Supplementary Information for
‘Machine learning and applications in microbiology’

Stephen J. Goodswen¹, Joel L. N. Barratt², Paul J. Kennedy³*, Alexa Kaufer¹, Larissa Calarco¹, John T. Ellis¹*

¹ School of Life Sciences, University of Technology Sydney (UTS), Ultimo, NSW, Australia.

² Parasitic Diseases Branch, Division of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, GA, USA.

³ School of Computer Science, Faculty of Engineering and Information Technology and the Australian Artificial Intelligence Institute, University of Technology Sydney (UTS), Ultimo, NSW, Australia.

* Correspondence: John Ellis (Microbiology) and Paul Kennedy (Machine Learning)

Email address: john.t.ellis@alumni.uts.edu.au, Paul.Kennedy@uts.edu.au

Preface

The topic of machine learning (ML) is complex and extensive. In the main article ‘Machine learning and applications in microbiology’, we have attempted to find that perfect balance between too much and too little detail to address this complexity. However, for those readers requiring that bit of extra background and detailed explanations this supplementary document complements the article.
Contents

Introduction ................................................................................................................... 5

Popular algorithms used in machine learning ......................................................... 5

Logistic regression ....................................................................................................... 5

Decision tree ................................................................................................................ 5

Random forest ............................................................................................................ 6

k-Nearest Neighbour (kNN) Classifier ................................................................. 6

Naive Bayes Classifier ............................................................................................. 6

Neural networks .......................................................................................................... 6

Support vector machine (SVM) .................................................................................. 7

Adaptive boosting ....................................................................................................... 7

Clustering ..................................................................................................................... 7

Dimensionality reduction .......................................................................................... 8

Decision tree example ............................................................................................... 8

Real life examples ....................................................................................................... 10

Example 1 (Rajaraman et al., 2018) | Protozoa | Clinical application | Malaria diagnosis using binary image classification ........................................... 10

Example 2 (Torrecilha et al., 2017) | Protozoa | Clinical application | Predicting parasite load for Leishmania infantum from clinical records when qPCR services are not available using radial basis artificial neural network ................................................................. 11

Example 3 (Mathison et al., 2020) | Protozoa | Clinical application | Gastrointestinal parasite diagnosis using binary image classification ..................... 12

Example 4 (Goodswen et al., 2014) | Protozoa | Vaccine discovery | Predicting candidates against Apicomplexan pathogens using an ensemble of classifiers ................................................................. 13

Example 5 (Verma et al., 2008) | Protozoa | Drug/Vaccine discovery | New method based on amino acid composition for predicting secretory proteins from malarial parasites using support vector machines ................................................................. 15

Example 6 (Sinha et al., 2017) | Protozoa | Drug/Vaccine discovery | Identify potential targets against Leishmania parasites using Naive Bayes Probabilistic Classifier ................................................................. 16

Example 7 (Maindola et al., 2015) | Protozoa | Drug discovery | Identifying compounds in traditional Chinese medicine to use as anti-malarials using various binary classification algorithms ................................................................. 17

Example 8 (Levatic et al., 2018) | Protozoa | Drug discovery | Identifying derivatives of Primaquine (antimalarial drug) that will retain or improve its efficacy, while reducing its toxicity using QSAR modelling – a regression problem ................................................................. 19
Example 9 (Plucinski et al., 2015) | Protozoa | Epidemiology and antimicrobial resistance | Distinguishing between a relapse of illness due to failure of antimicrobial treatment and infection with a new malaria strain ................................................................. 21

Example 10 (Perdiguero-Alonso et al., 2008) | Protozoa | Taxonomy | Fish discrimination given parasite community data as biological markers using random forest (a multi-classification problem) .................................................................................................................. 21

Example 11 (Yang et al., 2018) | Bacteria | Antimicrobial resistance | Identification of drug-resistant tuberculosis (TB) .................................................................................................................. 23

Example 12 (Burckhardt et al., 2019) | Bacteria | Epidemiology and Antimicrobial resistance | Confirmation of Streptococcus pneumoniae capsular serotypes on the basis of fourier-transformed infrared spectra ............................................................. 24

Example 13 (Ramirez et al., 2018) | Bacteria | Macroecology | Assessing bacterial communities across global soils using supervised and unsupervised random forest .............................................................. 24

Example 14 (Sohn & Li, 2018) | Bacteria | Microbial ecology and microbiomes | Distinguishing microbiome signatures collected from different body sites based on their bacterial 16S rDNA sequence composition .............................................................. 25

Example 15 (Midani et al., 2018) | Bacteria | Microbiome | Predicting susceptibility to Vibrio cholerae infection .............................................................................................................................. 26

Example 16 (Poore et al., 2020) | Bacteria | Microbiome | Cancer diagnostics using microbiome analyses of human blood and tissues ................................................................................................................ 27

Example 17 (Colubri et al., 2016) | Viruses | Clinical applications | Ebola Virus Disease (EVD) prognosis prediction using various classification algorithms .................................................................................................................. 28

Example 18 (Zorn et al., 2019) | Viruses | Drug discovery | Identifying new compounds against HIV with various classification algorithms ............................................................................................................ 29

Example 19 (Silva et al., 2017) | Viruses | Plant microbes | Multi-classification of large volumes of Geminiviridae sequences ............................................................................................................................ 29

Example 20 (Sperschneider et al., 2016) | Fungi | Plant pathogens | Fungal effector prediction from secretomes using Naive Bayes classifier .............................................................................................. 30

Example 21 (Kalafi et al., 2016) | Helminths | Taxonomy | Identification of Monogenean species via image classifications using kNN ............................................................................................................................ 31

Example 22 (Barratt et al., 2019, Barratt & Sapp, 2020, Nascimento et al., 2020) | Helminths and Protozoa | Taxonomy and Epidemiology | Assessment of genetic relationships in the context of complex genotypes using an ensemble learning procedure ............................................................................ 32

Example 23 (Cieslak et al., 2020) | Copepods | Microbial ecology and microbiomes | Distinguishing different groups of copepods in various transcriptional states based on RNA-Seq data .......................................................................................... 33

Tables ................................................................................................................. 34

Real-life published examples of supervised and unsupervised machine learning applied to microbiology ................................................................. 34
Introduction

It is unlikely that Arthur Samuel, the designer of a checkers-playing program back in 1959, had the realization his program design concepts would decades later have such a monumental impact on our everyday lives. An impact so great it has potential to rival that of the internet and change every industry. Samuel’s program was one of the world's first self-learning applications (Samuel, 1959) and ‘machine learning’ was a term he coined to encapsulate concepts underpinning his program. Machine learning (ML) now serves as a core subfield under artificial intelligence (AI).

Widely publicized and everyday examples of ML applications are online targeted advertisements (such as those from Amazon (Linden et al., 2003) and Netflix (Zhou et al., 2008)), credit card fraud detection (Bhattacharyya et al., 2011) (PayPal), facial recognition for friends’ suggestions (Shan et al., 2009) (Facebook), spam filters on e-mail (Guzella & Caminhas, 2009) (Gmail and Outlook), and the driverless-car (Gerla et al., 2014), which is possibly the most familiar application epitomizing the spirit of ML.

Popular algorithms used in machine learning

There are many machine learning algorithms. These are the sets of rules that steer the ML process to identify patterns in past data, build models, and make predictions or decisions without having explicit pre-programmed rules and set models. An outline of popular algorithms is now described.

Logistic regression – an algorithm based around the traditional statistical logistic function (also called the sigmoid function) and used for predicting probabilities and/or classification. Logistic regression uses a linear equation with independent predictors to predict a dependent variable (the target value). In theory, a predicted value from a linear equation can be anything from negative infinity to positive infinity. A logistic function, however, squashes the output of the linear equation into a probability range of 0 to 1, which when plotted is a common ‘S’ shape (sigmoid curve). Therefore, a model trained with logistic regression estimates how the probability of a binary dependent event (target value) may be affected by one or more input values (i.e. predictors). Strengths: uses traditional statistical regression techniques. Fast training and prediction times Limitations: sensitive to outliers and tends to be unstable with small datasets. Example application: classification tasks such as classifying whether a tumour is malignant (1) or benign (0) based on an estimated probability (p). If ‘p’ is closer to 1 than 0, then the tumour would be recorded as malignant.

Decision tree – this algorithm uses a recursive partitioning approach (see Decision tree example). A tree structure is created by recursively splitting the example dataset into subsets based on a statistical test on the value of an input variable. The approach is repeated until the splitting adds no further value to the prediction i.e. a point at which the subset has all or almost all of the same value of the target variable (e.g. a YES or NO candidacy). In effect a built decision tree represents a set of rules that can be used to predict classifications of new data with the same input variables. Strengths: decision trees can be easily understood and interpreted. They are flexible in terms of the data type of input and output variables which can be categorical, binary and numeric value. Limitations: decision trees tend to be the least accurate in comparison to the other algorithms described here, and so represent a trade-off between accuracy and a user’s ability to understand the process. Also, it is a concrete binary decision at each split point and only considers one input attribute at a time for the decision criteria. Furthermore, when new example data arrives, a new decision tree needs to be created by retraining every data from scratch.
**Random forest** – the general idea behind this algorithm is the combining of multiple decision trees (a forest of trees) into a single ensemble of models i.e. the expectation is that multiple models should perform better than any one single model. The ‘random’ aspect of the algorithm is the randomness in which it selects input variables and observations from the example data to build each individual decision tree. This means different trees generate different classification results. The final outcome is the average of the results for regression and a majority rules vote for classification. **Strengths:** addresses the limitations of decision trees and automatically identifies the importance of variables towards solving the problem. The algorithm is easy to use because there are only two parameters to optimise (number of trees and number of independent variables used at each split). **Limitations:** high prediction accuracy at the expense of explainability.

**k-Nearest Neighbour (kNN) Classifier** – based on assigning to an unclassified observation the same classification as the majority of nearest k classified example data points in a multidimensional feature space (See Fig. 1). The symbol k is a user-defined positive integer indicating the number of nearest neighbours to consider, and Euclidean distance is often used as the nearest distance metric. **Strengths:** easy to interpret output. It uses a non-parametric technique, which means that it makes no assumptions on the underlying data distribution (unlike linear regression for example). Comparatively low calculation time because kNN does not need to learn any model. **Limitations:** sensitive to outliers and noise and allows for numerical input only. Also, the entire training data are stored during the kNN execution, which makes it a computationally expensive algorithm.

**Naive Bayes Classifier** – this algorithm computes a *posterior* probability for the classification of new data based on applying Bayes’ theorem with ‘naïve’ independence assumptions i.e. it assumes that each input variable associated with the example dataset independently contributes to the classification. **Strengths:** it is highly scalable and can learn incrementally by updating the probability distribution table with any new counted observed variables. **Limitations:** expects that data is categorical, although some versions are able to deal with numeric data. Poor performance if the assumption of conditional independence is violated or if there are not enough data points to calculate the probabilities. **Example application:** detecting spam e-mails. For example, the frequency of words and/or patterns (features) from previous e-mails (past data) known to be spam is determined for the training data. Each feature is assigned a spam association probability e.g. the word ‘cheap’ assigned 70%, a spelling mistake 65%, a missing subject 80%. The Naive Bayes trained model then predicts/classifies a new e-mail as spam or not spam.

**Neural networks** – neural networks are conceptually artificial neurons inspired by the way biological neural networks in the human brain process information e.g. as humans, certain neurons in our brain are stimulated (activated) when we see an image, sending signals to other neurons which send signals to even more neurons, ultimately resulting in certain neurons being activated that help us to distinguish (or classify) the seen image. A basic neural network starts with an input layer of neurons, which activate neurons in the hidden layers, which then activate neurons in the output layer. The basic unit of computation is the **neuron**, often called a **node** or **unit**. Each neuron in effect is a logistic regression unit which receives multiple numeric inputs (from other nodes or from an external source) and computes one numeric output, typically between 0 and 1. The network of multiple neurons (nodes) is arranged in **layers**. Nodes from adjacent layers have **connections** or **edges** between them. All these connections have a **weight** associated with them. That is, each numerical input into the neuron has a weight and bias associated with it (one bias per node, one weight per connection) that influences the training of the network. The neuron output is computed by a non-linear function called the **Activation Function**. If the weighted sum of the inputs exceeds an internal threshold value within
the neuron, the output is activated, otherwise it is inactivated. The network learns by repeatedly adjusting the weights on the inputs to reinforce correct classifications (i.e. activation of the neuron) and to discourage wrong classifications (i.e. inactivation of the neuron). **Strengths:** the multi-layer model structure enables a neural network to learn non-linear relationships between input and output. **Limitations:** a black-box model that makes it difficult to understand how it solves the problem. Expects numerical input, which means categorical data requires transformation.

**Support vector machine (SVM)** – this algorithm engages vectors positioned at the edge of an area in feature space to present a boundary between two classes of observations e.g. vaccine and non-vaccine candidates (see Fig. 2). The general idea is to identify a straight line in feature space that separates the classes. New observations can then be classified depending on which side of the line it falls on (Cortes & Vapnik, 1995). However, example data may not be distributed in such a way that it can be linearly separated. In such cases, there are various methods coined kernels (e.g. Radial Basis ‘Gaussian’, polynomial, linear, hyperbolic tangent, Laplacian, Bessel, ANOVA RBF, and spline) that transform the data into a high-dimensional feature space i.e. in effect create a margin between the two classes allowing linear separation by a hyperplane. Fig. 3 shows the principle behind the kernel method. SVMs use linear optimisation (rather than gradient descent) to find the best separating line i.e. it maximises the margin through the feature space. There are also several model types (e.g. C, nu, and bound-constraint classifications) which determine the hyperplane. **Strengths:** only concerned with data close to boundary i.e. robust to outliers and can work with small datasets. **Limitations:** determining which combination of kernel and model type will build the best SVM model tends to be a trial and error process. **Example applications:** can be used in both classification and regression challenges, but mainly used in classification.

**Adaptive boosting** (AdaBoost) – a meta-algorithm in the sense that it is used in conjunction with other ML algorithms to build multiple models with the objective to perform better than any one single model. For example, consecutive individual decision trees from the same training dataset can be built. During the training process the AdaBoost algorithm first applies an initial equal weight to each row of evidence. Then, after each tree is built, the algorithm increases (boosts) the weight to any row that is incorrectly classified. The boosted weight has the effect that the misclassified evidence row is multiplied to have more representation during the building of the next tree and is more likely to be correctly classified. However, the weight is boosted even further if the classification is still incorrect and hence the ‘adaptive’ nature of the AdaBoost algorithm. Each model at the end of the training procedure has an associated weight and the larger the weight the greater the expected accuracy. The ensemble of built models is used in deployment to predict the classification of a new protein. A probability prediction made by each model is multiplied by the weight associated with the model and the final classification is the average of the predictions (i.e. a majority vote decides the final classification). **Strengths:** AdaBoost per se requires less tweaking of parameters or settings in comparison to most other classifiers. **Limitations:** dependent on the ML algorithms chosen to build the models, but adaptive boosting should reduce their effects.

**Clustering** – is an unsupervised machine learning procedure, also referred to as **unsupervised classification.** Clustering can be hierarchical or partitional, the difference being that clusters identified by partitional clustering do not overlap (they are un-nested); while clusters identified using a hierarchical clustering procedure are subsets of larger clusters (nested). Hierarchical clustering does not require the user to supply a value for k, while for partitional clustering a value for k must be supplied. K-means clustering and agglomerative nested clustering are popular partitional and hierarchical clustering approaches, respectively, though there are numerous clustering methods to
choose from (Xu & Tian, 2015). The objective of each algorithm is generally the same; to identify patterns from a complex, highly dimensional dataset typically provided in the form of a dissimilarity matrix. **Strengths:** Clustering produces a highly intuitive representation of relationships between each sample – a likely reason for the techniques widespread use. **Limitations:** With respect to hierarchical clustering, massive datasets may be difficult to visualise due to crowding, while in the case of partitional clustering selection of an appropriate value for \( k \) can be a challenge. There are a plethora of different clustering procedures and distance metrics to choose from and it can be difficult to identify the most appropriate method for a particular dataset – this will usually require some experimentation.

**Dimensionality reduction** – much like clustering, dimensionality reduction (also referred to as ordination) is an unsupervised machine learning technique that aims to identify patterns from a complex, highly dimensional dataset typically provided in the form of a dissimilarity matrix. However, the objective of dimensionality reduction differs slightly to clustering, in that the goal is to represent a highly dimensional dataset as points in two- or three-dimensional space, such that specimens (data points) that fall closer together share a series of common characteristics. In this way, ordination methods are often used as a visualisation aid for highly complex datasets. If the user chooses, the resulting set of two- or three-dimensional coordinates may also be clustered using a partitional approach to highlight specimens possessing shared features. There are various ordination methods to choose from (van-der-Maaten & Hinton, 2008, M & Li, 2018, Hawinkel et al., 2019, Sun et al., 2019) and selection of the most appropriate may require some experimentation on the part of the user. **Strengths:** Visualisations are highly intuitive and straightforward to interpret, even for massive datasets. **Limitations:** Visualisations are less granular than hierarchical clustering and there is a loss of information when the dataset is reduced to two- or three-dimensions. There are also a plethora of potential dimensionality reduction procedures to choose from and selecting the most appropriate can be confounding.

**Decision tree example**

This algorithm uses a recursive partitioning approach. A tree structure is created by recursively splitting the example dataset into subsets based on a statistical test on the value of an input variable. The approach is repeated until the splitting adds no further value to the prediction i.e. a point at which the subset has all or almost all of the same value of the target variable (e.g. a YES or NO candidacy). In effect a built decision tree represents a set of rules that can be used to predict classifications of new data with the same input variables. **Example:** Fig. A shows plotted points representing the example data. Only two input features are used here, S1 and S2. The example data are binary labelled. In this instance ‘Yes’ or ‘NO’ designated by blue triangles and green circles respectively (the label could be ‘1’ or ‘0’, ‘vaccine candidate’ or ‘non-vaccine candidate’ etc.). Binary splitting is at the core of constructing decision trees, where the aim is to split the data to only have one class in each region. Fig. B shows the first binary split denoted by a dashed line and the first decision criterion e.g. \( S2 > 1 \). The data points in the region \( S2 < 1 \) (a ‘no’ decision) represent only one class and consequently there is no need to split the region further. The next splits are shown in Fig. C and D in which more decisions are created representing a branching tree structure. The decision tree translates into rules, where each rule corresponds to one pathway through the tree. For example, if we have a new unclassified data point shown here as a red star, the point would be classified ‘Yes’ because it is governed by rule two (in this case, majority wins).
Real life examples

The following 23 examples are taken from published studies. The aim here is to present the essence of ML-related strategies taken to overcome specific biological problems.

Example 1 (Rajaraman et al., 2018) | Protozoa | Clinical application | Malaria diagnosis using binary image classification

**Background:** examination of microscopic thick and thin blood smears remains the ‘gold standard’ for Malaria diagnosis (Rajaraman et al., 2018).

**Problem setting:** The accuracy of a malaria diagnosis depends on the expertise of the human microscopist. As a consequence, diagnosis has been attempted using ML techniques with image analysis-based computer-aided diagnosis (CADx) software (Ross et al., 2006, Poostchi et al., 2018). Nonetheless, this automated process still requires human expertise to analyse variations of the region of interest (ROI) on the images (this requirement is termed ‘hand-engineered features’) e.g. there can be variability in size, background, angle, and position of the images. To overcome the ‘hand-engineered features’ requirement, studies (Liang et al., 2016, Bibin et al., 2017, Dong et al., 2017) have applied CNN models with significant success towards classifying parasitized and uninfected cells. However, none of these and other known DL-based studies have evaluated their predictive models at the patient level i.e. evaluation has been on relatively small image sets and/or randomized train/test splits (Rajaraman et al., 2018).

**Aim of Study:** cross-validation at the patient level of pre-trained CNN models as feature extractors toward classifying parasitized and uninfected cells.

**Training Data:** consisted of 27,558 cell images with equal instances of parasitized (i.e. positive samples containing *Plasmodium*) and uninfected cells (i.e. negative samples containing no *Plasmodium*). These images were obtained from a National Library of Medicine (NLM) archive. The source of the images was photographed from slides of thin blood smears extracted from 150 *P. falciparum*-infected and 50 healthy patients in Bangladesh.

**Algorithms used:** customized and pre-trained CNNs. Feature extraction was performed using pre-trained models from the Large Scale Visual Recognition Challenge (ILSVRC). An annual competition evaluating algorithms for object detection and image classification at large scale: AlexNet (Krizhevsky et al., 2017) (2012 winner), VGG-16 (winner of ILSVRC's localization task in 2014) , and ResNet-50 (Huang et al., 2017) (2015 winner). Other pre-trained models used DenseNet-121 (winner of the best paper award in 2017 Conference on Computer Vision and Pattern Recognition) (Huang et al., 2017) and Xception (Chollet & Ieee, 2017).

**Validation of trained model:** Predictive models were evaluated through five-fold cross-validation at the patient level.

**Statistical evaluation measures used:** Predictive models evaluated in terms of accuracy, AUC, SN, SP, F1-score(Lipton et al., 2014) and Matthews correlation coefficient (MCC) (Matthews, 1975). Shapiro_Wilk test (Shapiro & Wilk, 1965) was used to check for data normality. For non-normality data, Kruskal-Wallis H test (Kruskal & Wallis, 1952) statistical analyses was performed to choose the best model for deployment. A post-hoc analysis was needed to identify the specific models that
demonstrate statistically significant differences in performance mean values (Kucuk et al., 2016). The optimal layer for feature extraction to aid in improved classification was determined empirically.

**Best model chosen:** ResNet-50 outperformed the customized and other pre-trained CNNs in all performance metrics observed in patient-level cross-validation toward the binary task of classifying parasitized and uninfected cells studies i.e. following Kruskal-Wallis H and post-hoc analyses, the pre-trained ResNet-50 obtained the highest mean ranks for accuracy, SP, F1-score, and MCC. For the evaluation of the optimal layer for feature extraction the results demonstrated that features from shallow layers performed better than deep features to aid in improved classification (the expected optimal layer had been the final layer).

**Application:** pilot studies were initiated to analyse the performance of customized and pre-trained DL models deployed in mobile devices. This type of deployment has the potential to minimize delays in disease-endemic/resource-constrained settings. So far, it has been shown that training models and making predictions on mobile devices, or training models offline and then uploading to mobile devices is feasible (G. Howard et al., 2017).

**Example 2** (Torrecilha et al., 2017) | Protozoa | Clinical application | Predicting parasite load for *Leishmania infantum* from clinical records when qPCR services are not available using radial basis artificial neural network

**Background:** Visceral leishmaniasis (VL) is a disease caused by the protozoan parasite *Leishmania infantum* and is most prevalent in India, Bangladesh, Sudan, South Sudan, Ethiopia and Brazil (Alvar et al., 2012). Domestic dogs are highly susceptible to *L. infantum* and are the most important urban reservoirs of VL (Dantas-Torres, 2007). Common clinical and laboratory alterations are observed in dogs with VL. These alterations are associated with increased parasite load (PL) in tissues such as lymph nodes, bone marrow, spleen (Manna et al., 2009) and other peripheral tissues (Torrecilha et al., 2016). Quantification of PL load is obtained by real-time quantitative polymerase chain reaction (qPCR). Knowledge of PL in dogs with VL provides information for diagnostics, surveillance, and therapeutics follow-up.

**Problem setting:** qPCR requires costly and specialised infrastructure and technical training that is not always available commercially or in public services.

**Aim of Study:** provide a method to predict PL from clinical records (e.g. physical signs, serological test, and biochemical markers) when qPCR services are impracticable or not available.

**Training Data:** historical clinical data with known PL were sourced from 55 owned dogs located in seven distinct endemic areas, which included 35 naturally infected dogs (positives) and 20 controls (negatives). The following six input sets available for ML training were derived from historical examples of physical signs, serological test and biochemical markers: 1) nine quantitative biochemical markers (named here the BIOCHEM1 input set); 2) the same previous nine biochemical markers but indicating levels above or below reference values (named BIOCHEM2 input set); 3) four dog owner provided predictors including sex, age, vaccination and use of any form of repellent (named NULL or base model); 4) one serological test predictor indicating positivity (named ELISA); 5) six physical sign predictors including alterations to skin, eye, claw, lymph nodes, pallor of mucous membranes, and emaciation (named SIGNS1); and 6) one predictor indicating the presence/absence of at least one of the latter physical signs (named SIGNS2). In total, 30 predictors were available but 18 different combinations of the input datasets were used to train and subsequently test the prediction model e.g.
predictors representing ELISA + SIGN1 input sets were used in one training and testing prediction scenario, and ELISA + BIOCHEM1 in another prediction scenario and so on.

**Algorithm used:** Radial Basis Artificial Neural Network (RB-ANN) (Warwick et al., 1996) i.e. an artificial neural network that uses non-linear radial basis functions as activation functions. In this particular prediction problem, the precise nature of the relationships between the various predictors and PL was largely unknown. RB-ANN is a suitable method because these relationships can be approximated even in the absence of biological knowledge about the underlying phenomenon i.e. no explicit modelling is required.

**Validation of trained model:** Leave-One-Out (LOO) cross-validation was used to evaluate the training dataset.

The power of the study was assessed through bootstrap sampling of the training data e.g. the performance of the RB-ANN model was tested with varying training data sizes from 10 to 55 dogs. The expected accuracy of predictions was then computed from the average across 100 bootstrap samples for each training data size (the power assessment revealed that a minimum sample size of 43 dogs (including infected and controls) would be required to achieve a prediction accuracy of at least 80% (Torrecilha et al., 2017).

**Statistical evaluation measures used:** Accuracy of PL value predictions was measured by mean squared errors and Pearson’s correlation coefficient between observed and predicted values.

**Best model chosen:** After LOO cross-validation of the RB-ANN models generated with all 18 different combinations of the input datasets, the most accurate predictions were obtained with ELISA + SIGN2 + BIOCHEM1. In this prediction scenario, the correlation between predicted and observed PL was 0.869 with a prediction mean squared error of ±38.2 parasites per unit of volume. When input datasets were tested separately as part of the 18 combination testing, BIOCHEM2 had the highest accuracy (0.718) and the base model was the least reliable (0.192).

**Application:** RB-ANN is a proposed alternative to the more financially and technically constraining qPCR. This proposal in comparison only requires an ordinary laptop computer and example data. RB-ANN also makes reasonable estimates in lymph node PL from a relatively small training set of clinical data. Importantly, RB-ANN may be the only option available to estimate parasitism in dogs in the types of countries where *Leishmania infantum* is most prevalent.

**Example 3** (Mathison et al., 2020) | Protozoa | Clinical application | Gastrointestinal parasite diagnosis using binary image classification

**Background:** examination of stained faecal smears remains the gold standard for diagnosing infections with gastrointestinal protozoa.

**Problem setting:** The accuracy of diagnosis when examining stained faecal smears for gastrointestinal protozoa depends on the skill of a trained microscopist. Manual examination of stained faecal smears is also time intensive and laborious, and in some instances no parasites are observed despite the time dedicated to examining the slide. This led to the application of a trained CNN model to examine scans of stained faecal smears for protozoa with the goal of excluding slides that are truly negative (Mathison et al., 2020). The challenge is that gastrointestinal protozoa encompass numerous taxa possessing a plethora of morphologic characteristics, where different life cycle stages of the same
species may also differ greatly in their morphology. In some cases, different individual cells of the same taxon and life-cycle stage can also be highly pleomorphic. Protozoa must be distinguished from other non-pathogenic gastrointestinal flora and/or artefacts such as plant and animal cells consumed by the patient, which may resemble also protozoa.

**Aim of Study**: cross-validation of a pre-trained CNN model as a feature extractor toward classifying trichrome stained faecal smears as containing protozoa or not containing protozoa.

**Training Data**: 127 slides reported as positive based on manual human examination were used to train the CNN model. Slides contained various gastrointestinal protozoa including *Giardia duodenalis*, *Blastocystis* sp., *Dientamoeba fragilis*, various *Entamoeba* spp., *Chilomastix mesnili*, *Endolimax nana*, *Iodamoeba buetschlii* in addition to preparations made from human red blood cells, white blood cells and yeasts. Thousands of individual images of these parasite and non-parasite cell types (classes) were manually identified from these slides and used to train the model.

**Algorithms used**: All layers in the model were trained using the TensorFlow Object Detection API developed by Google, which implements TensorFlow and Keras for CNN training and execution (Mathison et al., 2020).

**Validation of trained model**: the primary validation metric used in this study was determination of slide-level accuracy, where specimens containing parasites as determined by a human microscopist were identified by the model as containing parasites (true positives) and those not containing parasites (or reportable levels of human white or red blood cells – which should not be present in faeces in large numbers) were not (true negatives).

**Statistical evaluation measures used**: slide level accuracy and precision.

**Best model chosen**: Only one CNN model was assessed.

**Application**: an insufficient number of high-quality exemplars of all parasite species prevented use of this CNN model for an accurate species level diagnosis. However, the high slide level accuracy (simple presence or absence of protozoa) combined with manual human examination of identified protozoa make this approach useful for augmenting traditional microscopic diagnosis by reducing the workload of microscopists.

**Example 4** (Goodswen et al., 2014) | Protozoa | Vaccine discovery | Predicting candidates against Apicomplexan pathogens using an ensemble of classifiers

**Background**: Due to the increasing volume of genes and protein sequences of eukaryote pathogens in addition to freely available bioinformatic tools, an *in silico* approach to vaccine discovery against pathogens is now feasible. Protein sequences from the target pathogen are primary input data for the *in silico* approach. Such sequences contain information/signals for predicting informative protein characteristics. However, no distinguishing signal has so far been detected that clearly indicates a protein is immunogenic and will have the desired immune response in the host. Consequently, protein characteristics only provide circumstantial evidence that a protein might be immunogenic. An *in silico* approach is therefore not an attempt to replace experimental work but a complementary one to predict which pathogen proteins among thousands are worthy of further laboratory investigation (Goodswen et al., 2013).
**Problem setting:** A typical *in silico* pipeline output is a collection of different protein characteristics (i.e. evidence) predicted by various bioinformatic programs. The goal of a researcher is to make an informed decision from the evidence as to whether a protein is worthy of further laboratory investigation. The challenge is that this evidence can be in different formats, contradicting, and inaccurate culminating in large numbers of false positive and negative decisions.

**Aim of Study:** implement an automated decision making process that reduces the number of false candidates allocated for laboratory validation. A key question to be answered was whether ML could classify potential vaccine candidates based on evidence with hidden inaccuracies.

**Training Data:** All proteins used for training data were a combination from the eukaryotic pathogens *Toxoplasma gondii* and *Neospora caninum*. Positive training data consisted of 70 proteins reported in published studies to induce immune responses; and negative training data were 70 proteins selected in accordance with the subcellular location descriptor in Universal Protein Resource knowledgebase (UniProtKB at http://www.uniprot.org/). That is, negatives were proteins with locations indicating neither membrane-association nor secretion. This is because a protein that is either external to or located on, or in, the membrane of a pathogen is assumed more likely to be accessible to surveillance by the immune system than a protein within the interior of a pathogen (Flower *et al*., 2010).

There were 13 predictors/features in total: 11 predictors obtained from the output of five bioinformatics programs that essentially predict subcellular locations; and two predictors obtained from MHC-peptide binding (i.e. T-cell epitope) prediction programs. The values of these 13 predictors represented a mixture of data types corresponding to an accuracy measure, a perceived reliability, or a type of score for the protein characteristic being predicted. All input protein sequences required for the prediction programs were downloaded from UniProtKB.

**Algorithms used:** Six supervised ML algorithms were used: adaptive boosting, random forest, k-nearest neighbour classifier, naive Bayes classifier, neural network, and support vector machine.

**Algorithm implementation:** All algorithms were executed via R functions within packages downloaded from the Comprehensive R Archive Network (CRAN): ada (adaptive boosting), randomForest, knn function contained in the Class package, naiveBayes function contained in the e1071 package, mnet, and ksvm function in the kernlab package. An R function called ‘predict’ was used as a generic function for predictions. All algorithms required at least two arguments: a data frame of categorical and/or numeric input variables (i.e. the training dataset consisting of rows of evidence) and a class vector of ‘YES’ or ‘NO’ classification for each row of evidence i.e. target variable.

**Validation of trained models:** An in-house R function executed each specific algorithm 100 times to evaluate each model built by each algorithm. For each run, the function randomly selected 70% of the training data to build a model. The remaining 30% was used as test data for classification. The prediction outcomes from the 100 runs were averaged to calculate the final evaluation performance measures.

**Statistical evaluation measures used:** Sensitivity and Specificity

**Best model chosen:** there was no apparent difference in performance (i.e. sensitivity and specificity) between the six algorithms. No algorithm could consistently classify proteins without false predictions when using the training datasets. However, a high degree of accuracy was observed when collectively
using the classifications from each algorithm and applying a voting and majority rule decision e.g. an average sensitivity and specificity of 0.97 and 0.98 respectively.

Application: An ensemble of ML algorithms was shown to perform binary classification (i.e. yes or no vaccine candidacy decisions) of thousands of pathogen proteins based on protein characteristics more accurately than human generated decisions, despite inaccuracies and inconsistencies in the input data. Vaccine candidates from an in silico approach, however, can only be truly validated in a laboratory. The application of ML can ultimately save time and money by reducing the false candidates allocated for validation.

Available from https://github.com/goodswen/vacceed/releases

Example 5 (Verma et al., 2008) | Protozoa | Drug/Vaccine discovery | New method based on amino acid composition for predicting secretory proteins from malarial parasites using support vector machines

Background: After infecting red blood cells (RBCs) in a susceptible host, *Plasmodium falciparum* (a malarial parasite) secretes an array of proteins for its growth and survival within the infected RBC. These secreted proteins are important potential drug/vaccine targets against malaria.

Problem setting: experimental identification of secretory proteins is difficult due to the complex life cycle of malaria causing parasites. Furthermore, in silico prediction based on motifs (i.e. signal sequences) has proved challenging because no universal motif exists in the known *Plasmodium* secretory proteins. It has also been shown that predicting secretory proteins based on current subcellular localization methods is currently not reliable because of multiple secretory pathways (Lingelbach & Przyborski, 2006).

Aim of Study: develop a method based on amino acid composition for predicting secretory proteins, specifically for malarial parasites.

Training Data: a non-redundant dataset of 252 (positives) identified as secretory in the literature and 252 (negatives) identified as non-secretory proteins in Swiss-Prot and PlasmoDB. Sequences from the 504 proteins were transformed to fixed length feature vectors based on amino acid composition e.g. the frequency of amino acid alanine within a protein sequence was calculated and this frequency (amino acid composition) represented one predictor. Similarly, the frequency for each amino acid was calculated such that there were 20 predictors as per 20 amino acids (i.e. the input comprised a vector of 20 dimensions/predictors per protein representing the protein’s amino acid composition).

Four different training datasets were created: 1) 504 rows * 20 predictors (based on single amino acid composition); 2) 504 rows * 400 predictors (based on dipeptide composition); 3) 504 rows * 60 predictors (based on single amino acid composition but calculated on three different parts of the protein sequence: 20 predictors representing 25 amino acids of the N terminus, 20 predictors representing 25 amino acids of the C terminus, and 20 predictors representing the remaining central amino acids); and 4) 504 rows * 400 predictors (based on position specific scoring matrix (PSSM) composition. The PSSM profile for each protein was created using PSI-BLAST – a Position-Specific Iterative Basic Local Alignment search tool that derives a PSSM from multiple sequence alignments (MSA). An MSA in the form of a PSSM in essence represents the evolutionary conservation of an amino acid at a sequence position.
Algorithm used: Support vector machine (SVM).

Algorithm implementation: SVMlight, an implementation of SVM written in C (http://svmlight.joachims.org/). Linear, polynomial and RBF kernels were evaluated but RBF generated the best results.

Validation of trained models: All models were evaluated using 5-fold cross-validation technique.

Statistical evaluation measures used: Sensitivity, Specificity, Accuracy, and Mathew's correlation coefficient (MCC).

Best model chosen: The PSSM trained model outperformed the 3 other trained models with a 92.66% accuracy. The results showed that an MSA provides more ML information than the sequence itself. Also, the dipeptide trained model performed better (86.45% accuracy) than the single amino acid composition model (84%). The model based on splitting the amino acid composition into three parts had an accuracy of 88.22%.

Application: As an outcome of this study, a web server called PSEApred was developed for predicting secretory proteins of malarial parasites: http://crdd.osdd.net/raghava/pseapred/

Example 6 (Sinha et al., 2017) | Protozoa | Drug/Vaccine discovery | Identify potential targets against Leishmania parasites using Naïve Bayes Probabilistic Classifier

Background: Visceral leishmaniasis (VL) is a major health threat and fatal disease for humans caused by Leishmania parasites. The disease is prevalent in several parts of the world but mainly endemic in North-East India. There is an urgent need to identify novel therapeutic/vaccine targets against VL, especially due to the increasing resistance of parasites to current drug therapies.

Membrane proteins are considered potential targets. A previous study (Kumar et al., 2015) conducted a sub-proteome analysis of membrane-enriched protein(MEP) fractions of Leishmania donovani. Analysis of 95 MEP spots from a two dimensional (2-D) gel image through Matrix Assisted Laser Desorption/Ionization-time of flight mass spectrometry (MALDI-TOF/MS) identified and annotated 72 proteins via database searches and bioinformatic prediction programs. Annotation was mainly three biological function types: 1) actin binding, cell signalling; 2) cellular localization e.g. cytosolic, membrane; and 3) class/family e.g. Hydrolase, ATPase. The 72 proteins were classified based on the annotation in terms of their target importance. The six classes were drug resistance, drug target, immunogenic protein, Th1 stimulatory, vaccine candidate, and unknown.

Problem setting: Performing in vivo or in vitro experiments to identify potential targets is a time consuming process. An efficient automated computational approach is sought as a compelling alternative.

Aim of Study: computationally perform a binary classification of the 37 unknown of the 72 proteins identified by the 2-D gel and MALDI-TOF/MS experiment. The two classes are drug target (DT) and vaccine candidate (VC).

Training Data: consisted of 28 of the 72 proteins that were previously classified as DT or VC (i.e. the ML target labels). There were 21 DT and seven VC proteins with four predictors: molecular mass, isoelectric point, biological function (categorical data), and cellular localization (categorical data). It was assumed that each predictor independently contributed to the classification. Note that four of the
21 DT labelled proteins were also classified drug resistant in a previous study (Kumar et al., 2015), which potentially introduces noise.

**Algorithm used:** Naïve Bayes classifier. However, RF; decision tree (C4.5) and SVM were also evaluated.

**Validation of trained models:** All trained models were evaluated using 10-fold cross-validation and evaluations were repeated 10 times (i.e. 100 runs per algorithm and the results averaged)

**Statistical evaluation measures used:** Accuracy, Mean Absolute Error, and RMSE

**Best model chosen:** Naïve Bayes (NB) classifier outperformed all other algorithms evaluated. The accuracies were 76.17% (NB), 73% (RF), 63% (SVM) and 56.33% (C4.5).

**Application:** NB classifier was trained on entire 28 protein training dataset to classify the 37 unknown instances. Additional DTs were predicted targets, which could be explored further as novel therapeutics against VL.

**Example 7 (Maindola et al., 2015)** | Protozoa | Drug discovery | Identifying compounds in traditional Chinese medicine to use as anti-malarials using various binary classification algorithms

**Background:** Artemisinin-based combination therapy (ACT) or quinine-based compounds are the recommended chemotherapeutics by the World Health Organization (WHO) for the treatment of malaria. The challenge is that these drugs have started showing resistance (Noedl et al., 2008, Phyo et al., 2012, Ashley et al., 2014) and therefore anti-malarials with different mechanisms of action are urgently needed.

The causative agent of malaria is the Apicomplexan *Plasmodium*. Invasion of *Plasmodium* into erythrocytes is mediated by proteins secreted by apical organelles of the parasite, micronemes and rhoptries. Rhoptry Neck (RON) is a protein secreted from the neck region of rhoptries; and antigen (AMA1) is a protein secreted from micronemes.

**Problem setting:** AMA1 forms a complex with RON2 as part of the moving junction that develops between the host cell and the invading parasite (MacRaild et al., 2011). It has been suggested that disruption of AMA1-RON2 interaction can be effectively used in combination with existing anti-malarials to block host cell invasion (MacRaild et al., 2011).

**Aim of study:** Identify compounds in traditional Chinese medicine that can potentially be used as anti-malarial by inhibiting AMA1-RON2 interaction.

**Training Data:** derived from a biological assay targeting AMA1-RON2 interaction downloaded from PubChem database at the National Centre for Biotechnology Information (NCBI) (Wang et al., 2009). The assay consisted of compounds available in the NIH Molecular Libraries Small Molecule Repository (MLSMR) that were tested for their capability of inhibiting AMA1-RON2 interaction. The chemical structures of active and inactive compounds were downloaded as SDF format (SDF is one of a family of chemical-data file formats developed by MDL Information Systems). The downloaded data were then converted to molecular descriptors. ‘The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment’ (Consonni & Todeschini, 2010). For input requirements to ML algorithms, the molecular descriptors
of the compounds were presented in a single CSV (comma separated values) file consisting of a total of 179 2D-descriptors per line. These descriptors were a combination of numerical and binary values corresponding to molecular features: 147 were Pharmacophore fingerprints (based on bioisosteric principles), and 24 were Weighted Burden (based on burden connectivity matrix. In effect, these values are representative properties considered useful for judging the drug-like nature of a molecule. In ML terms, the 179 descriptors are features/predictors. An additional column was added to the CSV file showing ‘Active’ or ‘Inactive’ as an indication of the bioactivity of each row of compound values. This column is the predicted or target label. Therefore in ML context, this real life example is using labelled data consisting of both features (i.e. the 179 descriptors) and the label (i.e. active or inactive) and embodies supervised training.

The CSV file was split such that 80% and 20% of the rows were for training cum internal validation and external validation/testing, respectively. These splits in effect represented 588 instances of active (positive training data) and 284,968 instances of inactive (negative training data) molecules for training; and 147 active and 71,241 inactive molecules for validation.

**Algorithms used:** Naive Bayes classifier (NB), Random Forest (RF), SVM (linear kernel), and J48 (an open source Java implementation of the C4.5 decision tree algorithm) were trained to build classification models.

**Algorithm implementation:** Weka workbench was the environment chosen to run the algorithms and analyse results.

**Validation of trained model:** All classification models were built upon the training cum internal validation set with a stratified 5-fold cross-validation. ‘Stratified’ refers to the splitting of data into folds (groups) in such a way that each fold has the same proportion of observations with a given categorical value e.g. in this example, 0.21% are active and 99.79% are inactive instances in the full training dataset and these same class proportions were kept in each fold.

**Statistical evaluation measures used:** TPR, FPR, SN and SP, ROC and AUC (note that all these measures are appropriate for an imbalanced classification problem, unlike ACC which may yield misleading results with imbalanced data). To also account for this imbalance, a balanced classification rate defined as geometric mean (G-mean) of sensitivity and specificity was used (Kubat et al., 1997).

**Complications:** The training dataset was highly imbalanced (i.e. 0.21% active and 99.79% inactive instances) resulting in skewed predictions with an excessively high false negative rate. To improve predictions, Cost Sensitive Classifiers (Japkowicz, 2000) acted as a meta classifier by penalising false negative classifications made by the base algorithms (NB, RF, SVM, J48).

**Best model chosen:** RF on the basis of best statistical evaluation measures (SN = 0.82, SP =0.80, G-mean = 0.81, and AUC = 0.86). This model was further validated on a test dataset of 17 known anti-malarial drugs from Drug Bank (Wishart et al., 2006) and predicted 11 as active.

**Application of trained model:** the best model (RF) was used to classify compounds in the Traditional Chinese Medicine (TCM) database hosted at ZINC repository (Chen, 2011) with the aim to identify possible new inhibitors of AMA1-RON2 interaction and ideally, identify a TCM compound (or its source herb) that has been used for anti-malarial treatment.
In a similar way to the manipulation of data used for training, molecular descriptors were generated from the TCM data and then presented in a single CSV file. All compounds predicted active by the best model were arranged by their score in a probability distribution. The best model classified 216 compounds from TCM as AMA1-RON2 inhibitors after setting a cut-off threshold of 1.0 within the probability distribution. However, only 32 of the 216 passed the software filtering process used to detect potential toxic compounds. Two of the 32 compounds have herbal sources in use as antimalarials in traditional medicinal pharmacopeia (Lohombo-Ekomba et al., 2004, Ye et al., 2013).

Example 8 (Levatic et al., 2018) | Protozoa | Drug discovery | Identifying derivatives of Primaquine (antimalarial drug) that will retain or improve its efficacy, while reducing its toxicity using QSAR modelling – a regression problem

Background: Plasmodium falciparum is one of the five Plasmodium species that cause malaria in humans, but is the species most responsible for malaria-associated mortality worldwide. Primaquine (PQ) is a commonly used antimalarial drug active against all species causing human malaria, including multi-resistant P. falciparum strains. PQ has a unique ability to prevent P. falciparum transmission of infection to mosquitoes by killing the gametocytes produced during the sexual life stage of the parasite in the blood (Levatic et al., 2018).

IC_{50} is the concentration of the tested compound required for 50% growth inhibition of P. falciparum

Problem setting: PQ exhibits toxicity towards a large subset of patients with a glucose-6-phosphate dehydrogenase (G6PD) enzyme deficiency (White, 2013). This G6PD deficiency is a widespread genetic trait, particularly in malaria-endemic regions.

Aim of Study: identify derivatives of PQ that will retain or improve its efficacy, while reducing its toxicity. The study’s approach was to systematically link structural features of PQ derivatives to antiplasmodial activity by applying a quantitative structure-activity relationship (QSAR).

Training Data: a set of 56 compounds consisting of antimalarial activity measures from 23 PQ derivatives (sourced from the study’s in vitro activity experiments on 64 diverse PQ derivatives against the erythrocytic stage of the drug-sensitive P. falciparum NF54 strain) and 33 compounds (sourced from published data (Kaur et al., 2011, Kaur et al., 2011, Kaur et al., 2012)). The 56 compound structures were initially represented as SMILES strings. SMILES is a line notation for describing molecular structures using short ASCII strings (Toropov et al., 2005). The SMILES strings were preprocessed using an application called Instant JChem Standardizer (Kind, 2007). The 56 structures were then represented by 140 2D molecular descriptors calculated with Chemistry Development Kit (Steinbeck et al., 2003). These molecular descriptors are effectively the features/predictors for ML algorithms i.e. the ML training data was a table comprising 56 rows and 140 columns (features) with the antimalarial activity (log IC_{50}) as the target regression value.

Algorithms used: QSAR modelling. QSAR models are typically regression or classification. For the purpose of predicting antimalarial activity (log IC_{50}) against P. falciparum NF54 strain, the QSAR model was trained using the SVM algorithm for regression i.e. the QSAR regression model relates the set of predictor variables (the numerical descriptions of molecular structure) to the potency of the response variable (the biological activity of the molecule). In effect, the QSAR model summarises a supposed relationship between molecular structures and biological activity using the training dataset and can be used to predict the activities of new molecules.
A forward feature selection method was used to highlight which of the 140 molecular descriptors were the most important. This is a method whereby a model is iteratively trained and evaluated with subsets of features until the best subset is selected. For instance, let ‘n’ equal the number of features. On the first round, ‘n’ models with an individual feature are evaluated and the best predictive feature is selected (e.g. if there are five features, then five models, each containing only one feature, are evaluated. For this example, feature three has the best metric and is selected). On the second round, ‘n-1’ models with each feature and the previously selected feature are evaluated (e.g. four models are evaluated that contain features three and one, three and two, three and four, three and five). The rounds are repeated until the best subset of features is obtained (e.g. if the subset three and four performed the best; then the next round evaluates three models containing features three, four and one; three, four and two; and three, four and five. For the study, a cross-validation Pearson correlation coefficient was used to measure the relative importance of the features/descriptors.

**Algorithm implementation:** LIBSVM, which is a software library for support vector machines (Chang & Lin, 2011). The radial basis function (RBF, or Gaussian) kernel was used (as recommended by LIBSVM authors). There are two parameters for an RBF kernel: C and γ, which are not known beforehand. A grid search procedure was used on the cross-validation data to find the optimum parameters C and γ for the given problem e.g. various pairs of (C,γ) values from exponentially growing sequences of C and γ were tried and the best pair with the best cross-validation accuracy was picked (where C was from $2^{-5}$, $2^{-4}$, …, $2^{20}$ and γ from $2^{-15}$, $2^{-14}$, …, $2^{5}$).

**Validation of trained model:** The models were evaluated by using 10-fold cross validation which was repeated 10 times with different random initialization. Furthermore, the SVM model was evaluated against an independent set of 37 compounds (i.e. out-of-sample data) whose antimalarial activity were experimentally determined in previous studies (Pavić et al., 2014, Pavić et al., 2016, Perković et al., 2016).

**Statistical evaluation measures used:** coefficient of determination ($R^2$). In this case, $R^2$ was used as a statistical measure from 0 to 1 of how well the predicted antimalarial activity (log IC$_{50}$, μM) approximated the observed antimalarial activity (log IC$_{50}$, μM) given the training data. Also, root-mean-square error (RMSE) was used, which is the standard deviation of the prediction errors (residuals) i.e. it provides an indication of how spread out the residuals are from the regression line such that when $R^2$=1, RMSE = 0 as all data points lie on the line with no residuals.

**Best model chosen:** The highest evaluation measures when predicting activity (log$_{10}$ IC$_{50}$ units) against *P. falciparum* following cross-validation testing were $R^2 = 0.785$ and RMSE = 0.284 obtained with $C = 2^5$ and $γ = 2^{-9}$. However, more favourable parameter settings ($C = 2^1$, $γ = 2^{-6}$) were adopted with a predictive performance of $R^2 = 0.776$ and RMSE = 0.294. Lower values of C and γ were considered to result in simpler, more parsimonious models that are less prone to overfitting. The prediction accuracy based on the independent set of 37 PQ derivatives using the same trained SVM model yielded a RMSE = 0.403 log$_{10}$ units of IC$_{50}$. Additionally, the initial training set of 56 compounds was combined with the 37 derivatives to create a new SVM model. Following a C and γ parameter optimization procedure on these 93 compounds, the cross validation evaluation measures were $R^2 = 0.729$, RMSE = 0.312 using $C = 2^5$ and $γ = 2^{-7}$.

**Application of trained model:** the adopted SVM model was applied to infer the antimalarial activity of an in-house virtual library of 522 molecules. Seven PQ derivatives were identified based on their high predicted antimalarial activity and other considerations related to chemical synthesis. These PQ
derivatives were selected for synthesis and evaluation against \textit{P. falciparum} and an initial cytotoxicity screen using the L6 mammalian cell line derived from rat skeletal myoblasts. Furthermore, the 93-compound SVM model was used on 13,401 PQ-like molecular structures extracted from PubChem to predict antimalarial activity. The model predicted 199 compounds to be both more active and less cytotoxic than PQ itself. A caveat, however, came with these predictions as QSAR models are less reliable the more the PubChem molecular structures diverge from the training set of PQ derivatives.

\textbf{Example 9} (Plucinski \textit{et al.}, 2015) | Protozoa | Epidemiology and antimicrobial resistance | Distinguishing between a relapse of illness due to failure of antimicrobial treatment and infection with a new malaria strain

\textit{Background:} Antimalarial resistance in \textit{Plasmodium falciparum} is an increasing problem in endemic regions (Plucinski \textit{et al.}, 2015, Talundzic \textit{et al.}, 2016, Halsey \textit{et al.}, 2017). Assessing antimalarial efficacy involves genotyping malaria parasites when an infection is first identified and treated, followed by genotyping parasites from the same patient if they become ill with malaria again, by sequencing a set of well-defined \textit{P. falciparum} microsatellite repeats. Comparison of microsatellite profiles before and after treatment enables assessment of whether patients remained infected with the original strain due to a treatment failure, or if they are infected with a new strain (Plucinski \textit{et al.}, 2015).

\textit{Problem setting:} Manually performed human classification of malaria microsatellite profiles as a recrudescence or reinfection can lead to gross under- or over-estimation of drug efficacy due to the introduction of human bias (Jones \textit{et al.}, 2020). Furthermore, lower frequency alleles may not be detected at each sampling using current laboratory protocols, meaning that an undetected presence may be misinterpreted by a human examiner as the true absence of an allele (Plucinski \textit{et al.}, 2015).

\textit{Aim of Study:} To assess the performance of a Bayesian classifier when classifying \textit{P. falciparum} infections as a recrudescence (treatment failure) or reinfection with a new strain on the basis of their microsatellite profile.

\textit{Algorithm used:} unsupervised Bayesian classifier.

\textit{Statistical evaluation measures used:} Classification failure rate (rate of misclassification) on simulated data of known classification label (Jones \textit{et al.}, 2020).

\textit{Application:} Assessment of antimalarial drug efficacy in malaria endemic regions. Specifically, the algorithm generates a posterior probability that a recurrent infection is a recrudescence using a microsatellite profile as input.

\textbf{Example 10} (Perdiguer-Alonso \textit{et al.}, 2008) | Protozoa | Taxonomy | Fish discrimination given parasite community data as biological markers using random forest (a multi-classification problem)

\textit{Background:} Fish stocks are subpopulations of the same fish species and each stock have geographic distribution limits due to their specific tolerances to the surrounding environment. Determining the geographic regions of harvested individual fish is an important requirement by fisheries management to detect illegal fishing, protect endangered commercial fish, and/or avoid legal fishing disputes between countries. One method of stock discrimination/ separation is using fish parasites as biological markers (or tags). This is because local environmental factors can regulate the survival and transmission success of parasite infective stages, thus causing fish in different regions to contain
different parasite populations (Timi, 2007). Discrimination, in this context, refers to the identification of members of different stocks in the catches of mixed aggregations (Perdiguero-Alonso et al., 2008).

**Problem setting:** discrimination using parasites as biological tags is difficult because parasite communities in fish are influenced by many factors such as fish behaviours and feeding habits, and immunological responses to parasites (Timi, 2007).

**Aim of Study:** use a ML technique in the application of fish discrimination given parasite community data as biological markers.

**Training Data:** Source of parasite data was sampled over a two year period from 763 individual fish from cod populations located at five spring (spawning) and autumn (feeding) geographic regions in the North East Atlantic. These five sampling regions constituted the five classification targets: Baltic, Celtic, Irish Sea, North Sea, and Icelandic waters. From the parasite data, the prevalence of 31 parasite species within the fish host was determined. The training data, therefore, consisted of 763 rows with 31 predictors (independent variables) and a sampling region (dependent variable/ target label). Note that sample sizes from each region, season and year were not equal.

**Test Data:** A ‘blind’ mixed sample of 50 fish collected in spring.

**Algorithm used:** Random Forest (RF). This algorithm was particularly chosen here because it makes no assumptions concerning input data towards normality or independence, handles noisy data and/or data containing many zeros, less prone to overtraining than other ML algorithms, and can indicate the importance of individual variables towards classification (Perdiguero-Alonso et al., 2008). Two other algorithms, LDA and artificial neural network (ANN), were also evaluated using the same parasite community data, essentially to support the choice of RF.

**Algorithm implementation:** randomForest package of the statistical software R.2.5.1 (the only two input parameters used were ‘number of trees = 2000’ and ‘number of independent variables used at each split = 5’).

**Validation of trained models:** Five experiments were performed. Experiment 1 comprised 50 runs (model creations). For each run the training dataset was randomly split into 80% for training and 20% for validation. The results were averaged over these 50 independent models. Experiment 2 comprised a 10-fold stratified cross-validation on the entire dataset. Experiment 3 was as per experiment 2, except the predictors used were selected on the basis of variable importance shown in Experiment 1. Experiment 4 was as per experiment 2 but with a reduced training dataset of only fish within a standard length range. Experiment 5 comprised different training and test sets taken from the training dataset but based on seasonal and annual sampling variations e.g. Spring 2002 samples as the training set and autumn 2002 samples as the validation set.

**Statistical evaluation measures used:** classification error rate (in this case, the proportion of fish incorrectly classified in a geographic region, averaged over multiple runs).

The test data (i.e. the blind mixed sample) were evaluated by means of the McNemar test and the performance measures: accuracy, precision, recall, and F-measure.

**Best model chosen:** comparison between classification error rates from the first three experiments showed negligible differences, suggesting the different approaches in selecting the validation sets did
not affect the trained models, and reducing the number of variables (experiment 3) did not improve the accuracy. Experiment 4 generated the lowest classification error rate and therefore deemed the best model. This model, by having a standard fish length range, in effect removed the younger and older fish during the 10-fold stratified cross-validation. The purpose of experiment 5 was to mimic real world situations e.g. provide indication of the generalisation abilities of the trained models for possible seasonal/annual effects on parasite community variation. The classification errors were substantially worse than all previous experiments. However, this is expected due to the reduced training data.

Results from all five experiments showed that the classification accuracy for some regions was consistently more accurate than for other regions. The study suggests that accuracy here is indirectly determined by whether a classification geographical region possesses all suitable conditions for the completion of a parasite’s life cycle i.e. a fish can become infected with a given parasite only when it is within the endemic area (MacKenzie et al., 2008). Therefore, poor accuracy may be a reflection of the relative positioning of the selected sampling regions e.g. some regions may overlap in terms of endemic area.

The ‘blind’ sample data was used to compare the performance of the three algorithms: RF, LDA and ANN. Overall, RF outperformed the other algorithms.

**Application:** The study concluded that parasite community data can be used successfully to discriminate individual cod from different geographical regions using RF. Good discriminatory results were obtained by RF despite cod being a migratory fish species with largely overlapping parasite faunas (i.e. noisy data). This RF approach is therefore a promising tool in the development of sustainable harvest and monitoring strategies by fisheries management.

**Example 11** (Yang et al., 2018) | Bacteria | Antimicrobial resistance | Identification of drug-resistant tuberculosis (TB).

**Background:** drug-resistant tuberculosis (TB) is a major concern for global public health. Rapid, but accurate, identification of this resistance is essential for TB control.

**Problem setting:** existing methods identify drug resistance based on the presence of a single, well-studied nucleotide polymorphism (SNP). Latest research proposes exploring multivariate association between genetic variants as a more appropriate, although challenging, identification method (Zhang et al., 2013, Walker et al., 2018).

**Aim of Study:** explore multivariate association with different ML models to classify drug resistance against eight anti-TB drugs and to classify multi-drug resistance.

**Training Data:** source data derived from 23 gene sequences of 1839 Mycobacterium tuberculosis isolates (Walker et al., 2018). Each isolate underwent drug-susceptibility testing to a maximum of 11 anti-TB drugs. Two classes of isolates were determined: resistant and susceptible. The presence or absence of a SNP in an isolate sequence was represented in the training data by 1 or 0 respectively. In the 23 candidate genes of the 1839 isolates, 2629 SNPs were found. Three sets of features, F1–F3, where F1 contained all SNPs found e.g. training data comprised 1839 rows * 2629 predictors (binary variables). F2 contained 108 SNPs previously reported as resistance-determinants (Walker et al., 2018), and F3 contained a subset of F1 where genes with only resistance-determinants to a particular drug were included.
**Algorithms used:** LR, SVM, RF, product-of-marginals model (PM) (Yang et al., 2018), and class-conditional Bernoulli mixture model (CBMM) (Yang et al., 2018).

**Algorithm implementation:** MatLab for LR, SVM, and RF.

**Validation of trained models:** randomly selected equal number of class members and then an 80%, 20% split for training and testing respectively. A 5-fold cross-validation performed during training.

**Statistical evaluation measures used:** SN, SP and AUC

**Best model chosen:** PM and SVM (with a radial basis function kernel). These models outperformed existing methods in terms of sensitivity to resistance classification. Evaluation comparisons between F1–F3 highlighted that the resistance identification for some drugs is more challenging than others.

**Example 12** (Burckhardt et al., 2019) | Bacteria | Epidemiology and Antimicrobial resistance | Confirmation of *Streptococcus pneumoniae* capsular serotypes on the basis of fourier-transformed infrared spectra

**Background:** *Streptococcus pneumoniae* colonizes the nasopharynx of healthy individuals though may disseminate in the context of a compromised immune system (Mitchell & Mitchell, 2010). The *S. pneumoniae* polysaccharide capsule is a major cell surface structure, and differentiation of capsular serotypes is important for molecular surveillance (Gonzales-Siles et al., 2019). Traditionally, the Quellung reaction is used to differentiate serotypes, involving a reaction between antibody and its specific capsular polysaccharide motifs, causing cells to swell; a microscopically visible process.

**Problem setting:** The Quellung reaction test is relatively expensive and its use is typically restricted to reference laboratories (Burckhardt et al., 2019). Laboratories performing *S. pneumoniae* capsular serotype surveillance would benefit from the ability of rapid, more accessible serotyping procedures.

**Aim of Study:** To determine whether hierarchical clustering of FT-IR spectra can be used as an alternative to the Quellung reaction for identifying *S. pneumoniae* capsular serotypes.

**Algorithm used:** hierarchical clustering using the Euclidean—average—mean spectra algorithm.

**Statistical evaluation measures used:** Concordance between results of the Quellung reaction and clustering based on FT-IR spectra (Burckhardt et al., 2019).

**Application:** Hierarchical clustering of FT-IR spectra as an alternative to the Quellung reaction, that is accurate, rapid and potentially more accessible to laboratories wishing to serotype the *S. pneumoniae* based on the capsular membrane of clinical isolates.

**Example 13** (Ramirez et al., 2018) | Bacteria | Macroeocology | Assessing bacterial communities across global soils using supervised and unsupervised random forest.

**Background:** merging microbial data is a potential opportunity to address global-scale questions such as predicting the response of soil organisms to global environmental change (Garcia-Palacios et al., 2015). Furthermore, variation in microbial community structure is possibly more ecologically relevant than measures of diversity and abundances of major taxa (Ramirez et al., 2018). Community structure is largely defined by the relative abundances of individual taxa.
Problem setting: microbial community data are mainly disseminated in disparate published studies with inherent biases (e.g. different sampling and sequencing methodologies).

Aim of Study: apply ML to assess global soil microbial community patterns from merged independent taxonomy-based data sets.

Training Data: metadata were collected from 30 high-throughput sequencing (HTS) studies comprising 1998 individual soil samples containing 8287 bacterial taxa from 21 countries, representing all continents except Antarctica. Metadata included properties such as soil measurements, geographical location, and sampling and sequencing technical information. Datasets were merged using the taxonomic affiliations of individual operational taxonomic units (OTUs). Relative abundances of OTUs were determined per sample. These abundances were used as the ML predictor values i.e. training data consisted of 1998 rows (samples) and 8287 columns (OTUs). An extra column represented the target label containing the study reference number (STN) of the sample (STN is also the link to the metadata).

Algorithms used: unsupervised RF was used to test for the importance of different taxa in separating communities i.e. the importance of each taxon was quantified in a model that separated observed data from synthetic data. Synthetic data was randomly drawn from the observed distributions of relative abundances for each taxon i.e. a randomized distribution. Supervised RF was used to identify the impact of biases within disparate studies in order to assess the value of data integration i.e. a model was constructed that classified the study from which a sample was taken based on the abundances of the taxa it contained.

Algorithm implementation: RandomForest package for R.

Validation of trained models: k-fold cross-validation for supervised RF (‘k’ value not specified).

Statistical evaluation measures used: error rate for supervised RF.

Finding: despite biases between studies, many bacterial taxa are still informative about community structure. Rare bacterial taxa were more important than abundant taxa in differentiating communities.

Example 14 (Sohn & Li, 2018) | Bacteria | Microbial ecology and microbiomes | Distinguishing microbiome signatures collected from different body sites based on their bacterial 16S rDNA sequence composition.

Background: Ordination methods (i.e. dimensionality reduction procedures) are widely used to distinguish microbial communities based on bacterial 16S sequence composition (Hawinkel et al., 2019). Unfortunately, bacterial 16S microbiome datasets sometimes suffer from different dispersion effects and high statistical sparsity which can greatly impact the interpretation of results following dimensionality reduction using some ordination procedures.

Problem setting: Many ordination methods often do not account for negative dispersion effects and high statistical sparsity. Furthermore, failure to detect taxa (resulting in a zero count for that taxon), may represent a true absence or an undetected presence, though is considered a true absence by many ordination methods.
Aim of Study: To assess the performance of a generalized linear model (GLM)-based ordination procedure that can differentiate 16S rDNA profiles generated from bacterial communities in different microbial niches while accounting for negative dispersion effects and statistical sparsity.

Algorithm used: GLM-based Ordination Method for Microbiome Samples (GOMMS).

Statistical evaluation measures used: The GOMMS procedure was compared to other commonly used distance-based ordination procedures applied to 16S rDNA datasets generated from the nasopharynx and oropharynx (M & Li, 2018). The ability to differentiate microbial communities from these two groups was assessed for each ordination procedure, and GOMMS was found to perform better on datasets suffering from negative dispersion effects (M & Li, 2018).

Application: Differentiation of bacterial communities occurring in different microbial niches on the basis of their 16S rDNA profile.

Example 15 (Midani et al., 2018) | Bacteria | Microbiome | Predicting susceptibility to Vibrio cholerae infection

Background: observational studies have identified several clinical and epidemiologic risk factors that correlate with susceptibility to Vibrio cholerae infection in household contacts of patients with cholera (Harris et al., 2008)

Problem setting: an abnormal human gut–associated bacterial community (microbiota) is hypothesised to be another susceptibility factor(Kamada et al., 2012).

Aim of Study: demonstrate that microbiota is associated with susceptibility to V. cholerae infection.

Training Data: sourced from samples taken from 76 household contacts. Three model types: microbiota, clinical, and combined. Predictors for microbiota model were relative abundances of operational taxonomic units (OTUs) selected from the baseline microbiota of contacts. There were 4181 unique OTUs in the contacts. Predictors for clinical model included age, blood group O status, and vibriocidal antibody titer. Combined model contained both microbiota and clinical model predictors. Two class labels: infected (n = 22), uninfected (n = 54).

Algorithms used: SVM with a recursive feature elimination (RFE) algorithm. RFE recursively works out the optimum combination of features (in this case OTUs) for improving SVM.

Algorithm implementation: SVM and RFE through Scikit-learn.

Validation of trained models: Hold-out models were trained on 48 household contacts and tested on 28 contacts recruited at a later date. Also, 30 replicates of 10-fold cross-validation, and randomly splitting contacts into training (90%) and testing (10%) sets.

Statistical evaluation measures used: AUC, FPR, TPR.

Best model chosen: combined model, however the microbiota model outperformed the clinical model and was only marginally less accurate than the combined. The optimal microbiota model needed only 143 OTUs out of 4181 as determined by RFE.
Findings: cholera susceptibility can be characterized by differences in microbiota of infected and uninfected individuals. These differences can be detected using ML.

Example 16 (Poore et al., 2020) | Bacteria | Microbiome | Cancer diagnostics using microbiome analyses of human blood and tissues

Background: Traces of DNA and RNA from microorganisms can be found in various human tissues and blood, including within or around tumour cells (Robinson et al., 2017).

Problem setting: The extent and diagnostic implications of microbial contributions to different types of cancer remain unknown.

Aim of Study: To discriminate between cancer types and between different stages of the same cancer type, as well as between cancer and normal tissue using microbial cancer signatures (i.e. microbial community structure and dynamics).

Training Data: sourced from sequences representing 18,116 samples across 10,481 patients and 33 cancer types held in The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) online resource. The TCGA sequencing data were processed by Kraken (Wood & Salzberg, 2014) (an ultrafast program for assigning taxonomic labels to metagenomic DNA sequences). Discrete taxonomical counts (i.e. microbial abundances) were converted into log-counts per million per sample using Voom (Law et al., 2014) (i.e. Voom is a program designed to transform RNA-Seq data Ready for linear modelling) and supervised normalization was performed by the SNM (Mecham et al., 2010) R package (the goal of normalization is to separate biologically meaningful signal from other confounding signal sources such as unavoidable technical factors introduced during data collection e.g. batch effects (Mecham et al., 2010)). In summary, the filtered and normalized data prior to training comprised ‘summarized read counts at the genus taxonomic level’ of 17,625 samples, 10,183 patients, and 32 cancer types.

Algorithms used: Stochastic gradient-boosting ML (GBM) models (Friedman, 2002) (models that construct additive regression models).

Algorithm implementation: R packages GBM and Caret (Kuhn, 2008) (short for classification and regression training).

Validation of trained models: Training and testing occurred on separate, randomly selected, stratified sampling splits of 70% and 30% of the data, respectively. The 70% training data was used with 4-fold cross validation to determine optimum GBM parameters. A fixed random number seed was used to ensure reproducibility and comparability of model results e.g. set.seed(102). Final model performances (i.e. statistical evaluation measures on fully optimised trained model) were generated by applying the final model to the unseen 30% holdout test set.

To evaluate generalizability across data sets, raw TCGA microbial count were randomly sorted into two batches (split #1: \( n = 8,814 \); split #2: \( n = 8,811 \) samples). Each batch was independently normalised by Voom-SNM, trained and optimised, and then tested on the other normalised batch. A third ML model containing all normalised samples (\( n = 17,625 \)) was used in 50–50% training and testing splits.

Statistical evaluation measures used: ROC curve (receiver operating characteristic curve), PR curve (precision–recall curve), AUROC (area under the ROC curve) and AUPR (area under the PR curve)
calculated using the PRROC (Grau et al., 2015) R package; and confusion matrices calculated using the Caret (Kuhn, 2008) R package.

**Findings:** Overall, the models performed well in discriminating between cancer types and between cancer and normal tissue (this suggests that microbial communities are unique to each cancer type). However, the models performed less well when discriminating between different stages of the same cancer type as defined by host tissue (this suggests that microbial community structure dynamics may not correlate with cancer stages for all types of cancer) (Poore et al., 2020). Importantly, these results suggest that the ML proposed strategy in this study has the potential to be a microbiome-based cancer diagnostic tool that may complement existing circulating tumour DNA (ctDNA) assays for detecting and monitoring cancer.

**Example 17** (Colubri et al., 2016) | Viruses | Clinical applications | Ebola Virus Disease (EVD) prognosis prediction using various classification algorithms

**Background:** Clinicians currently favour scoring protocols such as Ebola Prognostic Score (EPS) based on symptom counts to estimate an EVD patient’s mortality risk. Viral load (PCR) is the primary patient prognosis indicator (Colubri et al., 2016).

**Problem setting:** Recent serious outbreaks of EVD indicate the need for improved patient prognosis tools for healthcare personnel in the field. Available EVD patient records with known outcomes are limited and often incomplete including no PCR measurements (Colubri et al., 2016).

**Aim of Study:** Predict outcome of Ebola patients from their initial clinical symptoms.

**Training Data:** A previous study (Schieffelin et al., 2014) provided clinical symptoms and laboratory tests of 65 patients with known outcomes for EVD. Only 21 of 65 confirmed cases had complete records. Set percentages of missing values were imputed using three Multiple Imputation programs. Exploratory analysis with Mirador (http://fathom.info/mirador/) identified 24 clinical and laboratory factors (predictors) that showed association with EVD outcome e.g. PCR, temperature, diarrhea. MINE (http://www.exploredata.net/) ranked the 24 associations. Two informative subsets were created containing 10 top ranking predictors each; one with PCR and the other without. Each ML algorithm was trained on all possible combinations of predictors from the PCR and non-PCR.

**Algorithms used:** ANN, LR, DT, and SVM

**Algorithm implementation:** ANN using an in-house python script; LR, DT, and SVM using Scikit-learn.

**Validation of trained models:** Each predictive model, defined by a prediction algorithm and a particular selection of input predictors from the PCR and non-PCR sets was trained and tested 100 times.

**Statistical evaluation measures used:** F1-score, AUC.

**Best model chosen:** LR and ANN.

**Application:** Ebola Care app from https://play.google.com/store/apps.
Example 18 (Zorn et al., 2019) | Viruses | Drug discovery | Identifying new compounds against HIV with various classification algorithms

Background: the most common target of human immunodeficiency virus (HIV) antiviral drugs is reverse transcriptase (RT), which was the motivation for generations of two types of drugs: nucleoside and non-nucleoside reverse transcriptase inhibitors (NRTI and NNRTI) (Zorn et al., 2019).

Problem setting: HIV has a high mutation rate (e.g. inadequate proofreading activity of RT (Svarovskaia et al., 2003)) and is predisposed to drug resistance. There is a need for multiple-target antiviral drugs.

Aim of Study: compare different ML methods with public data input to identify new compounds for testing against HIV.

Training Data: sourced from ChemDB HIV, Opportunistic Infection and Tuberculosis Therapeutics Database (https://chemdb.niaid.nih.gov/). A quantitative structure-activity relationship (QSAR) approach was followed to link chemical features to antiviral activity: cell-based EC_{50} (50% efficacy) and target-based HIV-RT IC_{50} (50% inhibition) measurements. There were 16 training and test dataset combinations e.g. specific cell-based (2991 active, 3015 inactive) and target-based (131 active, 338 inactive); and nonspecific cell-based (8687 active, 10067 inactive) and target-based (1394 active, 3356 inactive) – ‘specific’ or ‘nonspecific’ refers to whether the assay method and/or cell lines were considered or not considered for dataset preparation. RDKit (http://www.rdkit.org) was used to generate extended-connectivity fingerprints (ECFP6) (Rogers & Hahn, 2010), which essentially converted compounds to chemical features (i.e. ML predictors) for the QSAR model. The antiviral activity was the target regression value.

Algorithms used: NB, AdaBoost DT, RF, SVM (classification), kNN, ANN, and Consensus (modelling with all six algorithms).

Algorithm implementation: Scikit-learn.

Validation of trained models: a stratified 5-fold cross validation (stratified maintains similar ratios of active and inactive compounds among the splits).

Statistical evaluation measures used: Recall, Precision, Accuracy, AUC, F1-score (Lipton et al., 2014) and MCC.

Best model chosen: SVM, ANN, and Consensus (no significant difference).

Application: ML can accelerate HIV drug discovery by identifying potential drug-resistant compounds to be tested in vitro.

Example 19 (Silva et al., 2017) | Viruses | Plant microbes | Multi-classification of large volumes of Geminiviridae sequences

Background: Geminiviridae is a family of viruses divided among nine genera that infects a wide range of tropical and subtropical plants worldwide causing problematic economic losses (Silva et al., 2017). The genomes of this family collectively contain ten different known genes, which are commonly identified through open reading frames (ORFs). Advancement of high-throughput sequencing technologies has instigated massive data generation for many viruses including geminivirus.
**Problem Setting:** current genus classification is based on interpretation of factors such as insect vector, host range, and phylogeny reconstruction. Identification of ORFs requires expertise in using tools such as ORF finder tool (https://www.ncbi.nlm.nih.gov/orffinder/).

**Aim of Study:** apply ML methods to classify genera and ORFs based on viral genomic and related satellite sequences.

**Training Data:** source data for genus classification were complete genomes of species from the nine genera plus genome sequences for alphasatellite and betasatellite (11 class labels): 1333 sequences for training and 9188 for testing. Values for 24 predictors were computed: proportions of A, T, C, and G of the whole sequence; and proportions of A, T, C, G, and GC content when the sequence was split equally into four parts.

Source data for ORF classification were ORFs from the genome sequences used in genus classification (12 class labels as per 10 genes plus two non-coding genes from the satellite sequences): 4035 ORFs for training and 33918 for testing. Values for 24 predictors were computed: proportions of A, T, C, and G of the coding DNA sequence (CDS); and proportion of each one of the 20 amino acids in the translated CDS.

**Algorithms used:** Sequential Minimal Optimization (an algorithm for training SVM), RF, and Multilayer Perceptron (MLP) (a class of ANN)

**Algorithm implementation:** Weka

**Validation of trained models:** 10-fold cross-validation and LOO

**Statistical evaluation measures used:** accuracy, precision, recall, F-measure, AUC

**Best model chosen:** MLP for genus classification, and RF for ORF classification.

**Application:** best ML models are packaged within Geminivirus Data warehouse: http://geminivirus.org:8080/geminivirusdw/.

**Example 20** (Sperschneider et al., 2016) | Fungi | Plant pathogens | Fungal effector prediction from secretomes using Naive Bayes classifier

**Background:** effectors are proteins secreted from plant-associated organisms to the host that modulate the host cell to facilitate infection (Kamoun, 2006). Conversely, non-effectors are also secreted but not involved in infection.

**Problem Setting:** fungal effector, rule-based prediction from secretomes is poor due to a lack of consensus sequence motifs and the reliance on manually set thresholds such as small size and cysteine content.

**Aim of Study:** apply ML to fungal effector prediction.

**Training Data:** positive sequence set (effectors) derived from 58 fungal effectors from 16 fungal species that were reported in the literature. Negative sequence set (non-effectors) were derived by first predicting secretomes of the same 16 species and then filtering out sequences that share similarity to the positive set to obtain 14143 negative sequences. Positive and negative set numbers were
balanced for training by randomly selecting negative sequences. Features relevant for classification were extracted for each sequence. There were 10 features such as amino acid frequencies, amino acid class frequencies, molecular weight, sequence length, and protein net charge.

**Algorithms used:** evaluated LR, NB, DT, RF and ANN.

**Algorithm implementation:** Weka tool box (v.3.6.12).

**Validation of trained models:** 10-fold cross-validation

**Statistical evaluation measures used:** SN, SP, PPV, NPV, AUC.

**Best model chosen:** NB (84.5% SN, 82.8% SP, 62% PPV, 94.1% NPV, 0.898 AUC).

**Application:** EFFECTORP: http://effectorp.csiro.au/

**Example 21** (Kalafi et al., 2016) | Helminths | Taxonomy | Identification of Monogenean species via image classifications using kNN

**Background:** Monogenean parasites are flatworms primarily found on gills and skin of fishes. Differentiation of Monogenean species is based on morphological characteristics of sclerotized hard parts (termed haptoral attachment organs) and reproductive parts (Vignon, 2011).

**Problem setting:** automated differentiation using microscopic digital images is a challenge because of complex and overlapping anatomical structures.

**Aim of Study:** implement image processing techniques and ML methods for automated identification of Monogenean species.

**Training Data:** consisted of 40 digital images of four randomly chosen Monogenean species. There were 10 images per species showing each species unique hard anatomical structures (i.e. 10 images * 4 species = 40 images). Similarly, the test dataset consisted of 40 different images but still showing anatomical structures from the same four species (10 images per species).

**Image processing:** included three software executed steps: image pre-processing, image segmentation, and feature extraction. Ten features were extracted. Linear Discriminant Analysis (LDA) was used to reduce the number of features by finding the combination of features obtaining the best separability among the four Monogenean species. LDA is dimensionality reduction method before classification by finding a linear combination of features separating two or more classes; however, it can still be used strictly as a linear classifier (note that logistic regression is traditionally limited to only two-class classification problems).

**Algorithms used:** kNN (‘k’ parameter set to 10). Three selected features obtained from LDA were used as the input i.e. the input comprised vectors of image labels (e.g. species 1, 2, 3 or 4) prepared according to their features e.g. feature_1, species_1, feature_x, species_x.

**Validation of trained model:** Leave-One-Out (LOO) and 10 fold cross-validations

**Statistical evaluation measures used:** ACC
**Best model chosen:** test dataset with k = 10 resulted in 90% ACC when using the three LDA selected features. ACC was 75% when using all 10 extracted features; 91.25% ACC for LOO and 95.5% ACC for 10 fold cross-validation on training data.

**Application:** an automated, time saving identification system for monogenean species that is useful to taxonomists and non-taxonomists. The underlying methods could be extended to identify other microscopic specimens.

**Example 22** (Barratt et al., 2019, Barratt & Sapp, 2020, Nascimento et al., 2020) | Helminths and Protozoa | Taxonomy and Epidemiology | Assessment of genetic relationships in the context of complex genotypes using an ensemble learning procedure

**Background:** Genotyping of helminth and protozoan parasites using multi-locus sequence typing (MLST) strategies is a common approach used in the contexts of molecular epidemiology, and to address taxonomic questions (Barratt et al., 2019, Barratt & Sapp, 2020). However, these organisms are eukaryotes (as opposed to prokaryotes) that often possess multiple genomes (e.g. nuclear, mitochondrial, and plastid), with different mechanisms of inheritance (Barratt et al., 2019). Many protozoa and helminths reproduce sexually, so characterising genetic relationships between individuals is not straightforward because individuals can be heterozygous, possessing more than one sequence type at the same locus. This does not occur for pure cultures of prokaryotic microbes. Many protozoa and helminths cannot be cultured in vitro, so genotyping them relies on collecting individual parasite specimens from infected patients for extraction of DNA. In the case of single-celled protozoa, this is not a feasible approach so DNA is often extracted from a collection of protozoan cells in a clinical specimen (e.g. blood or faeces) which may constitute a mixed (non-clonal) population. In some instances, parasite load in clinical specimens can be very low, or the mass of specimen provided for parasite genotyping can be insufficient to extract enough DNA to amplify all typing loci in a panel prior to sequence typing. In these contexts, missing data can be a serious problem and limitations associated with analytic methods for multi-locus-sequence-typing (MLST) data (e.g. genetic distance statistics calculated as part of typical phylogenetic approaches) often result in exclusion of specimens with partial MLST data (Barratt et al., 2019, Barratt & Sapp, 2020, Nascimento et al., 2020).

**Problem setting:** Methods used to calculate genetic distances for phylogenetic applications cannot accommodate specimens with missing loci and do not satisfactorily address the problem of heterozygosity in specimens. Therefore, methods for calculating distance statistics for MLST datasets possessing these problematic features are required.

**Aim of Study:** This example constitutes three studies, where an unsupervised ensemble-based distance statistic is used in place of a traditional genetic distance metric commonly used in phylogenetic applications. Unlike commonly used phylogenetic procedures, the ensemble learning procedure calculates a distance metric even when specimens possess multiple alleles at any number of sequenced loci, and can accommodate specimens with data missing at some MLST loci. The algorithms generate a distance matrix that can be clustered downstream for interpretation.

**Algorithm used:** This approach is an unsupervised ensemble learning procedure with two underlying algorithms. One of these is an unsupervised Bayesian algorithm and the other is a heuristic algorithm.

**Statistical evaluation measures used:** Sensitivity, specificity, precision, and negative predictive value (Nascimento et al., 2020). These metrics were calculated using epidemiologically-defined clusters of illness as ground truth labels, where the agreement between genetic clustering using the
ensemble-based distance statistic and epidemiologically defined clusters was used to identify true positive, false positive, true negative, and false negative cluster (label) assignments.

Application: This ensemble-based distance statistic has been used to identify infections with the protozoan parasite *Cyclospora cayetanensis* that are genetically related as part of a molecular surveillance routine (Nascimento *et al.*, 2020). This procedure has also been applied to the human- and dog-infecting helminth *Strongyloides stercoralis*, to assess whether *S. stercoralis* populations from humans and dogs are genetically similar or distinct (Barratt & Sapp, 2020).

**Example 23** (Cieslak *et al.*, 2020) | Copepods | Microbial ecology and microbiomes | Distinguishing different groups of copepods in various transcriptional states based on RNA-Seq data

**Background:** For RNA-Seq experiments, differences in gene expression between groups is traditionally assessed by enrichment analyses based on the Gene Ontology (GO), and statistical approaches such as Fisher’s exact test (Dong *et al.*, 2016, Lenk *et al.*, 2019).

**Problem setting:** The Gene Ontology (GO) and statistical approaches such as Fisher’s exact test (Dong *et al.*, 2016, Lenk *et al.*, 2019) are effective at identifying differences in transcriptional states between test populations but procedures that identify more generalised transcriptional categories can be more conducive to hypothesis generation.

**Aim of Study:** T-distributed stochastic neighbour embedding (t-SNE) was assessed for its ability to differentiate RNA-Seq datasets from copepods collected at different sites, from different copepod developmental stages and, copepods cultured under different conditions (Cieslak *et al.*, 2020)

**Algorithm used:** t-distributed stochastic neighbour embedding (t-SNE)

**Statistical evaluation measures used:** Concordance of profile categories identified by t-SNE with those identified using conventional differential gene expression and GO analyses performed on the same datasets.

**Application:** This study supported that t-SNE can be effective at classifying complex RNA-Seq datasets into categories based on their transcriptional physiology (Cieslak *et al.*, 2020).
**Table 1**

Real-life published examples of supervised and unsupervised machine learning applied to microbiology

| Problem setting                                                                 | Organism type | Research area          | Algorithm(s) used¹                  |
|-------------------------------------------------------------------------------|---------------|------------------------|-------------------------------------|
| Identify compounds in traditional Chinese medicine to use as anti-malarials(Maindola et al., 2015). | Protozoa      | Drug discovery         | NB, RF, SVM, J48                     |
| Identify derivatives of Primaquine (antimalarial drug) that will retain or improve its efficacy, while reducing its toxicity(Levatic et al., 2018). | Protozoa      | Drug discovery         | QSAR modelling with SVM (regression) |
| New method based on amino acid composition for predicting secretory proteins from malarial parasites(Verma et al., 2008). | Protozoa      | Drug/vaccine discovery | SVM                                 |
| Identify potential targets against *Leishmania* parasites(Sinha et al., 2017). | Protozoa      | Drug/vaccine discovery | NB                                  |
| Predict parasite load for *Leishmania infantum* from clinical records when real-time quantitative polymerase chain reaction (qPCR) services are not available(Torrecilha et al., 2017). | Protozoa      | Clinical application   | ANN with non-linear radial basis functions |
| Fish discrimination given parasite community data as biological markers(Perdiguero-Alonso et al., 2008). | Protozoa      | Taxonomy               | RF                                  |
| Prediction of major livestock sources of *Salmonella* Typhimurium using whole-genome sequencing data(Zhang et al., 2019). | Bacteria      | Epidemiology           | RF                                  |
| Clustering of *Yersinia pseudotuberculosis* restriction fragment length polymorphism patterns(Voskresenskaya et al., 2014). | Bacteria      | Epidemiology           | Hierarchical clustering by unweighted pair group method with arithmetic mean (UPGMA) |
| Predict effector proteins for *Legionella pneumophila* using optimal set from 370 features (chemical, structural, and compositional properties of protein | Bacteria      | Drug discovery         | SVM                                 |
| sequences)(Ashari et al., 2019). |  |  |
|---|---|---|
| **Gram-positive subcellular localization** **prediction using evolutionary and structural features (EvoStruct-Sub)**(Uddin et al., 2018). | Bacteria | Drug discovery | SVM |
| **Predictor named QSpred-FL for the detection of quorum-sensing peptides (associated with cellular processes e.g. cell–cell communication, gene expression regulation)** (Wei et al., 2018). | Bacteria | Drug discovery | RF |
| **Predicting protective antigens by extracting features directly from pathogen protein sequences**(Rahman et al., 2019). | Bacteria | Vaccine discovery | RF |
| **Identify genetic signatures of antibiotic resistance via analysis of Mycobacterium tuberculosis pan-genome**(Kavvas et al., 2018). | Bacteria | Antimicrobial resistance | SVM |
| **Phenotypic antimicrobial susceptibility testing with deep learning video microscopy**(Yu et al., 2018). | Bacteria | Antimicrobial resistance | Deep learning (with images) |
| **Assessing the spread of antimicrobial resistance in different environmental samples using metagenomic sequencing**(Oh et al., 2018). | Bacteria | Metagenomics, Antimicrobial resistance | k-means, fuzzy c-means and k-medoid clustering. |
| **Prediction of infection on admission to hospital**(Rawson et al., 2019). | Bacteria | Clinical applications | SVM |
| **Automated classification for blood-culture Gram stains**(Smith et al., 2018). | Bacteria | Clinical applications | Deep CNN |
| **Fast identification of bacteraemia in systemic inflammatory response syndrome patients on care wards**(Ratzinger et al., 2018). | Bacteria | Clinical applications | RF, ANN |
| **Prediction of new in vitro leads for Ebola virus**(Anantpadma et al., 2019). | Viruses | Drug discovery | NB |
| **Predicting individuals most at risk of being exposed to a pathogen from ecological data**(Fountain-Jones et al.). | Viruses | Clinical applications | SVM, RF, Gradient boosting |
| **Modelling US geographic distribution of tuberculosis**(Mollalo et al., 2019). | Viruses | Epidemiology | ANN |
| Title                                                                 | Entity 1          | Entity 2          | Method 1         |
|----------------------------------------------------------------------|-------------------|-------------------|------------------|
| Monitoring influenza outbreaks using the social media platform Twitter(Allen et al., 2016). | Viruses           | Epidemiology      | SVM              |
| Detection of viral sequences in human metagenomic datasets(Bzhalava et al., 2018). | Viruses           | Metagenomics      | RF, ANN          |
| Integrating mass spectrometry-based metabolomic data: to screen patients with Zika Virus(Melo et al., 2018). | Viruses           | Metabolomics      | RF               |
| Assessing genetic relationships between various families of viruses(Dougan & Quake, 2019). | Virus             | Taxonomy          | t-SNE            |
| Predicting ascospore release of Monilinia vaccinii-corymbosi, which impacts the blueberry industry(Harteveld et al., 2017). | Fungi             | Plant microbes    | RF, ANN, LR      |
| Projecting global distribution of amphibian fungal pathogen from Intergovernmental Panel on Climate Change data(Xie et al., 2016). | Fungi             | Climate change    | RF               |
| Demonstrating immunomodulatory effect of helminth derivatives in autoimmune diseases(Ben-Amram et al., 2017). | Helminths         | Microbiome        | Nearest shrunken centroids |
| Chemical and genetic validation of the statin drug target to treat the helminth disease, schistosomiasis(Rojo-Arreola et al., 2014). | Helminths         | Drug discovery    | SVM              |
| Investigating host susceptibility to helminthiasis from repeated Schistosoma japonicum infection(Carlton et al., 2013). | Helminths         | Clinical applications | RF, kNN          |

Artificial Neural Networks (ANN), Convolutional Neural Network (CNN), Decision Tree (DT), Deep Learning is essentially working with large neural networks, k-nearest neighbour classification (kNN), J48 (an open source Java implementation of C4.5 decision tree algorithm), Linear logistic Regression (LR), Naive Bayes classifier (NB); Random Forest (RF), Support Vector Machine (SVM), T-distributed stochastic neighbour embedding (t-SNE).
Figures

**Figure 1.** k-Nearest Neighbour (kNN) classifier

‘k’ in kNN is the number of nearest neighbours to the unknown observation as determined by an appropriate metric e.g. Euclidean distance. The circle shown is an indication only of the ‘k’ nearest neighbours selected for voting process. Results to determine class membership of unknown observation are derived from majority voting within the enclosed circle.

**Figure 2.** Support vector Machine (SVM) example

The aim is to find the maximum margin between the separable classes. Margin is the space between the two class regions. Support vectors are used to identify the hyperplane, which is a straight line in two dimensions.
**Figure 3:** The kernel trick

The aim is to binary split the Yes and No data points. This is not possible in a 2-dimensional plane. However, the classes Yes and No can be separated based on a Z value in a 3-dimensional plane.
References

Allen C, Tsou MH, Aslam A, Nagel A & Gawron JM (2016) Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. Plos One 11.

Alvar J, Velez ID, Bern C, Herrero M, Desjeux P, Cano J, Jannin J, den Boer M & Team WHOLC (2012) Leishmaniasis Worldwide and Global Estimates of Its Incidence. Plos One 7.

Anantpadma M, Lane T, Zorn KM, Lingerfelt MA, Clark AM, Freundlich JS, Davey RA, Madrid PB & Ekins S (2019) Ebola Virus Bayesian Machine Learning Models Enable New in Vitro Leads. Acs Omega 4: 2353-2361.

Ashari ZE, Brayton KA & Broschat SL (2019) Using an optimal set of features with a machine learning-based approach to predict effector proteins for Legionella pneumophila. Plos One 14.

Ashley EA, Dhorda M, Fairhurst RM, et al. (2014) Spread of Artemisinin Resistance in Plasmodium falciparum Malaria. New England Journal of Medicine 371: 411-423.

Barratt JLN & Sapp SGH (2020) Machine learning-based analyses support the existence of species complexes for Strongyloides fuelleborni and Strongyloides stercoralis. Parasitology 146.

Barratt JLN, Park S, Nascimento FS, Hofstetter J, Plucinski M, Blank M, Shoenfeld Y & Koren O (2017) Tuftsin-Phosphorylcholine Maintains Normal Gut Microbiota in Collagen Induced Arthritic Mice. Frontiers in Microbiology 8.

Bhattacharyya S, Jha S, Tharakunnel K & Westland JC (2011) Data mining for credit card fraud: A comparative study. Decision Support Systems 50: 602-613.

Bibin D, Nair MS & Punitha P (2017) Malaria Parasite Detection From Peripheral Blood Smear Images Using Deep Belief Networks. Ieee Access 5: 9099-9108.

Burchhardt I, Sebastian K, Mauder N, Kostrzewa M, Burckhardt F & Zimmermann S (2019) Analysis of Streptococcus pneumoniae using Fourier-transformed infrared spectroscopy allows prediction of capsular serotype. Eur J Clin Microbiol Infect Dis 38: 1883-1890.

Bzhalava Z, Tampuu A, Bala P, Vicente R & Dillner J (2018) Machine Learning for detection of viral sequences in human metagenomic datasets. BMC Bioinformatics 19.

Carlton EJ, Hubbard A, Wang S & Spear RC (2013) Repeated Schistosoma japonicum Infection Following Treatment in Two Cohorts: Evidence for Host Susceptibility to Helminthiasis? Plos Neglected Tropical Diseases 7.

Chang C-C & Lin C-J (2011) LIBSVM: A Library for Support Vector Machines. Acm Transactions on Intelligent Systems and Technology 2.

Chen CY-C (2011) TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening In Silico. Plos One 6.

Chollet F & Ieee (2017) Xception: Deep Learning with Depthwise Separable Convolutions. 30th Ieee Conference on Computer Vision and Pattern Recognition.p,^pp. 1800-1807.

Cieslak MC, Castelfranco AM, Roncalli V, Lenz PH & Hartline DK (2020) t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis. Mar Genomics 51: 100723.

Colubri A, Silver T, Fradet T, Retzepi K, Fry B & Sabeti P (2016) Transforming Clinical Data into Actionable Prognosis Models: Machine-Learning Framework and Field-Deployable App to Predict Outcome of Ebola Patients. Plos Neglected Tropical Diseases 10.

Consonni V & Todeschini R (2010) Molecular Descriptor.

Cortes C & Vapnik V (1995) Support-vector networks. Machine Learning 20: 273-297.

Dantas-Torres F (2007) The role of dogs as reservoirs of Leishmania parasites, with emphasis on Leishmania (Leishmania) infantum and Leishmania (Viannia) braziliensis. Veterinary Parasitology 149: 139-146.

Dong X, Hao Y, Wang X & Tian W (2016) LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. Sci Rep 6: 18871.
Dong Y, Jiang Z, Shen H, Pan WD, Williams LA, Reddy VVB, Benjamin WH, Jr., Bryan AW, Jr. & Ieee (2017) Evaluations of Deep Convolutional Neural Networks for Automatic Identification of Malaria Infected Cells.

Dougan TJ & Quake SR (2019) Viral taxonomy derived from evolutionary genome relationships. *Plos One* **14**.

Flower DR, Macdonald IK, Ramakrishnan K, Davies MN & Doytchinova IA (2010) Computer aided selection of candidate vaccine antigens. *Immune research* **6 Suppl 2**: S1-S1.

Fountain-Jones NM, Machado G, Carver S, Packer C, Recamonde-Mendoza M & Craft ME How to make more from exposure data? An integrated machine learning pipeline to predict pathogen exposure. *Journal of Animal Ecology*.

Friedman JH (2002) Stochastic gradient boosting. *Computational Statistics & Data Analysis* **38**: 367-378.

G. Howard A, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M & Adam H (2017) *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*.

Garcia-Palacios P, Vandegehuchte ML, Shaw EA, Dam M, Post KH, Ramirez KS, Sylvain ZA, de Tomasel CM & Wall DH (2015) Are there links between responses of soil microbes and ecosystem functioning to elevated CO2, N deposition and warming? A global perspective. *Global Change Biology* **21**: 1590-1600.

Gerla M, Lee E-K, Pau G, Lee U & Ieee (2014) *Internet of Vehicles: From Intelligent Grid to Autonomous Cars and Vehicular Clouds*.

Gonzales-Siles L, Salva-Serra F, Degerman A, Norden R, Lindh M, Skovbjerg S & Moore ERB (2019) Identification and capsular serotype sequotyping of Streptococcus pneumoniae strains. *J Med Microbiol* **68**: 1173-1188.

Goodswen SJ, Kennedy PJ & Ellis JT (2013) A novel strategy for classifying the output from an in silico vaccine discovery pipeline for eukaryotic pathogens using machine learning algorithms. *Bmc Bioinformatics* **14**.

Goodswen SJ, Kennedy PJ & Ellis JT (2014) Vaccceed: a high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology. *Bioinformatics* **30**: 2381-2383.

Grau J, Grossel I & Keilwagen J (2015) PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**: 2595-2597.

Guzella TS & Caminhas WM (2009) A review of machine learning approaches to Spam filtering. *Expert Systems with Applications* **36**: 10206-10222.

Halsey ES, Venkatesan M, Plucinski MM, et al. (2017) Capacity Development through the US President's Malaria Initiative-Supported Antimalarial Resistance Monitoring in Africa Network. *Emerg Infect Dis* **23**.

Harris JB, LaRocque RC, Chowdhury F, Khan AI, Logvinenko T, Faruque ASG, Ryan ET, Qadri F & Calderwood SB (2008) Susceptibility to Vibrio cholerae Infection in a Cohort of Household Contacts of Patients with Cholera in Bangladesh. *Plos Neglected Tropical Diseases* **2**.

Harteveld DOC, Grant MR, Pscheidt JW & Peever TL (2017) Predicting Ascospore Release of Monilinia vaccinii-corymbosi of Blueberry with Machine Learning. *Phytopathology* **107**: 1364-1371.

Hawinkel S, Kerckhof FM, Bijnen L & Thas O (2019) A unified framework for unconstrained and constrained ordination of microbiome read count data. *PLoS One* **14**: e0205474.

Huang G, Liu Z, van der Maaten L, Weinberger KQ & Ieee (2017) Densely Connected Convolutional Networks. *30th Ieee Conference on Computer Vision and Pattern Recognition*, pp. 2261-2269.

Japkowicz N (2000) *The class imbalance problem: Significance and strategies*.

Jones S, Plucinski M, Kay K, Hodel EM & Hastings IM (2020) A Computer Modelling Approach To Evaluate the Accuracy of Microsatellite Markers for Classification of Recurrent Infections during Routine Monitoring of Antimalarial Drug Efficacy. *Antimicrob Agents Chemother* **64**.

Kalafi EY, Tan WB, Town C & Dhillon SK (2016) Automated identification of Monogeneans using digital image processing and K-nearest neighbour approaches. *Bmc Bioinformatics* **17**.

Kamada N, Kim YG, Sham HP, Vallance BA, Puente JL, Martens EC & Nunez G (2012) Regulated Virulence Controls the Ability of a Pathogen to Compete with the Gut Microbiota. *Science* **336**: 1325-1329.
MacRaild CA, Anders RF, Foley M & Norton RS (2011) Apical Membrane Antigen 1 as an Anti-Malarial Drug Target. Current Topics in Medicinal Chemistry 11: 2039-2047.

Maindola P, Jamal S & Grover A (2015) Cheminformatics Based Machine Learning Models for AMA1-RON2 Abrogators for Inhibiting Plasmodium falciparum Erythrocyte Invasion. Molecular Informatics 34: 655-664.

Manna L, Reale S, Vitale F & Gravino AE (2009) Evidence for a relationship between Leishmania load and clinical manifestations. Research in Veterinary Science 87: 76-78.

Mathison BA, Kohan JL, Walker JF, Smith RB, Ardon O & Couturier MR (2020) Detection of Intestinal Protozoa in Trichrome-Stained Stool Specimens by Use of a Deep Convolutional Neural Network. J Clin Microbiol 58.

Matthews BW (1975) Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochimica Et Biophysica Acta 405: 442-451.

Mecham BH, Nelson PS & Storey JD (2010) Supervised normalization of microarrays. Bioinformatics 26: 1308-1315.

Melo C, Navarro LC, de Oliveira DN, et al. (2018) A Machine Learning Application Based in Random Forest for Integrating Mass Spectrometry-Based Metabolomic Data: A Simple Screening Method for Patients With Zika Virus. Frontiers in Bioengineering and Biotechnology 6.

Midani FS, Weil AA, Chowdhury F, et al. (2018) Human Gut Microbiota Predicts Susceptibility to Vibrio cholerae Infection. Journal of Infectious Diseases 218: 645-653.

Mitchell AM & Mitchell TJ (2010) Streptococcus pneumoniae: virulence factors and variation. Clin Microbiol Infect 16: 411-418.

Mollalo A, Mao L, Rashidi P & Glass GE (2019) A GIS-Based Artificial Neural Network Model for Spatial Distribution of Tuberculosis across the Continental United States. International Journal of Environmental Research and Public Health 16.

Nascimento FS, Barratt J, Houghton K, et al. (2020) Evaluation of an ensemble-based distance statistic for clustering MLST datasets using epidemiologically defined clusters of cyclosporiasis. Epidemiology and Infection under peer review.

Noedl H, Se Y, Schaecher K, Smith BL, Socheat D, Fukuda MM & Consortium ARCS (2008) Evidence of Artemisinin-Resistant Malaria in Western Cambodia. New England Journal of Medicine 359: 2619-2620.

Oh M, Pruden A, Chen C, Heath LS, Xia K & Zhang L (2018) MetaCompare: a computational pipeline for prioritizing environmental resistome risk. FEMS Microbiol Ecol 94.

Pavić K, Perković I, Cindrić M, Pranjić M, Martin-Kleiner I, Kralj M, Schols D, Hadjipavlou-Litina D, Katsori AM & Zorc B (2014) Novel semicarbazides and ureas of primaquine with bulky aryl or hydroxyalkyl substituents: Synthesis, cytostatic and antioxidative activity. European Journal of Medicinal Chemistry 86: 502-514.

Pavić K, Perković I, Gilja P, Kozlina F, Ester K, Kralj M, Schols D, Hadjipavlou-Litina D, Pontiki E & Zorc B (2016) Design, Synthesis and Biological Evaluation of Novel Primaquine-Cinnamic Acid Conjugates of the Amide and Acylsemicarbazide Type. Molecules 21: 1629.

Perdiguero-Alonso D, Montero FE, Kostadinova A, Raga JA & Barrett J (2008) Random forests, a novel approach for discrimination of fish populations using parasites as biological tags. International Journal for Parasitology 38: 1425-1434.

Perković I, Antunović M, Marijanović I, et al. (2016) Novel urea and bis-urea primaquine derivatives with hydroxyphenyl or halogenphenyl substituents: Synthesis and biological evaluation. European Journal of Medicinal Chemistry 124: 622-636.

Phyo AP, Nkhoma S, Stepniewska K, et al. (2012) Emergence of artemisinin-resistant malaria on the western border of Thailand: a longitudinal study. Lancet 379: 1960-1966.

Plucinski MM, Morton L, Bushman M, Dimbu PR & Udhayakumar V (2015) Robust Algorithm for Systematic Classification of Malaria Late Treatment Failures as Recrudescence or Resurgence Using Microsatellite Genotyping. Antimicrob Agents Chemother 59: 6096-6100.

Poore GD, Kopylova E, Zhu Q, et al. (2020) Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 579: 567-+.

Pooschti M, Silamut K, Maude RJ, Jaeger S & Thoma G (2018) Image analysis and machine learning for detecting malaria. Translational Research 194: 36-55.
Rahman MS, Rahman MK, Saha S, Kaykobad M & Rahman MS (2019) Antigenic: An improved prediction model of protective antigens. *Artificial Intelligence in Medicine* **94**: 28-41.

Rajaraman S, Antani SK, Poostchi M, Silamut K, Hossain MA, Maude RJ, Jaeger S & Thoma GR (2018) Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *Peerj* **6**.

Rajaraman S, Antani SK, Poostchi M, Silamut K, Hossain MA, Maude RJ, Jaeger S & Thoma GR (2018) Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *Peerj* **6**.

Ramirez KS, Knight CG, de Hollander M, et al. (2018) Detecting macroecological patterns in bacterial communities across independent studies of global soils. *Nature Microbiology* **3**: 189-196.

Ratzinger F, Haslacher H, Perkmann T, Pinzan M, Anner P, Makristathis A, Burgmann H, Heinz G & Dorffner G (2018) Machine learning for fast identification of bacteraemia in SIRS patients treated on standard care wards: a cohort study. *Scientific Reports* **8**.

Rawson TM, Hernandez B, Moore LSP, et al. (2019) Supervised machine learning for the prediction of infection on admission to hospital: a prospective observational cohort study. *Journal of Antimicrobial Chemotherapy* **74**: 1108-1115.

Robinson KM, Crabtree J, Mattick JSA, Anderson KE & Hotopp JCD (2017) Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data. *Microbiome* **5**.

Rogers D & Hahn M (2010) Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50**: 742-754.

Rojo-Arreola L, Long T, Asarnow D, Suzuki BM, Singh R & Caffrey CR (2014) Chemical and Genetic Validation of the Statin Drug Target to Treat the Helminth Disease, Schistosomiasis. *Plos One* **9**.

Ross NE, Pritchard CJ, Rubin DM & Duse AG (2006) Automated image processing method for the diagnosis and classification of malaria on thin blood smears. *Medical & Biological Engineering & Computing* **44**: 427-436.

Samuel AL (1959) Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* **3**: 211-6.

Schieffelin JS, Shaffer JG, Goba A, et al. (2014) Clinical Illness and Outcomes in Patients with Ebola in Sierra Leone. *New England Journal of Medicine* **371**: 2092-2100.

Shan C, Gong S & McOwan PW (2009) Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing* **27**: 803-816.

Shapiro SS & Wilk MB (1965) An analysis of variance test for normality (complete samples). *Biometrika* **52**: 591-6.

Sohn MB & Li H (2018) A GLM-based latent variable ordination method for microbiome samples. *Biometrics* **74**: 448-457.

Sperschneider J, Gardiner DM, Dodds PN, Tini F, Covarelli L, Singh KB, Manners JM & Taylor JM (2016) EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytologist* **210**: 743-761.

Steinbeck C, Han YQ, Kuhn S, Hloracher O, Luttmann E & Willighagen E (2003) The Chemistry Development Kit (CDK): An open-source Java library for chemoinformatics. *Journal of Chemical Information and Computer Sciences* **43**: 493-500.

Sun S, Zhu J, Ma Y & Zhou X (2019) Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol* **20**: 269.

Svarovskaia ES, Cheslock SR, Zhang WH, Hu WS & Pathak VK (2003) Retroviral mutation rates and reverse transcriptase fidelity. *Frontiers in Bioscience-Landmark* **8**: D117-D134.

Talundzic E, Plucinski MM, Biliya S, Silva-Flannery LM, Arquin PM, Halsey ES, Barnwell JW, Vannberg F & Udhayakumar V (2016) Advanced Molecular Detection of Malarone Resistance. *Antimicrob Agents Chemother* **60**: 3821-3823.
Timi JT (2007) Parasites as biological tags for stock discrimination in marine fish from South American Atlantic waters. *Journal of Helminthology* **81**: 107-111.

Toropov AA, Toropova AP, Mukhamedzhanova DV & Gutman I (2005) Simplified molecular input line entry system (SMILES) as an alternative for constructing quantitative structure-property relationships (QSPR). *Indian Journal of Chemistry Section a-Inorganic Bio-Inorganic Physical Theoretical & Analytical Chemistry* **44**: 1545-1552.

Torrecilha RBP, Utsunomiya YT, Bosco AM, et al. (2016) Correlations between peripheral parasite load and common clinical and laboratory alterations in dogs with visceral leishmaniasis. *Preventive Veterinary Medicine* **132**: 83-87.

Uddin MR, Sharma A, Farid DM, Rahman MM, Dehzangi A & Shatabda S (2018) EvoStruct-Sub: An accurate Gram-positive protein subcellular localization predictor using evolutionary and structural features. *Journal of Theoretical Biology* **443**: 138-146.

van-der-Maaten LJP & Hinton GE (2008) Visualizing Data Using t-SNE. *Journal of Machine Learning Research* **9**: 2579–2605.

Verma R, Tiwari A, Kaur S, Varshney GC & Raghava GPS (2008) Identification of proteins secreted by malaria parasite into erythrocyte using SVM and PSSM profiles. *Bmc Bioinformatics* **9**.

Vignon M (2011) Putting in shape - towards a unified approach for the taxonomic description of monogenean haptorial hard parts. *Systematic Parasitology* **79**: 161-174.

Voskresenskaya E, Savin C, Leclercq A, Tseneva G & Carniel E (2014) Typing and clustering of *Yersinia pseudotuberculosis* isolates by restriction fragment length polymorphism analysis using insertion sequences. *J Clin Microbiol* **52**: 1978-1989.

Walker TM, Kohl TA & Omar SV (2018) Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study (vol 15, pg 1193, 2015). *Lancet Infectious Diseases* **18**: 21-21.

Wang Y, Xiao J, Suzek TO, Zhang J, Wang J & Bryant SH (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research* **37**: W623-W633.

Warwick K, Craddock R & iee (1996) An introduction to Radial Basis Functions for system identification - A comparison with other neural network methods. *Proceedings of the 35th Ieee Conference on Decision and Control, Vols 1-4, p.* pp. 464-469.

Wei L, Hu J, Li F, Song J, Su R & Zou Q (2018) Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. *Briefings in Bioinformatics*.

Weinstein JN, Collison EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM & Canc Genome Atlas Res N (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **45**: 1113-1120.

White NJ (2013) Primaquine to prevent transmission of falciparum malaria. *Lancet Infectious Diseases* **13**: 175-181.

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z & Woolsey J (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* **34**: D668-D672.

Wood DE & Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**.

Xie GY, Olson DH & Blaustein AR (2016) Projecting the Global Distribution of the Emerging Amphibian Fungal Pathogen, *Batrachochytrium dendrobatidis*, Based on IPCC Climate Futures. *Plos One* **11**.

Xu D & Tian Y (2015) A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science* **2**: 165-193.

Yang Y, Niehaus KE, Walker TM, et al. (2018) Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics* **34**: 1666-1671.

Ye Z, Van Dyke K & Rossan RN (2013) Effective treatment with a tetrandrine/chloroquine combination for chloroquine-resistant falciparum malaria in Aotus monkeys. *Malaria Journal* **12**.
Yu H, Jing WW, Iriya R, Yang YZ, Syal K, Mo MN, Grays TE, Haydel SE, Wang SP & Tao NJ (2018) Phenotypic Antimicrobial Susceptibility Testing with Deep Learning Video Microscopy. *Analytical Chemistry* **90**: 6314-6322.

Zhang HT, Li DF, Zhao LL, et al. (2013) Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. *Nature Genetics* **45**: 1255-U1217.

Zhang SK, Li ST, Gu WD, et al. (2019) Zoonotic Source Attribution of Salmonella enterica Serotype Typhimurium Using Genomic Surveillance Data, United States. *Emerging Infectious Diseases* **25**: 82-91.

Zhou Y, Wilkinson D, Schreiber R & Pan R (2008) Large-scale parallel collaborative filtering for the Netflix Prize. *Algorithmic Aspects in Information and Management, Proceedings*, Vol. 5034 (Fleischer R & Xu JH, eds.), pp. 337-348.

Zorn KM, Lane TR, Russo DP, Clark AM, Makarov V & Ekins S (2019) Multiple Machine Learning Comparisons of HIV Cell-based and Reverse Transcriptase Data Sets. *Molecular Pharmaceutics* **16**: 1620-1632.