Found: machine learning model comparison in ncbi

Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol:

Salod et al., 2019

Robustness: stratified k-cross fold validation; correlation based feature selection, parameter tuning

Run sets on both full and selected features

Significance for public health: … an artificial intelligence tool can eventually be created using highest performing model(s). clinicians can input patient-specific biomarkers into the tool. The tool would output the likelihood of patients having BC (Breast Cancer), with a certain level of accuracy. This envisioned process could serve to eventually revolutionize the early prediction of BC in patients and consequently, a reduction in BC mortality rate.

<overview of the problem at hand: breast cancer>

<traditional / state of the art methods for screening, detecting BC>

<Ml intro>

>ML and BC> includes how the data was collected

<proposed first approach for screening using ML>

Research Questions:

1. Which ML model is best using this data
2. Do our model beat the best available models out there

Hypotheses:

H01) the performance of all different models are the same

HA1) the performance of at least one is different

H02) all the best performing models from our study do not outperform the best out there

HA2) at least one of the best performing ML models created here outperforms the best of the world

Research objectives:

1. Investigate which subset of features from the BCCD result in the best performing models, based on the elimination of highly positively correlated features, implying redundant features, using Correlation-based Feature selection algorithm
2. To determine the correct values for different parameters of the respective ML models, that result in the best performance per model using parameter tuning
3. To identify the best performing ML model(s) in this study that optimally enable(s) the prediction of BC
4. To examine whether the best performing ML model(s) in this study also outperform(s) the existing ML models from the literature concerning BC prediction

<materials and methods>

Data source

Framework of the modeling process

Feature selection

Train test validation set

ML algorithms used

1. LR: whip out function
2. SVM were not doing that
3. KNN neither
4. DT neither
5. RF: ensemble algorithm for DT. Dt is easy to interpret and efficient. Rule based flowchart structure is easily understandable (or so we think 😉). Traversing a list of rules / questions down to the prediction
6. Boosting algorithm: we should consider. Basically adjusts the weights of observations from earlier classifications. If an observation is misclassified, the boosting attempts to increase the weight of this observation and vice versa. XGBoost or ADABoost or GBM

Stratified k fold cross validation: the accuracy from the validation set indicates whether or not the model is robust ( but up to a certain point lol 😊)

Explain what confusion matrix is, with formulas. TP top left; FN top right; FP bottom left, TN bottom right. Use accuracy, precision, recall, specificity, F1. Also ROC and AUROC.

Whip out on what specs you ran the model

<statistical analysis for model comparison>

Friedman test: using all metrics collectively; all stats test evaluated at 5% level of significance

Conclusions

Future work: share results with clinicians, so they can be sort of validated. This also helps future data collection.

Comparing different supervised machine learning algorithms for disease prediction

Uddin et al., 2019

Disease prediction: potential application area for ML methods

Here only looking at supervised methods

This study follows PRISMA guidelines; More specifically, this article considered only those studies that used more than one supervised machine learning algorithm for a single disease prediction in the same research setting.

In 48 articles in Pubmed and Scopus: search ML algo for disease predictions

Most used: LR, but RF usually has better results. 29 studies use SVM, 23 use NB, 17 use RF. But RF best accuracy followed by SVM

ML is applied to medicine:

8. Yao D, Yang J, Zhan X. A novel method for disease prediction: hybrid of random forest and multivariate adaptive regression splines. J Comput. 2013;8(1):170–177. doi: 10.4304/jcp.8.1.170-177. [[CrossRef](https://dx.doi.org/10.4304%2Fjcp.8.1.170-177)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=J+Comput&title=A+novel+method+for+disease+prediction:+hybrid+of+random+forest+and+multivariate+adaptive+regression+splines&author=D+Yao&author=J+Yang&author=X+Zhan&volume=8&issue=1&publication_year=2013&pages=170-177&doi=10.4304/jcp.8.1.170-177&)] [[Ref list](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6925840/#CR8)]

LR to supervised classification:

22. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Wiley; 2013. [[Ref list](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6925840/#CR22)]

DT

24. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1):81–106. [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Mach+Learn&title=Induction+of+decision+trees&author=JR+Quinlan&volume=1&issue=1&publication_year=1986&pages=81-106&)] [[Ref list](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6925840/#CR24)]

RF

26. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Mach+Learn&title=Random+forests&author=L+Breiman&volume=45&issue=1&publication_year=2001&pages=5-32&)] [[Ref list](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6925840/#CR26)]

NB

27. Lindley DV. Fiducial distributions and Bayes’ theorem. J Royal Stat Soc. 1958;1:102–107. [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=J+Royal+Stat+Soc&title=Fiducial+distributions+and+Bayes%E2%80%99+theorem&author=DV+Lindley&volume=1&publication_year=1958&pages=102-107&)] [[Ref list](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6925840/#CR27)]

28. Rish I. IJCAI 2001 workshop on empirical methods in artificial intelligence. New York: IBM; 2001. An empirical study of the naive Bayes classifier; pp. 41–46. [[Google Scholar](https://scholar.google.com/scholar_lookup?title=IJCAI+2001+workshop+on+empirical+methods+in+artificial+intelligence&author=I+Rish&publication_year=2001&)] [[Ref list](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6925840/#CR28)]

This theorem can describe the probability of an event based on the prior knowledge of conditions related to that event. This classifier assumes that a particular feature in a class is not directly related to any other features although features for that class could have interdependence among themselves. By considering the task of classifying a new object to either green or red class:

Prior probability: there are 40 green dots, 20 red. The probability of green is 40/60 = 2/3, for red it is 20/60 = 1/3. These are the prior probabilities.

To classify a new object: consider a number of points that are similar. What fi we consider 4 points, 3 of the are red and one is green – so the likelihoods are 1/40 for green and 3/20 for red. The final classifier is produced as likelihood value times prior probability. This is the posterior probability, which is for red = 2/3 \* 1/40 = 0.017 while for red it is 1/3\*3/20 = 0.049. The highest posterior probability yields the class of the object we are trying to classify.

ANN: In the biological brain, neurons are connected to each other through multiple axon junctions forming a graph like architecture. These interconnections can be rewired (e.g., through neuroplasticity) that helps to adapt, process and store information. ANN algorithms ca be represented as an interconnected group of nodes. The output of one node goes as input to another node for subsequent processing according to the interconnection. Nodes are normally grouped into a matrix called a layer depending on the transformation they perform. Apart from the input and output layer, there can be one or more hidden layers in an ANN framework. Nodes and edges have weights that enable to adjust signal strengths of communication which can be amplified or weakened through repeated training. Based on the training and subsequent adaptation of the matrices, node and edge weights, ANNS can make a prediction for the test data.

Comparison

ROC: An ROC is one of the fundamental tools for diagnostic test evaluation and is created by plotting the true positive rate against the false positive rate at various threshold settings.

Roc

37. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27(8):861–874. doi: 10.1016/j.patrec.2005.10.010. [[CrossRef](https://dx.doi.org/10.1016%2Fj.patrec.2005.10.010)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Pattern+Recogn+Lett&title=An+introduction+to+ROC+analysis&author=T+Fawcett&volume=27&issue=8&publication_year=2006&pages=861-874&doi=10.1016/j.patrec.2005.10.010&)] [[Ref list](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6925840/#CR37)]

RMSE: running (root?) mean square error: for different pairs of actual and predicted values, RMSE represents the mean value of all square errors. An error is the difference between an actual and its corresponding predicted value. Another such measure is the mean absolute error. For an actual and its predicted value, MAE indicates the absolute value of their difference.

Advantages and limitations of different supervised machine learning algorithms:

ANN: can detect complex non-linear relationships between dependent and independent variables; requires less formal statistical training; classification AND regression; but is black box, computationally expensive if complex problem, and predictor or independent variables require pre=processing

DT: resultant classification tree is easy to understand and interpret (or so we think), data preparation is easy, different data types are supported, can be validated using statistical tests, requires classes to be mutually exclusive, depends on the order of attributes / variables, poor performance

LR: easy to implement, straightforward, can be updated easily, does not make assumptions regarding the distribution of independent variables; has a probabilistic interpretation of model parameters; but does not have good accuracy for complex relationships, does not consider the linear relationship between variables, logic models are vulnerable to overconfidence, may overstate the prediction accuracy due to sampling bias

Naïve bayes: simple, useful for large sets, requires less training data; but classes must be mutually exclusive, presence of dependency between attributes negatively affects the classification performance; assumes the normal distribution of numeric attributes.

RF: lower chance of variance and overfitting than DT, since RF takes the average value from the outcomes of its constituent decision trees; perform well for large datasets, but can be computationally expensive, number of base classifiers needs to be defined, favors those variables or attributes that can take high number of different values in estimating variable importance (high cardinality); can overfit easily (although you can get an estimate of this using oob error)