

## Article

# Pitfalls of using multinomial regression analysis to identify class-structure relevant variables in biomedical datasets: Why a mixture of experts (MOE) approach is better.

Jörn Lötsch <sup>1,2,\*</sup>  and Alfred Ultsch <sup>3</sup>

<sup>1</sup> Institute of Clinical Pharmacology, Goethe - University, Theodor Stern Kai 7, 60590 Frankfurt am Main, Germany

<sup>2</sup> Fraunhofer Institute for Translational Medicine and Pharmacology ITMP, Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany

<sup>3</sup> DataBionics Research Group, University of Marburg, Hans – Meerwein - Straße, 35032 Marburg, Germany

\* Correspondence: j.loetsch@em.uni-frankfurt.de (J.L.)

**Abstract:** Recent advances in mathematical modelling and artificial intelligence have challenged the use of traditional regression analysis in biomedical research. This study examined artificial and cancer research data using binomial and multinomial logistic regression and compared its performance with other machine learning models such as random forests, support vector machines, Bayesian classifiers, k-nearest neighbours and repeated incremental clipping (RIPPER). The alternative models often outperformed regression in accurately classifying new cases. Logistic regression had a structural problem similar to early single-layer neural networks, which limited its ability to identify variables with high statistical significance for reliable class assignment. Therefore, regression is not always the best model for class prediction in biomedical datasets. The study emphasises the importance of validating selected models and suggests that a "mixture of experts" approach may be a more advanced and effective strategy for analysing biomedical datasets.

**Keywords:** data science; artificial intelligence; machine-learning; digital medicine

## 1. Introduction

Data analysis of an association between clinical class outcomes, such as patient versus controls, or responders versus non-responders, and complex underlying information, such as omics data, is often performed using classical methods such as binary logistic regression, or in case of  $k > 2$  classes, multinomial logistic regression. These methods are believed to be effective in achieving the desired goals, making it unnecessary to broaden the methodological spectrum on the data analysis side. Hence, reviewers and editors of scientific papers occasionally require restricting the data analysis to this perceived standard, up to demanding deletion of alternative approaches on the grounds these would merely produce confusion without adding relevant information. This is in stark contrast to the demand for a wider range of data analysis methods covering different facets of a disease and the use of more laboratory methods to address complex clinical or preclinical questions.

However, restricting data analysis to, for example, calculating differential expression and performing a regression makes implicit assumptions such as (i) the relevant information is fully contained in the statistically significant variables, and (ii) the regression always adequately captures the relevant structure in a data set, and therefore analysis of which variables were relevant in the regression model is sufficient. In the dynamic and rapidly evolving field of data science, which encompasses diverse experts ranging from biomedical researchers to computer scientists and fosters collaborative efforts among individuals driven by their shared passion for data, it becomes apparent that the above assumptions are only partially valid. First, and it can be shown that the best predictive variables for subgroup assignment, such as a medical diagnosis, are not necessarily the variables that

**Citation:** Lötsch, J.; Ultsch, A. Pitfalls of using multinomial regression analysis to identify class-structure relevant variables in biomedical datasets: Why a mixture of experts (MOE) approach is better.. *Biomedinformatics* **2023**, *1*, 1–15.

<https://doi.org/>

Received:

Revised:

Accepted:

Published:

**Copyright:** © 2023 by the authors. Submitted to *Biomedinformatics* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

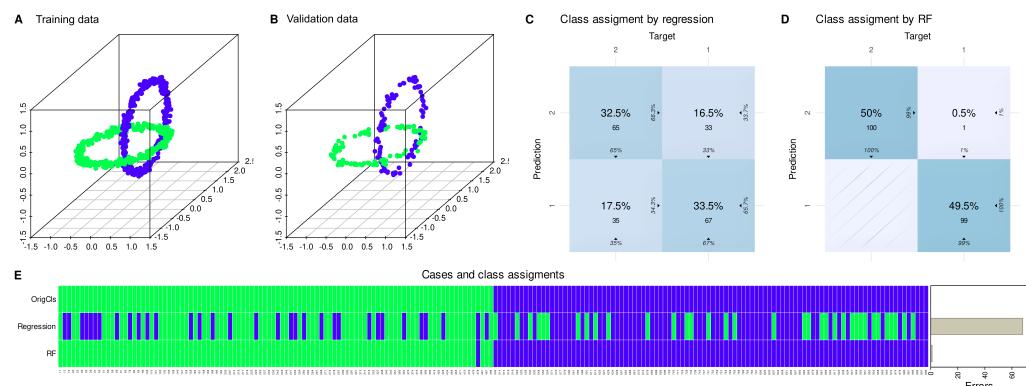
differ most significantly between subgroups or may not even show statistically significant group differences at all [1]. Second, the assertion that regression necessarily adequately describes data sets with a class structure, and therefore can always provide a reliable basis for identifying relevant variables, is challenged by a canonical example presented below.

## 2. Introductory example case

As an example of the limitations of relying solely on a regression model, consider a data set consisting of two classes that form two orthogonally intertwined rings in three dimensions (variables) ("X", "Y", and "Z") [2]. These rings are clearly visually distinguishable, and the variables are independent (Pearson's correlation [3]  $r = -0.064$  to  $r = 0.026$ ). Logistic regression analysis produced a highly significant result indicating that the variable "Y" carried the relevant information for class membership of the data points, with a p-value of  $p < 2 \cdot 10^{-16}$  (Table 1).

**Table 1.** Results of a multivariate linear regression analysis of the FCPS-based "Chainlink" data set. For reproducibility, the code used to perform this analysis is available at [https://github.com/JornLotsch/MisClassificationRegressionNN/blob/main/Chainlink\\_Reg\\_paper.R](https://github.com/JornLotsch/MisClassificationRegressionNN/blob/main/Chainlink_Reg_paper.R).

Variable	Regression				
	Estimate	Std. Error	Z-value	Pr(> z )	Signif.
(Intercept)	0.01541	0.08657	0.178	8.59E-01	
X	0.03846	0.0827	0.465	0.642	
Y	1.56726	0.11461	13.674	<2e-16	***
Z	-0.06873	0.0824	-0.834	0.404	



**Figure 1.** Weakness of the regression model fit on 80% of the data (training) to predict the remaining 20% of the data (validation) of the FCPS-based "ChainLink" data set. **A:** Training data set. **B:** Validation data set. **C:** Confusion matrix plot of the assignment to each of the  $k = 2$  classes using the regression model. In the center of each tile is the normalized number (total percentage) of cases assigned to each class, and below is the number of cases per class. The column-wise percentages are shown at the bottom, and the row-wise percentages are shown to the right of each tile. **D:** Confusion matrix plot for the results when singing a trained random forest model instead of the regression model. **E:** Matrix heat plot of the class assignment of unseen cases (20% hold-out validation sample) by either regression or random forests trained on 80% of the data set. For comparison, the original class assignments are shown in the top row. The figure has been created using the R software package (version 4.3.1 for Linux; <https://CRAN.R-project.org/> [4]) and the R libraries "ggplot2" (<https://cran.r-project.org/package=ggplot2> [5]), "scatterplot3d" (<https://cran.r-project.org/package=scatterplot3d> [6]), "cvms" (<https://cran.r-project.org/package=cvms> [7]) and "ComplexHeatmap" (<https://www.bioconductor.org/packages/ComplexHeatmap> [8]).

Regression was run on 80% of the  $n = 1000$  data points (400 of each class), while a 20% class-proportional holdout sample of the data (100 data points of each class) was separated as a validation data subset prior to analysis. When applying the fitted regression model

on this validation sample, it becomes clear that the model captured the class structure in the data set with only 65% accuracy. In contrast, there are alternative algorithms, such as random forests [9], that can much better separate the two classes (Figure 1).

This raises the question of whether a research result in terms of relevant variables should be based on a model that incompletely describes the data, in particular when an obviously better model is available. By contrast, it seems preferable to validate several models on data not used for model fitting, and then to use the best models for identification of relevant variables on which the main conclusions of a research project can be based. It will be shown this problem applies not only to the artificial data set chosen here as an introductory example because its true class structure is known by design, but also to real biomedical data sets.

### 3. Materials and Methods

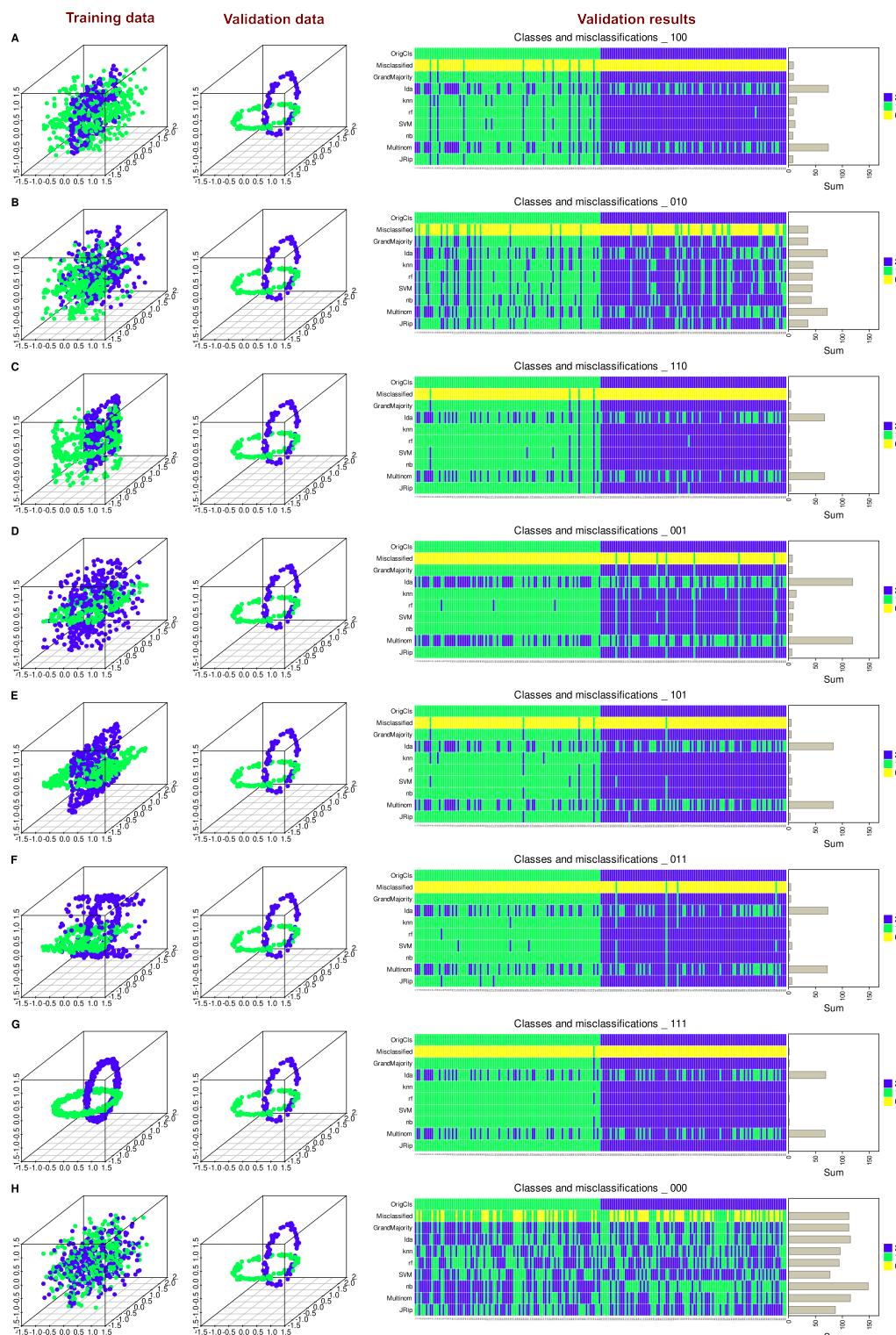
#### 3.1. Sample data sets

The introductory example data set is an artificial data set published as part of the "Fundamental Clustering Problems Suite (FCPS)" [2]. It is freely available at <https://www.mdpi.com/2306-5729/5/1/13/s1> or in the similarly named R library "FCPS" (<https://cran.r-project.org/package=FCPS> [10]). The "ChainLink" data set contains two distinct classes arranged as interlinked rings in the variables X, Y and Z, each class containing  $n = 500$  members. Two other data sets from the FCPS collection, "Atom" and "Tetra", were additionally used. The "Atom" data set contains 3D data visually resembling an atomic nucleus and its hull, each composed of  $n = 200$  data points. The "Tetra" data set consists of four classes of  $n = 100$  points each, and each point has three variables indicating the position in a three-dimensional space and drawn independently and uniformly distributed. Additionally, a canonical machine learning data set was generated consisting of 4 normal distributions with means  $[(5,-5), (-5,5), (-5,-5), (5,5)]$  and standard deviations of 1 in both x and y directions with  $n = 400$  observations in each Gaussian. This results in four heaps of points in an xy coordinate system. In both dimensions, the variables were normalized, i.e., standardized to mean 0 and unit variance. The upper left and lower right corners belong to class 1, and the lower left and upper right corners belong to class 2. This "Four Gaussians" data set provides a classic "XOR" structured data set [11].

A large biomedical data set was from a multiomics data analysis in the context of non-small cell lung cancer [12]. The data set contained information from a total of  $n = 566$  patients, including 672 differentially expressed mRNAs, 9 microRNAs, 719 gene methylation data, and 153 protein expression data. In the original study, spectral clustering was used to identify  $k = 5$  clusters that correlated with patient survival. For the present analysis, we downloaded Supplementary Table S17 from the publication site at [https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-023-31426-w/MediaObjects/41598\\_2023\\_31426\\_MOESM2\\_ESM.xlsx](https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-023-31426-w/MediaObjects/41598_2023_31426_MOESM2_ESM.xlsx). The file contains cluster membership information and preprocessed omics data used by the authors of the original paper to train and validate machine learning-based classifiers. In the present demonstration, the data set was used to test the sufficiency of regression as an analysis method of biomedical data sets in comparison to alternative methods, with no intention to replicate or challenge the published results.

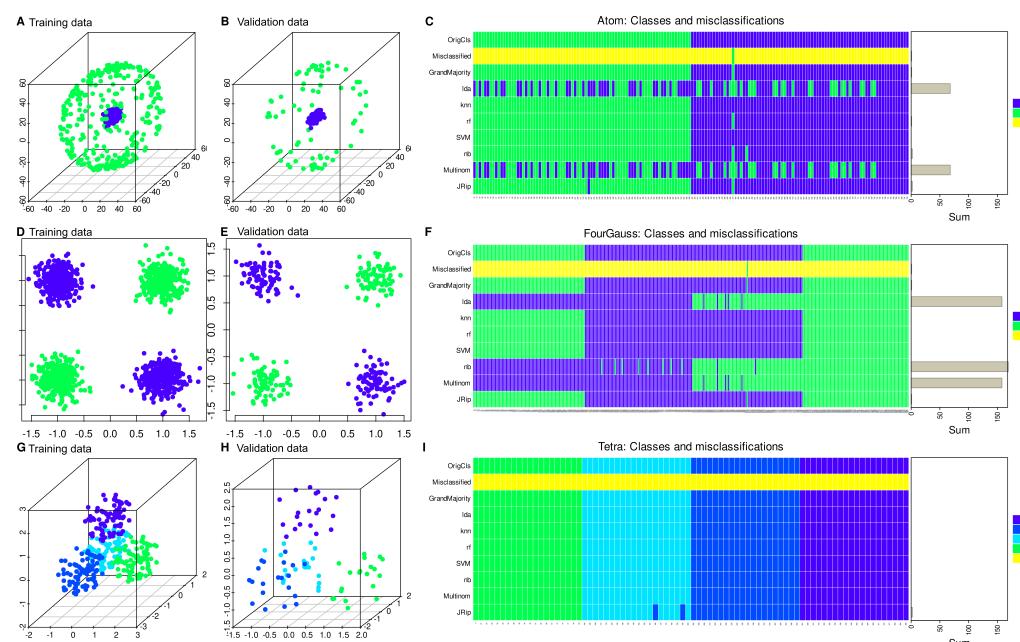
#### 3.2. Experimentation

Software coding was done in the R language [13] and Python [14] languages on Linux. The main packages were "caret" (<https://cran.r-project.org/package=caret> [15]) for R and "scikit-learn" (<https://scikit-learn.org/stable/> [16]) for Python. A set of algorithms covering various machine learning and statistical approaches was trained on an 80% training sample of the data set, including regression techniques, specifically binary or multinomial logistic regression for data sets with  $k = 2$  or  $k > 2$  classes, respectively. Additionally, linear discriminant analysis (LDA) [17], a commonly applied statistical procedure, was employed.



**Figure 2.** Comparative class assignment error rates for unseen cases (20% hold-out validation sample) of the FCPS-based "ChainLink" data set by  $d = 6$  different classifiers trained on 80% of the respective data sets (lda = linear discriminant analysis, knn = k-nearest neighbors, rf = random forest, SVM = support vector machines, nb = naive Bayes, multinom = logistic regression, binary for  $k = 2$  classes, jRip = repeated incremental clipping for error reduction). An exhaustive approach of permutations of 1 - 3 variables was used to identify the variables on which the class assignment can best be based. The plots show the different scenarios, with a 3-dimensional visualization of the actual data set on the left and the classification results on the right. Which variables were permuted or not is coded in the titles of the matrix heat plots as 0 or 1, e.g., "111" means all variables in original form, "000" all permuted, "010" only the 2<sup>nd</sup> of three variables, i.e., "Y", permuted, and so on. For comparison, the original class assignments are shown in the top row, the number of finally misclassified cases is shown in the second row, based on the majority vote of all  $d = 6$  classifiers shown in the third row.

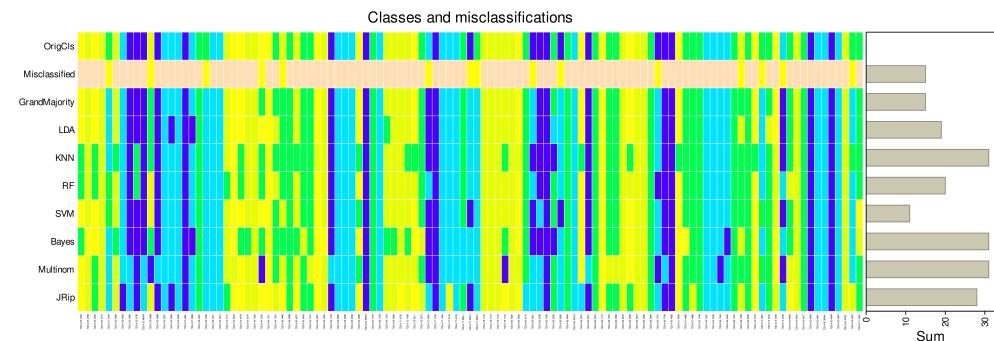
Furthermore, a non-compressive collection of supervised machine learning algorithms was utilized, chosen methods often used for biomedical data analysis and generally known to usually well perform classification tasks. These included random forests [9,18] as a robust tree-based bagging classifier, support vector machines (SVM) [19] as a hyperplane separation-based method, k-nearest neighbors [20] as a distance-based classifier, naïve Bayes [21] as a posterior probabilities-based method, and furthermore, decision rules were created based on repeated incremental clipping for error reduction (RIPPER [22]). Hyperparameter tuning, such as kernel shape for SVM, number of trees for random forests, and others, was done in a grid search, and classifier training and validation was done using nested cross-validation. Details of the R code are available at [https://github.com/JornLotsch/MisClassificationRegressionNN/blob/main/MisClassificationAnalysis\\_MainFunctions\\_paper.R](https://github.com/JornLotsch/MisClassificationRegressionNN/blob/main/MisClassificationAnalysis_MainFunctions_paper.R).



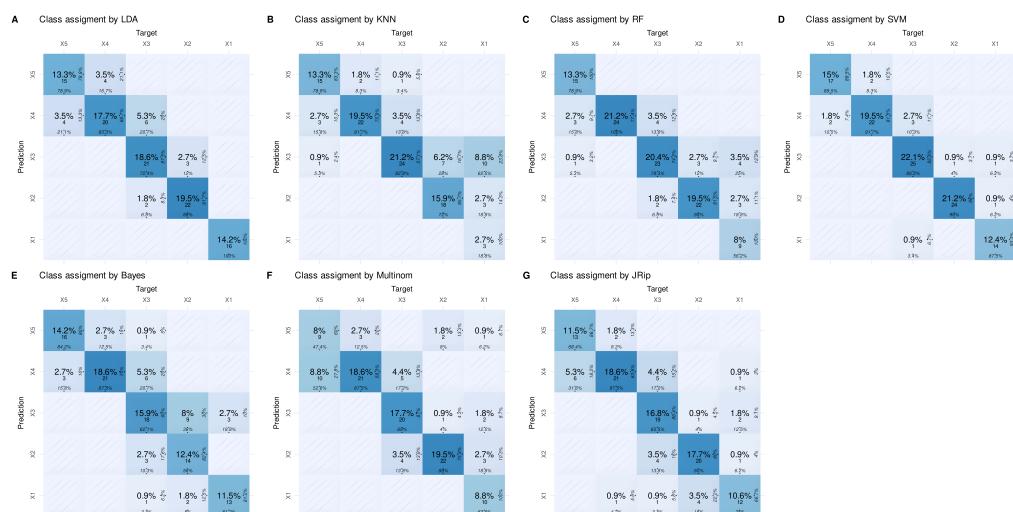
**Figure 3.** Comparative class assignment error rates for unseen cases (20% hold-out validation sample) by  $d = 6$  different classifiers trained with 80% of the respective data sets (Lda = linear discriminant analysis, knn = k-nearest neighbors, rf = random forest, SVM = support vector machines, nb = naïve Bayes, Multinom = logistic regression, binary for  $k = 2$  classes and multinomial for  $k > 2$  classes, jRip = repeated incremental clipping for error reduction). **A:** Training data from the FCPS based “Atom” data set. **B:** Validation data. **C:** Matrix heat plot of the class assignment of unseen cases in the hold-out validation sample. For comparison, the original class assignments are shown in the top row, the number of finally misclassified cases is shown in the second row, based on the majority vote of all  $d = 6$  classifiers shown in the third row. **D - F:** As panels A - C but for the (standardized) “Four Gaussians” data set. **G - I:** As panels A - C but for the FCPS based “Tetra” data set. The figure has been created using the R software package (version 4.3.1 for Linux; <https://CRAN.R-project.org/> [4]) and the R libraries “ggplot2” (<https://cran.r-project.org/package=ggplot2> [5]), “scatterplot3d” (<https://cran.r-project.org/package=scatterplot3d> [6]), “cvms” (<https://cran.r-project.org/package=cvms> [7]) and “ComplexHeatmap” (<https://www.bioconductor.org/packages/ComplexHeatmap> [8]).

The methods were then compared in terms of their classification performance based on balanced accuracy [23] in a 20% hold-out validation sample separated from the data set before the analysis. In addition, the area under the receiver operator curve (roc-auc) was calculated [24]. A “mixture of experts” approach was used to identify the most suitable variables based on the best-performing classification algorithms. These were identified by item categorization into subsets labeled “A,” “B,” and “C”, implemented as computed ABC analysis (cABC) [25]. Algorithms falling in category “C”, which is generally considered

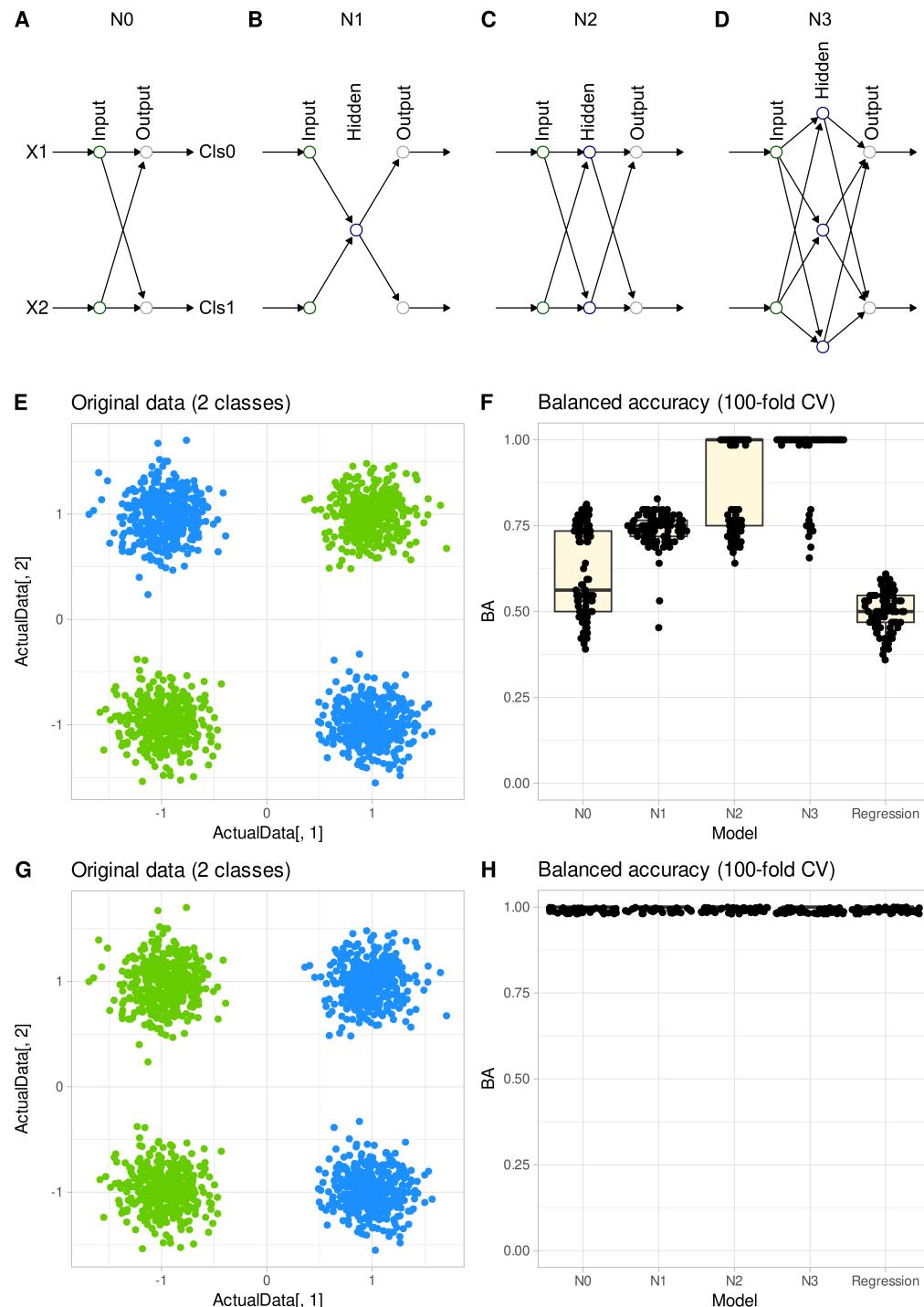
as “the trivial many” [26], were rejected as a basis of variable selection. The informative variables were then selected for the best models using generic permutation importance as a feature selection method applicable to any type of classifier.



**Figure 4.** Multiomics data set [12]: Comparative class assignment error rates for unseen cases (20% hold-out validation sample) by  $d = 6$  different classifiers trained with 80% of the respective data sets (lda = linear discriminant analysis, knn = k-nearest neighbors, rf = random forest, SVM = support vector machines, nb = naïve Bayes, Multinom = multinomial logistic regression, for  $k = 5$  classes, JRip = repeated incremental clipping for error reduction). Matrix heat plot of the class assignment of unseen cases in the hold-out validation sample. For comparison, the original class assignments are shown in the top row, the number of finally misclassified cases is shown in the second row, based on the majority vote of all  $d = 6$  classifiers shown in the third row. The figure has been created using the R software package (version 4.3.1 for Linux; <https://CRAN.R-project.org/> [4]) and the R libraries "ggplot2" (<https://cran.r-project.org/package=ggplot2> [5]), "scatterplot3d" (<https://cran.r-project.org/package=scatterplot3d> [6]), "cvms" (<https://cran.r-project.org/package=cvms> [7]) and "ComplexHeatmap" (<https://www.bioconductor.org/packages/ComplexHeatmap/> [8]).

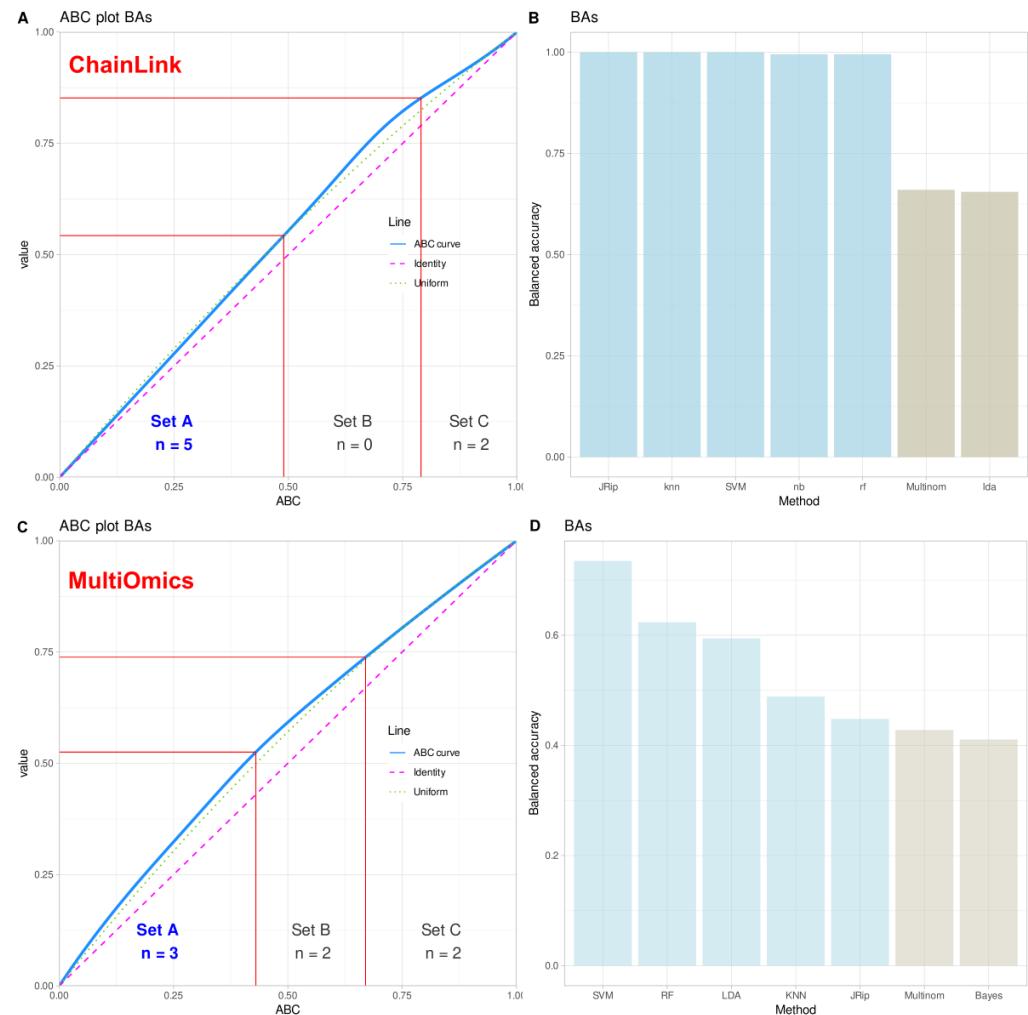


**Figure 5.** Multiomics data set [12]: Confusion matrix plot of the assignment to each of the  $k = 2$  classes using the logistic regression model. In the center of each tile is the normalized number (total percentage) of cases assigned to each class, and below is the number of cases per class. The column-wise percentages are shown at the bottom, and the row-wise percentages are shown to the right of each tile. A: LDA = linear discriminant analysis, B: KNN = k-nearest neighbors, C: RF = random forest, D: SVM = support vector machines, E: Bayes = naïve Bayes, F: Multinom = logistic regression, (multinomial for  $k = 5$  classes). The figure has been created using the R software package (version 4.3.1 for Linux; <https://CRAN.R-project.org/> [4]) and the R libraries "ggplot2" (<https://cran.r-project.org/package=ggplot2> [5]) and "cvms" (<https://cran.r-project.org/package=cvms> [7]).



**Figure 6.** Structural limitations of regression to separate classes in a data set, exemplified by the "Four Gaussians" data (standardized). **A - D:** Neural network architectures with 0 - 3 neurons in a hidden layer (N0 - N3). **E:** Original two-class data set. **F:** Balanced accuracies obtained on a 20% hold-out validation sample after training the algorithms in a 100-fold cross-validation scenario. The box plots show the median and the 25th and 75th percentiles and are overlaid with individual balanced accuracies as single points. **G:** Data set switched to a simpler problem. **H:** Balanced accuracies obtained by running the same code used in panel F on the data set in panel G. The figure has been created using the R software package (version 4.3.1 for Linux; <https://CRAN.R-project.org/> [4]) and the R library "ggplot2" (<https://cran.r-project.org/package=ggplot2> [5]) and "nnet" (<https://cran.r-project.org/package=nnet> [27]).

Since this was not a feature selection benchmarking assessment, other methods (for an overview, see [28]) were not considered. Finally, the selected variables were used to train algorithms in a  $5 \times 20$  nested cross-validation scenario using randomly selected subsets of 67% of the original training data set and applied to each 80% of the hold-out validation data subset.



**Figure 7.** Identification of the best algorithms based on the balanced accuracy (BA) of class assignment for unseen cases (20% hold-out validation sample) by  $d = 6$  different classifiers trained with 80% of the respective data sets (lda = linear discriminant analysis, knn = k-nearest neighbors, rf = random forest, SVM = support vector machines, nb = naïve Bayes, Mutinom = logistic regression, (multinomial for  $k = 5$  classes, jRip = repeated incremental clipping for error reduction). **A:** Item categorization via cABC analysis of the balanced accuracies in the FCPS based “ChainLink” data set. The ABC plots (blue lines) show the cumulative distribution function of the importance variables together with the identity distribution,  $x_i = \text{constant}$  (magenta line), and the uniform distribution, i.e., as a stopping criterion for the repetitions of the cABC analysis. The red lines show the boundaries between the ABC subsets "A", "B" and "C". **B:** Bar plot of the performance measures, quantified as balanced accuracy, of the  $d = 6$  algorithms. **C and D:** Similar to panels A and B but for the real-life multiomics data set [12]. The figure has been created using the R software package (version 4.3.1 for Linux; <https://CRAN.R-project.org/> [4]) and the R libraries "ggplot2" (<https://cran.r-project.org/package=ggplot2> [5]) and "ABCAnalysis" (<https://cran.r-project.org/package=ABCAnalysis> [25]).

## 4. Results

### 4.1. Regression occasionally generalizes poorly compared to alternative methods

As shown in the introductory example, the regression result for the variable "Y" in the "ChainLink" data set was highly statistically significant (Table 1). However, cautious interpretation is needed when considering "Y" as the primary determinant of class membership since regression had only modest ability to assign class membership of new cases, not seen during regression model building (Figure 1).

To further validate this, six different models were trained on all possible combinations of preserved and randomly permuted variables (Figure 2). The results showed that most algorithms successfully classified unseen data when trained on the three variables, with regression and LDA performing worst as mentioned in the introductory example. By contrast, classification constantly failed when all variables were randomly permuted as the expected outcome in this scenario (Figure 2 G and H, respectively). Surprisingly when considering the statistical results of regression analysis, training with only the variable "Y", i.e., the variable emerging as a highly significant result from regression analysis, resulted in poor classification performance for new data across all algorithms (Figure 2 B).

Regression was also found to be inferior to machine learning algorithms in the FCPS based "Atom" data set as well as in the "Four Gaussians" data set (Figure 3). However, to highlight that regression can be a successful method, the FCPS based "Tetra" data set was included in which it expectedly provided a perfect classification similar to the other methods (Figure 3).

The weakness of regression was not arbitrarily produced by the actual choice of artificial data sets expected to pose problems, but extends to real-life data sets, such as a large multiomics data set from cancer research (Figure 4).

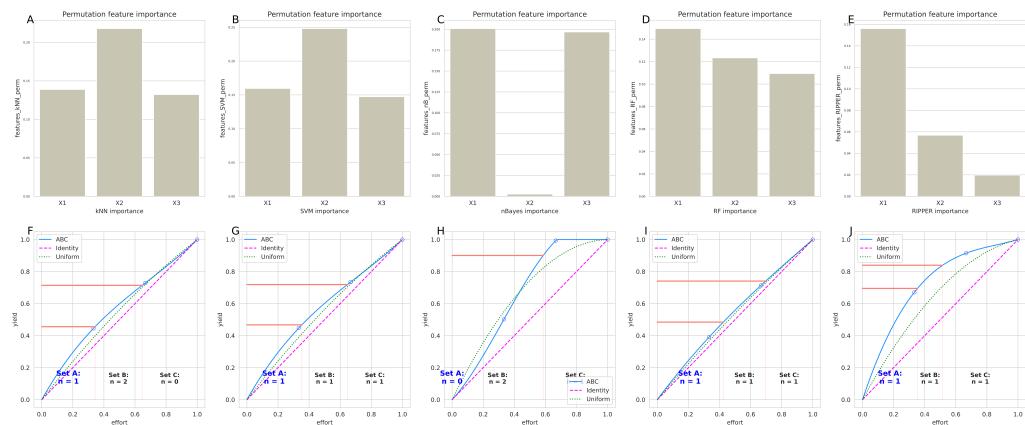
There, for example support vector machines outperformed regression considerably. This data set identified once again regression as one of the weakest predictors producing least accurate class assignment of unseen (validation) cases (Figure 5).

### 4.2. Regression inadequately captures the structural characteristics of certain data sets

In the "Four Gaussians" data set, regression analysis failed to provide the correct classification when the classes in the data set are linear non-separable (Figure 6 E and F). Its inability to draw a line between the two original classes was shared by a single-layer neural network (Figure 6 A). The balanced accuracy obtained with binary logistic regression or with a single-layer perceptron was about 50%, i.e., guessing level in two classes. Only the addition of a hidden layer with at least 1 and up to 3 neurons (Figure 6 B - D) enabled the algorithms to perform the classification successfully, eventually reaching 100% accuracy. When the class labels were changed so that the classes were linearly separable (Figure 6 G), both regression and simple artificial neural networks were able to perform classification with 100% accuracy (Figure 6 H). The R code for this experiment can be available at <https://github.com/JornLotsch/MisClassificationRegressionNN>.

### 4.3. Variables chosen by the most successful algorithms are more generalizable

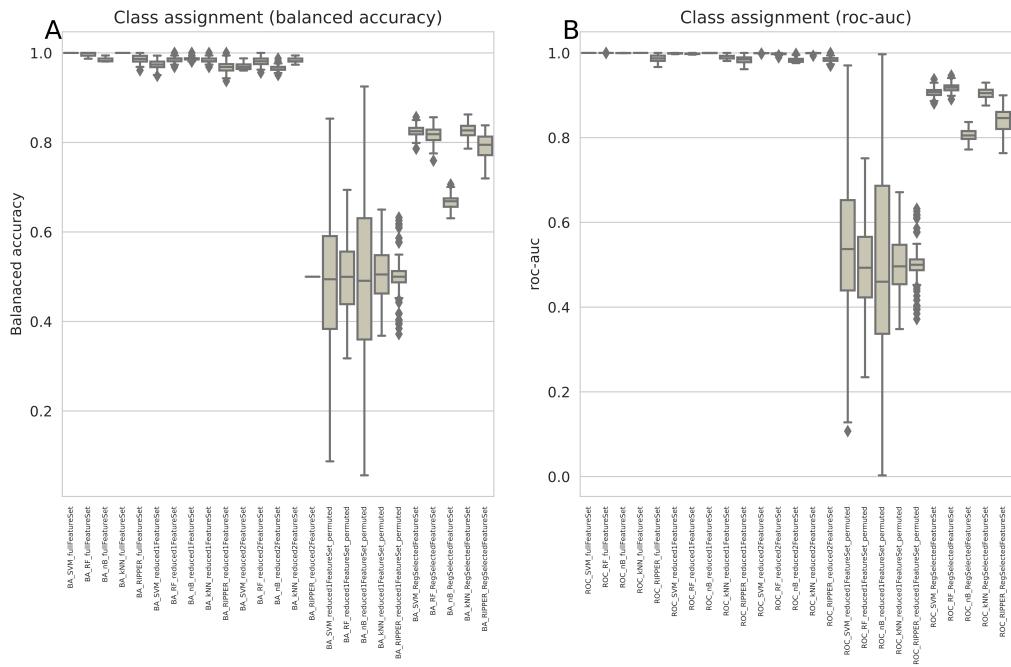
Above experiments cast doubt on the validity of the statistically significant result of the logistic regression analysis identifying the variable 'Y' in the FCPS "ChainLink" data set as the relevant variable for class assignment. Regression was the poorest model for assigning new cases to the correct classes based on the available variables, and therefore may not be the most reliable "expert" to indicate which variables are relevant in the data set. Using cABC analysis and the permutation importance method on the balanced accuracy, LDA and regression were placed in category "C" by cABC analysis, indicating that these methods provided the poorest description of the "ChainLink" data set and should therefore not be explored further for important variables (Figure 7).



**Figure 8.** Feature selection for the FCPS based “ChainLink” data set, using the  $d = 4$  best performing classification algorithms, with generic permutation importance and balanced accuracy as the criterion. **A - D:** Bar plots of permutation features importance (variables “X1”, “X2”, “X3” corresponding to the x, y and z coordinates of the three-dimensional data set, resulting from different classifiers as indicated at the x-axis legends (kNN = k-nearest neighbors, SVM = support vector machines, nBayes = naïve Bayes, RF= random forests, RIPPER = repeated incremental clipping for error reduction). **E – H:** Corresponding ABC plots of feature importance. Variables in sets “A” or “B” were retained for the majority vote. The ABC plots (blue lines) show the cumulative distribution function of the importance variables together with the identity distribution,  $x_i = \text{constant}$  (magenta line), and the uniform distribution, i.e., as a stopping criterion for the repetitions of the cABC analysis. The red lines show the boundaries between the ABC subsets “A”, “B” and “C”. The figure was created using Python version 3.10.12 for Linux (<https://www.python.org>) with the seaborn statistical data visualization package (<https://seaborn.pydata.org> [29]) and our Python package "cABCAnalysis" (<https://pypi.org/project/cABCAnalysis/>).

The brute force experiment reported above suggested that variable "Y" was consistently worse than other combinations as a basis for training any model to perform the classification task on new data. When this issue was addressed using feature selection based on permutation variable importance, variable "X" was consistently found to be important for most of the (well-performing) algorithms, while "Y" and "Z" were each important for three out of five algorithms (Figure 8), which was further addressed by testing classifiers when trained with different combinations of variables. Training the models with all variables or combinations of "XZ" or "XY" resulted in good classification performance (Figure 9), but with "XZ" slightly better than "XY". More importantly, using only variable "Y" for training resulted in significantly worse classification performance on new data. As a control, training the algorithms with permuted and thus nonsensical information led, as expected, to guessing class membership (balanced accuracy in the median around 50%), supporting that observed correct classification was not due to overfitting.

In the real-life multi-omics data set, the highest balanced accuracies of the  $k = 5$  classes identified by the authors of the original paper [12] were support vector machines, along with random forest, k-nearest neighbors, LDA and RIPPER (Figure 7). In fact, the first two algorithms currently identified as the best (SVM, random forests) were used by the authors of the original report [12]. Thus, from the range of methods available, these should be used to identify the variables relevant to the cluster structure in this data set. This will not be done again here, as it is not the intention of this report to verify the original results, but rather to use this data set as an example to discourage reliance on regression alone in this type of data set.



**Figure 9.** Comparative class assignment error rates for unseen cases (20% hold-out validation sample) by  $d = 6$  different classifiers trained with 80% of the FCPS based “ChainLink” data sets (kNN = k-nearest neighbors, SVM = support vector machines, nBayes = naïve Bayes, RF= random forests, RIPPER = repeated incremental clipping for error reduction). The classifiers were trained in a 100-fold nested cross validation setting with 2/3-sized subsets randomly drawn from the training data set separated at the beginning of the analyses. Training was performed with all  $d = 3$  variables (“full” feature set),  $d = 2$  variables that had resulted from the feature selection steps as “reduced1” feature set (variables “X” and “Z”) or as “reduced2” feature set (variables “X” and “Y”), and with the variable “Y” that had emerged as the only variable found by regression to be significant. In addition, the inference of class assignment was repeated using permuted data for algorithm training, performed with variant “reduced1”. The boxes show the 25<sup>th</sup>, 50<sup>th</sup> and 75<sup>th</sup> percentiles balanced accuracy (BA) and roc-auc for the classification performance in the 20% validation data separated before feature selection and classifier training from the whole data set and not used for feature selection and algorithm training. The figure was created using Python version 3.10.12 for Linux (<https://www.python.org>) with the seaborn statistical data visualization package (<https://seaborn.pydata.org> [29]).

## 5. Discussion

Present findings challenge the traditional reliance on regression as the standard for multivariate data analysis in biomedical research aimed at identifying variables most relevant for a given class structure. The widespread use of regression in the past may have been due to the lack of alternatives or computational limitations. However, with the advancement of machine learning techniques and the availability of computational power, it is now possible to employ a collection of different methods for data analysis and select those that accurately capture the underlying structure of the data.

This paper attempts to clarify a common misconception about regression, namely that regression algorithms for classification are as powerful as machine learning methods and therefore, suffice all requirements of the analysis of biomedical data. This is not true. By contrast, this paper points out a structural weakness of regression in separating classes in data sets, comparable to early neural networks. The breakthrough to address linear non-separable classifications of data sets was the introduction of hidden layers. In this context, regression provides the equivalent of a simple perceptron, which for good reasons has been discarded for data analysis in favor of more sophisticated model architectures. One can show that a single artificial neuron of a neural network can be described as nonlinear (logistic) regression and vice versa. However, to generalize this to “machine

learning is nothing more than the application of regression models under a different name" is a mistake. Regression models can handle the same type of classification problems as neural networks without hidden layers. This type of model was proposed by Rosenblatt in 1958 under the name "Perceptron" [30]. These algorithms were used until 1970 for many problems, including classification. In 1969, Minsky and Papert published a book showing that single-layer neural networks are, in principle, incapable of classifying data sets that are linearly non-separable [11].

The data set "Four Gaussians" illustrates this property for empirical data in a nutshell: for this type of data sets (Figure 6) there is no single straight line, i.e., a "hyperplane in two dimensions", that can correctly separate the two classes. Regression models used for classification (e.g., logistic regression) can only achieve a classification accuracy of about 50%, which is a guessing level as shown in the present analyses. Current experiments emphasize that regression is equivalent to a simple perceptron with respect to classification. A practical example of such data is a normal versus abnormal body mass index (BMI): a person is in the "normal BMI" (the green group in Figure 6) group if they have either a high weight and height or a low weight and height, otherwise they are in the "abnormal BMI" group (the blue group in Figure 6).

Thus, regression models, which are computationally equivalent to neural networks with one layer, are in principle unable to classify data sets that are linearly non-separable. Indeed, the other algorithms used in this report, except LDA, were well able to classify the "Four Gausian" data set (Figure 3 D - F). However, testing whether a given empirical multivariate data set is linearly non-separable is not an easy task and often requires the use of more powerful machine learning algorithms [31]. A simple example to demonstrate this is the "ChainLink" data set used in the present experiments (see introductory example and Figure 2). Machine learning (ML) models such as deep learning, i.e., hidden layer networks, support vector machines (SVN), random forests, and other ML algorithms can easily handle linear, non-separable data. Some of these algorithms, such as SVN, are specifically designed to handle linearly non-separable classification problems.

The current experiments have shown that relying on a single algorithm to select informative variables in a data set has potential risks in scientific findings. If the chosen algorithm fails for any reason to capture the class structure, although statistics may deliver highly significant p-values the results could be unreliable. Regression can of course be successful, as in the "Tetra" data set and many other reported in the scientific literature, however, this should not be assumed as a given fact but verified in every actual data set. The present workflow included alternative methods creating a "mixture of experts" on which conclusions can be relied with greater confidence that on a single model that has not even been tested in unseen data, as it is a frequent approach in biomedical research. This is likely to provide a result closer to the ground truth than when relying on a single model and basing research results on a particular model architecture. Furthermore, the addition of other methods provides internal validation, and there is major no disadvantage compared to a pure regression analysis.

However, the abandonment of an apparent standard that provided reassurance about the results of data analysis could be perceived as a disadvantage. In fact, selecting relevant variables requires more effort than simply accepting the statistical results of a regression. There are many methods for selecting features, of which the permutation importance method chosen here is only one [28]. These include univariate feature selection methods such as effect size calculations, the false positive rate, the family-wise error rate, selection of the k best variables based on F-statistics and others, available for example in the "sklearn.feature\_selection" module of the "scikit-learn" Python package ([https://scikit-learn.org/stable/modules/feature\\_selection.html](https://scikit-learn.org/stable/modules/feature_selection.html) [16]). Model-based methods are applied after training the algorithms and include methods such as "SelectFromModel" (SFM), which selects features based on importance weights in the trained algorithm, and (vii) Recursive Feature Elimination (RFE), which selects features by recursively considering smaller and smaller feature sets and generating a feature ranking, and many others.

Neural network based feature selection was not even considered in these experiments [32]. The choice of algorithms also influences the final result. Currently, algorithms of different types have been compiled, but alternatives are certainly possible. In addition, as shown in a simple biomedical case [33], the current setting of the hyperparameters can directly affect the classification performance and thus the ranking of the algorithms, resulting in different sets of algorithms from which to derive the relevant features. While this may be a difficult task to test, reverting to a simple single method, with the uncertainty of whether disease even captures the class structure in a dataset, does not seem to be a better solution.

The mixture of experts makes the model predictions more robust. In the present experiments, a majority vote was always included in the presentation of the results ("grand majority" columns in Figure 2, Figure 3 and Figure 4), showing that occasional weakness of one or the other model can often be compensated by the votes contributed by stronger models, although this might not be a general rule. In fact, research has shown that using a "mixture of experts" (MOE) approach, where multiple models are combined, can yield more accurate results in the analysis of biomedical data. The cited multiomics example paper was produced using support vector machines, random forests, and additionally a feed-forward neural network [12], but not regression, consistent with the present observations in other data sets and the referenced examples. Further examples include, a customized electrocardiogram (ECG) beat classifier combined with a global classifier to form an MOE classifier structure, which provided significant performance improvement over single method approaches [34]. Similarly, a wide variety of machine learning algorithms have been proposed to select the best combination of biomarkers to predict categorical outcomes from highly unbalanced (biomedical) data sets [35–37], and own molecular research also points to the advantage of using multiple models, of which regression can be one, but should not be the only one [38].

## 6. Conclusions

Regression is only one of several possible mathematical models to describe a data set; it can be valid, as in the present "Tetra" data set, but this cannot be taken for granted and used to demand that scientific reports be based on regression analysis. Regression is not necessarily always the best and most correct model, as shown in other data sets where it failed completely, as in three artificial data sets, or was one of the poorest predictors in an actual real-life multiomics data set from cancer research. This means that regression has been outperformed by most alternative machine-learning based algorithms on almost every dataset, yet it is still considered a standard that can be requested in scientific publications. A regression model needs to be validated like any other model, although many scientific papers do not do this, so it cannot be taken for granted that the results reported are valid. In contrast, this paper highlights a structural problem with logistic regression that it has in common with early neural network architectures (perceptrons), which were abandoned for good reason. Since it is difficult or even impossible to judge whether a given real-life data set is linearly non-separable, this paper calls for the use of a mixture of experts (MOE) of machine learning models, which may include regression, to handle classification problems in biomedical data. Requiring authors of scientific reports to use regression to classify a biomedical dataset, or even limiting the analysis to regression without considering alternatives, is clear malpractice.

**Author Contributions:** JL – Conceptualization of the project, programming, performing the experiments, writing of the manuscript, data analyses and creation of the figures, funding acquisition. AU – Theoretical background, mathematical implementation, writing of the manuscript.

**Funding:** JL was supported by the Deutsche Forschungsgemeinschaft (DFG LO 612/16-1).

**Institutional Review Board Statement:** Not applicable. All data sets have been taken from public sources.

**Informed Consent Statement:** Not applicable. All data sets have been taken from public sources.

**Data Availability Statement:** Data sets except the 4-Gaussians-example have been taken from public sources that are precisely referenced in the report. The data of the 4-Gaussians example and the R code for this experiment can be available at <https://github.com/JornLotsch/MisClassificationRegressionNN>. Please note that the unmodified R code runs only on Unix-like systems (e.g., Linux, MacOS) due to the particular kind of implementation of parallel processing.

**Acknowledgments:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

**Sample Availability:** Not applicable

## References

1. Lo, A.; Chernoff, H.; Zheng, T.; Lo, S.H. Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences of the United States of America* **2015**, *112*, 13892–13897. <https://doi.org/10.1073/pnas.1518285112>.
2. Ultsch, A.; Lötsch, J. The Fundamental Clustering and Projection Suite (FCPS): A Dataset Collection to Test the Performance of Clustering and Data Projection Algorithms. *Data* **2020**, *5*, 13.
3. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901**, *2*, 559–572. <https://doi.org/10.1080/14786440109462720>.
4. Team, R.D.C. *R: A Language and Environment for Statistical Computing*; Springer-Verlag New York, 2008.
5. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer-Verlag New York, 2009.
6. Ligges, U.; Mächler, M. Scatterplot3d - an R Package for Visualizing Multivariate Data. *Journal of Statistical Software* **2003**, *8*, 1–20–1–20.
7. Olsen, L.R.; Zachariae, H.B. cvms: Cross-Validation for Model Selection, 2023.
8. Gu, Z.; Eils, R.; Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **2016**, *32*, 2847–9. <https://doi.org/10.1093/bioinformatics/btw313>.
9. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. <https://doi.org/10.1023/a:1010933404324>.
10. Thrun, M.; Stier, Q. Fundamental clustering algorithms suite. *SoftwareX* **2021**, *13*, 100642–100642.
11. Minsky, M.; Papert, S. *Perceptrons; an Introduction to Computational Geometry*; MIT Press, 1969.
12. Khadiraikar, S.; Shukla, S.; Prasanna, S.R.M. Machine learning based combination of multi-omics data for subgroup identification in non-small cell lung cancer. *Sci Rep* **2023**, *13*, 4636. <https://doi.org/10.1038/s41598-023-31426-w>.
13. Ihaka, R.; Gentleman, R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **1996**, *5*, 299–314. <https://doi.org/10.1080/10618600.1996.10474713>.
14. Van Rossum, G.; Drake Jr, F.L. *Python tutorial*; Vol. 620, Centrum voor Wiskunde en Informatica Amsterdam, 1995.
15. Kuhn, M.; Max. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **2008**, *28*, 1–26. <https://doi.org/10.18637/jss.v028.i05>.
16. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
17. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **1936**, *7*, 179–188. <https://doi.org/10.111/j.1469-1809.1936.tb02137.x>.
18. Ho, T.K. Random decision forests. In Proceedings of the Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, Vol. 1, pp. 278–282 vol.1. <https://doi.org/10.1109/ICDAR.1995.598994>.
19. Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273–297. <https://doi.org/10.1023/a:1022627411411>.
20. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.* **1967**, *13*, 21–27. <https://doi.org/10.1109/tit.1967.1053964>.
21. Bayes, M.; Price, M. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions* **1763**, *53*, 370–418. <https://doi.org/10.1098/rstl.1763.0053>.
22. Cohen, W.W. Fast Effective Rule Induction. In Proceedings of the ICML, 1995.
23. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. The Balanced Accuracy and Its Posterior Distribution. In Proceedings of the Pattern Recognition (ICPR), 2010 20th International Conference on, 2010, pp. 3121–3124. <https://doi.org/10.1109/ICPR.2010.764>.
24. Peterson, W.; Birdsall, T.; Fox, W. The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory* **1954**, *4*, 171–212. <https://doi.org/10.1109/TIT.1954.1057460>.
25. Ultsch, A.; Lötsch, J. Computed ABC Analysis for Rational Selection of Most Informative Variables in Multivariate Data. *PLoS One* **2015**, *10*, e0129767. <https://doi.org/10.1371/journal.pone.0129767>.
26. Juran, J.M. The non-Pareto principle; Mea culpa. *Quality Progress* **1975**, *8*, 8–9.
27. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, fourth ed.; Springer: New York, 2002. ISBN 0-387-95457-0.

28. Guyon, I. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182. 379
29. Waskom, M.L. seaborn: statistical data visualization. *Journal of Open Source Software* **2021**, *6*, 3021. 380
30. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* **1958**, *65*, 386–408. 381
31. Elizondo, D. The linear separability problem: some testing methods. *IEEE Transactions on Neural Networks* **2006**, *17*, 330–344. 382
- <https://doi.org/10.1109/TNN.2005.860871>. 383
32. Verikas, A.; Bacauskiene, M. Feature selection with neural networks. *Pattern Recognition Letters* **2002**, *23*, 1323–1335. [https://doi.org/https://doi.org/10.1016/S0167-8655\(02\)00081-8](https://doi.org/https://doi.org/10.1016/S0167-8655(02)00081-8). 384
33. Löttsch, J.; Mayer, B. A Biomedical Case Study Showing That Tuning Random Forests Can Fundamentally Change the Interpretation of Supervised Data Structure Exploration Aimed at Knowledge Discovery. *BioMedInformatics* **2022**, *2*, 544–552. 385
- <https://doi.org/10.3390/biomedinformatics2040034>. 386
34. Hu, Y.H.; Palreddy, S.; Tompkins, W.J. A patient-adaptable ECG beat classifier using a mixture of experts approach. *IEEE Trans Biomed Eng* **1997**, *44*, 891–900. <https://doi.org/10.1109/10.623058>. 387
35. Leclercq, M.; Vittrant, B.; Martin-Magniette, M.L.; Scott Boyer, M.P.; Perin, O.; Bergeron, A.; Fradet, Y.; Droit, A. Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. *Front Genet* **2019**, *10*, 452. 388
- <https://doi.org/10.3389/fgene.2019.00452>. 389
36. Miettinen, T.; Nieminen, A.I.; Mäntyselkä, P.; Kalso, E.; Löttsch, J. Machine Learning and Pathway Analysis-Based Discovery of Metabolomic Markers Relating to Chronic Pain Phenotypes. *Int J Mol Sci* **2022**, *23*. <https://doi.org/10.3390/ijms23095085>. 390
37. Kringsel, D.; Kaunisto, M.A.; Kalso, E.; Löttsch, J. Machine-learned analysis of global and glial/opioid intersection-related DNA methylation in patients with persistent pain after breast cancer surgery. *Clin Epigenetics* **2019**, *11*, 167. <https://doi.org/10.1186/s13148-019-0772-4>. 391
38. Löttsch, J.; Schiffmann, S.; Schmitz, K.; Brunkhorst, R.; Lerch, F.; Ferreiros, N.; Wicker, S.; Tegeder, I.; Geisslinger, G.; Ultsch, A. Machine-learning based lipid mediator serum concentration patterns allow identification of multiple sclerosis patients with high accuracy. *Sci Rep* **2018**, *8*, 14884. <https://doi.org/10.1038/s41598-018-33077-8>. 392

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content. 403

404  
405  
406