Erasmus
School of
Economics

Turing Students Machine and Deep Learning 2024

Final Report

Detecting Malaria Using Machine Learning Algorithms

Antonios Dimitriou [599877]

Maria Pawlak [616541]

Rana Uzunoğlu [598031]

Word count: [1735] | Date: 07-04-2024

Erasmus University Rotterdam
Make it happen.

Abstract

In this study, Machine Learning (ML) algorithms are used to train a model on microscope images of cells, some infected with malaria while others not. This is done in order for it to be able to identify whether an individual is a carrier from a microscope image of a cell. The results show that the model can indeed identify, to some extent, if an individual is infected or not. This model has the potential of aiding in the identification of carriers, but it can also prove to be beneficial in more ways (financially, etc.) wherever implemented as discussed.

Introduction

Even though medicine has evolved exponentially over the past couple of centuries, the world is still suffering from health issues, caused by diseases, such as malaria. Malaria is a life-threatening disease and it is very prominent in tropical regions and countries (World Health Organisation, Malaria (WHO, 2024) such as Nigeria and the Democratic Republic of Congo. With there being approximately 250 million malaria cases a year (World Health Organisation, Malaria (WHO, 2024), one would wonder how this situation can be improved, perhaps through a technological method. This is the question that is attempted to be answered in this report; Is it possible to identify malaria particles using ML algorithms?

While attempting to use ML algorithms to detect malaria, two rudimentary hypotheses are formed. This includes the null hypothesis (Ho) which states that ML algorithms cannot accurately identify malaria, and the alternative hypothesis (Ha), which states that ML algorithms can accurately identify malaria.

With malaria being a disease that can be detected from microscope images of the blood cells of an individual, the model created aims to differentiate between images of infected blood samples and uninfected ones. In other words, the model has been created to carry out a classification task.

The dataset used is the Tensorflow 'malaria' dataset. The data is composed of microscope images of cells compiled from various sources by researchers of the National Library of

Medicine (NLM), and it has been used in the development of a "Malaria Screener" smartphone application by researchers at the Lister Hill National Center for Biomedical Communications (LHNCBC, see appendix for details on the data). These images are labelled as either "parasitised (0)" or "uninfected (1)". The dependent variable that the model aims to predict is the label of the images, and the independent variable used is the images themselves.

Methods

We use two different approaches in our analysis, namely K-means clustering and Principal Component Analysis (PCA) combined with Support Vector Machine (SVM) classification. In the second approach, we use PCA in order to reduce the dimensionality of our data later which we train a SVM model with.

K-means clustering partitions a dataset into K clusters by iteratively assigning each point to the nearest cluster centroid while updating the means of the centroids. PCA is a dimension reduction technique that works by identifying directions (principal components) that explain most of the total variation in the data. SVM is a supervised learning algorithm that classifies data points into separate classes by maximising the distance between them.

The rationale behind using K-means clustering is its simplicity and efficiency in handling large datasets. By clustering similar images together, K-means can potentially group malaria-infected and uninfected cells separately. SVM was proposed for its high performance with high-dimensional data. However, due to the very high dimensionality of our data due to its image nature, we run into the risk of "curse of dimensionality." Using PCA as a dimensionality reduction measure addresses these issues by reducing the dimensionality of the data while retaining most of its variance.

Before using any machine learning methods, we perform several preprocessing steps on the image dataset. We first resize the images to a consistent size of 100x100 pixels, convert them to grayscale to reduce complexity, flattening the images into one-dimensional Numpy arrays that are suitable for input to the algorithms used. For the SVM model, we standardise the arrays to have mean zero and standard deviation one, and apply PCA to the standardised features. We then use this to train our SVM.

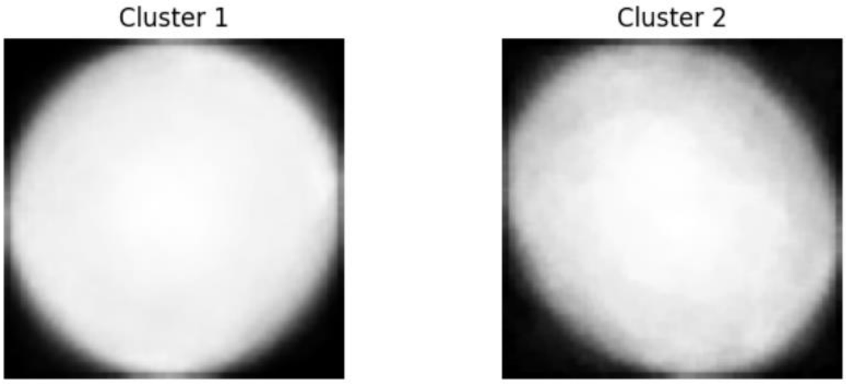
Results

The classification accuracy achieved with K-means clustering was approximately 40%. This relatively low accuracy suggests that K-means clustering alone may not be suitable for accurately classifying malaria-infected cells. While K-means is a simple and unsupervised approach for clustering data, its performance heavily relies on the inherent structure and separability of the data. This might not be well-suitable for this task since the structure of all image files are relatively similar and the separability is limited in the sense that cell images may exhibit subtle variations and complexities that make it challenging to distinguish between infected and uninfected cells.

In contrast, the SVM classifier used together with PCA resulted in a significantly higher accuracy of about 75%. This indicates the effectiveness of using a supervised learning approach with dimensionality reduction for malaria-infected cell classification. By reducing the dimensionality of the data with PCA and training an SVM classifier on the transformed feature space, our model was able to make a significantly better separation between infected and uninfected cell images.

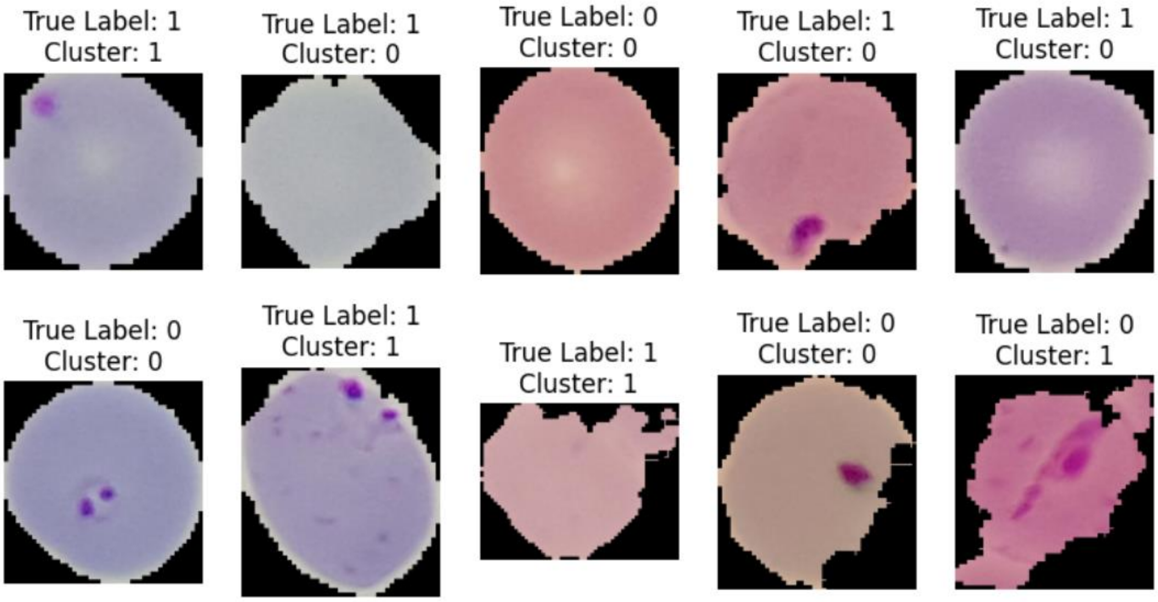
We use both accuracy and recall as our preferred evaluation metrics. Accuracy is defined as the ratio of correct predictions to the total number of predictions while recall focuses on the model’s ability to correctly identify positive instances, which is especially important in the medical context to maximise the number of true positives. For both approaches, we calculate both metrics and find that recall is equal to the accuracy. While this can be the case if the two clusters identified by the K-means clustering perfectly corresponds to the classes of our classification task (infected and uninfected) or in the case where the dataset has an equal number of instances of each class and the models used are effective at separating these classes. In the dataset that we used, an equal number of infected and uninfected images are present. However, we believe that this alone may not explain why accuracy and precision are equal for both approaches, and that this is a point for future exploration.

Centroids of K-means clustering:



Visualisation of the labelling assigned by K-means:

1: healthy and 0:infected



Discussion

Both K-means clustering and SVM with PCA algorithms struggled with accurately identifying malaria-infected cells due to the nuanced complexity of the procedure, scoring accuracy of 40% and 75% respectively. SVM with PCA significantly outperformed K-means clustering. The discrepancy between them underscores the usefulness of utilising supervised ML algorithms in medical classification. Moreover, the results imply effectiveness of dimensionality reduction in enhancement of overall accuracy.

Our research concludes that it is possible to identify malaria particles using ML algorithms to a certain extent, hence aligning with the null hypothesis. Accuracy and recall were used as the key indicators of the model performance in identifying malaria-infected cells. The superior performance of SVM with PCA algorithm reinforces findings of this work.

To evaluate null hypothesis, both supervised and unsupervised ML algorithms were utilised. It allowed for a comprehensive comparison of different approaches and their performance in medical classification tasks.

One such unsupervised model was K-means clustering, known for its independence from labelled data for training and outstanding computational efficiency for primary analysis. These characteristics make it desirable for tasks like malaria identification. However, the model's performance turned out to be far from satisfactory, likely due to the focus on the intrinsic structure of the data. This led to difficulty in capturing subtle variations between images of infected and uninfected cells.

On the contrary, SVM with PCA tested supervised learning methods. The biggest asset of this model in this research is reduction of dimensionality in feature space, enabling easier distinction of the cell. Yet, it sacrificed the interpretability of the relationship between original features and the classification decision.

While our model shows promise, it is essential to acknowledge its limitations. In medical diagnoses, especially for diseases like malaria, precision is invaluable. Therefore, it is recommended to complement this model with medical tests like PCR or RTD to boost overall accuracy of the treatment.

As for the ML model, typically, a larger dataset leads to improved accuracy and reduced risk of selection bias. Therefore, if it was to be upgraded, significantly more data should be used. Additionally, utilised data could be transformed through rotating the images, scaling or adding noise. This could potentially capture underlying patterns more effectively, mitigating the risk of overfitting or underfitting. Finally, as we were utilising K-means and PCA algorithms, exploring variations in their hyperparameters, such as the number of clusters or initialization methods, could enhance the model's performance.

Possible explanation for unexpected results can be attributed to the dataset characteristics such as balanced class distribution and separability. However, further research is crucial for deeper understanding of the impact of underlying factors on the outcome.

Conclusion

Our research aimed to uncover feasibility of identifying malaria-infected cells using supervised and unsupervised machine learning methods. By conducting a thorough analysis, we gained valuable insights into efficiency of both approaches. From our outcomes, it can be deduced that supervised learning methods outperform unsupervised ones in terms of accuracy and recall. The success of this model can be attributed to reduction of dimensions offered by SVM with the PCA algorithm.

Thanks to the comprehensibility of this approach, we managed to fully answer the research question. Utilising machine learning models enables identification of malaria-infected cells. Nevertheless, it is important to bear in mind that the model can, and should be tested on more ML methods to substantially increase the overall accuracy of the performance.

The recommendations for further research include subjecting the model to a variety of supervised and unsupervised algorithms to assess the accuracy and recall score they can obtain. This approach could provide more reliable and groundbreaking results. Among preferred methods are Decision Tree Models and Convolutional Neural Networks (CNNs).

Decision Tree Models can effectively capture nonlinear relationships between features and target variables, a trait useful for medical image classification. Convolutional Neural Networks on the other hand, are outstanding in learning hierarchical features directly from raw image data making it particularly well-suited for this model.

References

World Health Organization. (n.d.). *Fact sheet about malaria*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/malaria, Accessed on 06/04/2024

Link to TensorFlow Malaria Dataset

https://www.tensorflow.org/datasets/catalog/malaria

https://lhncbc.nlm.nih.gov/LHC-downloads/downloads.html#malaria-datasets