	0.136	0.187	0.908	0.856
CB	0.949	0.714	0.227	0.345	0.924	0.802
SMOTE+SVM	0.919	0.278	0.227	0.250	0.921	0.771

SMOTE+RF	0.922	0.348	0.364	0.356	0.917	0.805
SMOTE+MLP	0.746	0.125	0.545	0.203	0.917	0.674
SMOTE+Ada	0.922	0.360	0.409	0.383	0.907	0.803
SMOTE+GB	0.927	0.400	0.455	0.426	0.907	0.804
SMOTE+CB	0.881	0.211	0.364	0.267	0.924	0.714

Appendix A. Supplementary Table 1. Query list of parasitic disease

Parasitic disease list	Query list
Chagas disease	((Trypanosoma OR Chagas Disease OR trypanosomiasis) AND America) AND case report
Sleeping sickness	((Trypanosoma OR Sleeping Sickness OR trypanosomiasis) AND Africa) AND case report
Acanthamoeba Infection	(Acantamoeba OR Granulomatous Amebic Encephalitis) AND case report
Angiostrongyliasis	(Angiostrongylus OR Angiostrongyliasis) AND case report
Anisakiasis	(Anisakis OR Anisakiasis OR Pseudoterranova) AND case report
Ascariasis	(Ascaris OR Ascariasis OR Intestinal Roundworms) AND case report
Babesiosis	(Babesia OR Babesiosis) AND case report
Balantidiasis	(Balantidium OR Balantidiasis) AND case report
Baylisascariasis	(Baylisascaris OR Baylisascariasis OR Raccoon Roundworm) AND case report
Bed Bugs	(Bed Bugs OR Tropical bedbug OR Triatomid bug OR Cimex OR Panstrongylus megistus OR Rhodnius OR Triatoma protracta) AND case report
Capillariasis	(Capillaria OR Capillariasis) AND case report
Cercarial Dermatitis	(Cercaria OR Cercarial Dermatitis OR Swimmer's Itch) AND case report
Clonorchiasis	(Clonorchis OR Clonorchiasis)AND case report

Cryptosporidiosis	(Cryptosporidiosis OR Cryptosporidium) AND case report
Cyclosporiasis	(Cyclospora OR Cyclosporiasis) AND case report
Cysticercosis	(Cysticercosis OR Neurocysticercosis OR Taenia OR Cysticercus) AND case report
Cystoisosporiasis	(Cystoisosporiasis OR Isospora OR Cystoisospora) AND case report
Dientamoeba fragilis Infection	(Dientamoeba fragilis) AND case report
Dipylidium caninum Infection	(Diphyllobothriasis OR Diphyllobothrium OR tapeworm) AND case report
Dirofilariasis	(Dirofilaria OR Dirofilariasis) AND case report
Echinococcosis, Hydatid Disease	(Echinococcus OR Echinococcosis OR Hydatid Disease OR Hydatidosis) AND case report
Nonpathogenic Intestinal Protozoa	(Endolimax nana OR Entamoeba coli OR Entamoeba dispar OR Entamoeba hartmanni OR Entamoeba polecki OR Nonpathogenic Intestinal Protozoa OR Harmless Intestinal Protozoa OR Iodamoeba buetschlii OR Entamoeba gingivalis) AND case report
Amoebiasis	(Entamoeba OR Amebiasis) AND case report
Enterobiasis	(Enterobiasis OR Pinworm OR Enterobius) AND case report
Fascioliasis	(Fasciola OR Fascioliasis) AND case report
Fasciolopsiasis	(Fasciolopsiasis OR Fasciolopsis) AND case report
Filariasis	(Filaria OR Filariasis OR Elephantiasis OR Wuchereria bancrofti OR Brugia) AND case report
Myiasis	(fly OR Myiasis OR Dermatobia hominis OR bot fly OR Cochliomyia hominovorax OR screwworm fly OR Chrysomya bezziana OR screwworm OR Cordylobia anthropophaga OR tumbu fly OR Cuterebra OR Oestrus OR Wohlfahrtia) AND case report
Giardiasis	(Giardia OR Giardiasis) AND case report
Gnathostomiasis	(Gnathostoma OR Gnathostomiasis) AND case report
Dracunculiasis	(Guinea OR Dracunculiasis OR Dracunculus medinensis) AND case report

Heterophyiasis	(Heterophyes OR Heterophyiasis) AND case report
Ancylostomiasis/Hookworm	(Hook worm OR Ancylostomiasis OR Cutaneous larva migrans OR Ancylostoma OR Necator americanus) AND case report
Hymenolepiasis	(Hymenolepis OR Hymenolepiasis) AND case report
Leishmaniasis	(Leishmania OR Leishmaniasis OR Kala-azar) AND case report
Loiasis	(Loa loa OR Loiasis) AND case report
Lice Infestation	(Louse OR Body Lice OR Pediculosis OR Pthiriasis OR Pubic lice OR Pubic crab lice OR Head Lice OR Phthirus pubis OR Pediculus humanus corporis) AND case report
Malaria	(Malaria OR Plasmodium) AND case report
Microsporidiosis	(Microsporidiosis OR Microsporidia OR Anncaliia algerae OR Anncaliia connori OR Anncaliia vesicularum OR Brachiola OR Anncaliia OR Encephalitozoon cuniculi OR Encephalitozoon hellem OR Encephalitozoon intestinalis OR Septata intestinalis OR Tubulinosema acridophagus OR Enterocytozoon bieneusi OR Microsporidium ceylonensis OR Microsporidium africanum OR Nosema ocularum OR Pleistophora OR Trachipleistophora hominis OR Trachipleistophora anthropophthera OR Vittaforma corneae OR Tubulinosema acridophagus) AND case report
Mite Infestation	(mite OR Scabies OR Sarcoptes) AND case report
Naegleria Infection	(Naegleria OR brain eating amoeba OR primary amebic meningoencephalitis) AND case report
Onchocerciasis	(Onchocerciasis OR Onchocerca OR River Blindness) AND case report
Opisthorchiasis	(Opisthorchis OR Opisthorchiasis) AND case report
Paragonimiasis	(Paragonimus OR Paragonimiasis OR flatworm OR) AND case report
Sappinia	(Sappinia OR amebic encephalitis) AND case report
Sarcocystosis	(Sarcocystosis OR Sarcocystis) AND case report
Schistosomiasis	(Schistosomiasis OR Schistosoma OR Bilharzia) AND case report

Strongyloidiasis	(Strongyloidiasis OR Strongyloides) AND case report
Taeniasis	(Taenia OR Taeniasis OR Tapeworm OR cysticercosis) AND case report
Toxocariasis	(Toxocara OR Ocular Larva Migrans OR Toxocariasis OR Roundworm OR Visceral Larva Migrans) AND case report
Toxoplasmosis	(Toxoplasma gondii OR Toxoplasmosis OR Toxoplasma) AND case report
Trichinosis	(Trichinella OR Trichinellosis OR Trichinosis) AND case report
Trichomoniasis	(Trichomonas OR Trichomoniasis) AND case report
Trichuriasis	(Trichuris OR Whipworm OR Trichuriasis) AND case report
Balamuthia	Balamuthia AND case report
Chilomastix mesnili Infection	Chilomastix mesnili AND case report

520

Highlights

- A machine learning model is proposed to predict malaria using patient information from parasite case reports.
- Feature importance analysis indicates that the nationality and region of travel are important factors to diagnose malaria.
- SMOTE application does not achieve any considerable improvement

Conflict of interest

The authors have no competing interests to declare.