Exploring Classification Tools for the Prediction of in-Vitro Microbial Pyruvate Yield From Organic Carbon Sources

Manish Pant (manishpant4444@gmail.com)
Govind Ballabh Pant University of Agriculture and Technology
Tanuja Joshi
Kumaun University
Pramod Kumar Omre
Govind Ballabh Pant University of Agriculture and Technology
Triveni Prasad Singh
Govind Ballabh Pant University of Agriculture and Technology
Anil Kumar
Govind Ballabh Pant University of Agriculture and Technology
Hemlata Pundir
Kumaun University
Subhash Chandra
Kumaun University

Research Article

Keywords: Machine learning, Classification, Learning curve, Multilayer perceptron, Pyruvate

Posted Date: January 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1232136/v1

Exploring classification tools for the prediction of *in-vitro* microbial pyruvate yield from organic carbon sources

Manish Pant[a]*, Tanuja Joshi[b], Pramod Kumar Omre[a], Triveni Prasad Singh[a], Anil Kumar[a], Hemlata Pundir[b], Subhash Chandra[b]

[a]Govind Ballabh Pant University of Agriculture and Technology, India
[b]Kumaun University, India
*Corresponding Author.

Abstract

The *in-vitro* microbial production is based on screening of parameters and statistical validation of parameters using regression analysis. The recent trends have shifted from full factorial design towards more complex response surface methodology designs such as Box-Behnken design, Central Composite design. The problem with these designs is apart from optimisation, the designs are not flexible enough in deducing properties of parameters in terms of class variables. Machine learning algorithms have unique visualisations for the dataset presented with appropriate learning algorithms. The classification algorithms cannot be applied on all datasets and selection of classifier is essential in this regard. The aim of the current study was to investigate the data-mining accuracy on the dataset developed using *in-vitro* pyruvate production using organic sources for the first time. The attributes were subjected to comparative classification on various classifiers and based on accuracy, multilayer perceptron was selected as classifier. As per the results, the model showed significant results for prediction of classes and a good fit. The learning curve developed also showed the datasets converging and were linearly separable.

Keywords: Machine learning, Classification, Learning curve, Multilayer perceptron, Pyruvate.

Results

Experimental summary. In the experimental procedure, Corn Steep Liquor (CSL) was used as nitrogen alternative in the culture medium. Moreover, two carbon sources viz. rice straw and jackfruit rind were selected as the carbon alternative to glycerol for the comparisons in overall pyruvate yield [1]. The different carbon sources based media formulations were investigated separately on varied levels of CSL concentration (%v/v) accompanied with processing parameters i.e. incubation time (h) and incubation temperature (°C) in the full factorial design. Pyruvate concentration (g/L), reducing sugars (g/L), and biomass concentration (g/L) were considered as the response variables. All experimental runs were performed independently in triplicates and the average values were analysed.

$$Number\ of\ experiments\ for\ each\ carbon\ source\ = 5\ levels \times 3\ levels \times 3\ levels = 45\ experiments$$

The individual run in the experiments were considered as data points for the three datasets in each of the carbon sources, respectively (Supplementary Table S1.).

Table 1. Statistical summary of classifier algorithms on the dataset (glycerol)

Algorithms/Scheme	Statistics summary						
	Correctly classified instances	Incorrectly classified instances	Kappa value	Mean absolute error	RMSE	ROC area	PRC value
KStar	3 (6.6667%)	42 (93.3333%)	-0.1667	0.3408	0.4566	0.420	0.199
Multilayer perceptron	31 (68.8889%)	14 (31.1111%)	0.6111	0.1898	0.3062	0.891	0.728
REPTree	13 (28.8889%)	32 (71.1111%)	0.1111	0.2962	0.3999	0.575	0.289
BayesNet	1 (2.2222%)	44 (97.7778%)	-0.2222	0.3381	0.4313	0.233	0.167
IBk	0 (0%)	45 (100%)	-0.25	0.3912	0.5983	0.375	0.185
RandomForest	3 (6.6667%)	42 (93.3333%)	-0.1667	0.3393	0.4647	0.448	0.230
SimpleLogistic	21 (46.6667%)	24 (53.3333%)	0.3333	0.2254	0.3383	0.817	0.565
SMO	1 (2.2222%)	44 (97.7778%)	-0.2222	0.3644	0.4713	0.211	0.160
J48	8 (17.7778%)	37 (82.2222%)	-0.0278	0.3077	0.447	0.534	0.240
DecisionStump	13 (28.8889%)	32 (71.1111%)	0.1111	0.2869	0.3817	0.557	0.265
RandomTree	11 (24.4444%)	34 (75.5556%)	0.0556	0.3022	0.5497	0.528	0.253
ZeroR	5 (11.1111%)	40 (88.8889%)	-0.1111	0.3208	0.401	0.414	0.185
PART	7 (15.5556%)	38 (84.4444%)	-0.0556	0.3173	0.4572	0.503	0.219
OneR	12 (26.6667%)	33 (73.3333%)	0.0833	0.2933	0.5416	0.542	0.220
DecisionTable	4 (8.8889%)	41 (91.1111%)	-0.1389	0.3208	0.4506	0.468	0.214
JRip	11 (24.4444%)	34 (75.5556%)	0.0556	0.2893	0.3992	0.553	0.297
LWL	3 (6.6667%)	42 (93.3333%)	-0.1667	0.3153	0.4241	0.464	0.232

Screening of classifiers. The unbiased estimate of a classifying model's efficiency, the statistical measures have to be evaluated from the dataset used in the model building process [2]. Common statistical measures of classification are sensitivity, specificity, accuracy and the area under the ROC curve (or, equivalently, the c-index). These values are essential in the determination of a comparatively superior for model building process [3]. The class variable, CSL concentration (class value '1') was used to investigate the accuracy of various applicable classifiers. The experimental design was screened and modelled with glycerol as standard carbon source. The classifier comparison was investigated for pyruvate concentration with results using *in-vitro* experimental data. The performance of each classifier was evaluated with statistical summary as shown in **Table 1** representing model performance chart (Supplementary Figure S1). The classifier selected at this step was applied to the other carbon sources for classification of CSL concentration. **Table 1** represents important statistical measures from the text viewer, such as correctly classified instances, incorrectly classified instances, Cohen's kappa value, Receiver Operating Characteristic (ROC) area and Precision Recall Curve (PRC) value for the classifier algorithms.

$$Cohen's\ Kappa\ Statistic\ (\kappa) = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e}$$

P_o is the relative observed agreement among raters and P_e is the hypothetical probability of chance agreement. As shown, the highest accuracy corresponds to multilayer perceptron (68.8889%) under 10X cross validation fold-maker, followed by simple logistic (46.6667%) and the lowest for IBk (0%). It is shown that simple logistic regression predictive classifier is weaker than MLP in terms of data mining accuracy. Generally, both models perform on about the same level with the more flexible neural networks generally outperforming logistic regression [4]. In this experimental comparison, we can say that multilayer perceptron (MLP) was the best scheme in all applicable classifiers. Thus, multilayer perceptron (MLP) was selected as the classification algorithm for three carbon sources to create models and summary for predictive classification of datasets in the microbial production. The generic object editor of multilayer perceptron was built with hidden layers as wildcard value "a" (attributes + classes), learning rate as 0.3, momentum 0.2 and validation threshold 20.

Multilayer perceptron. The multilayer perceptron algorithm consists of a simple system of artificial neurons connected by weights and output signals, which are a function of the sum of inputs for the modified neuron from a linear activation function [5] and is described as.

$$y(v_i) = \tanh(v_i)\ \ and\ y(v_i) = (1 - e^{-v_i})^{-1}$$

The hyperbolic tangent function ranges from -1 to 1, and the logistic function is similar in shape but ranges from 0 to 1. Here y_i is the output of the i^{th} node (neuron) and v_i is the weighted sum of the input connections. The network is divided into three layers: input, hidden, and output. The input layer receives the value vector for network initialization, the hidden layer performs training, and the output layer receives the output vector [6]. The main adjustable parameters are the maximum amount of iterations, learning rate, momentum, and the number of neurons in the hidden layer [7]. The datasets from all the carbon sources were modelled, measured and analysed with training set validation under multilayer perceptron from the classifier tab. The graphic user interface with intermittent weighted functional nodes in multilayer perceptron provides model summary, accuracy by class and confusion matrix developed under multilayer perceptron and the statistical measures for all carbon sources.

Table 2. Model summary developed under multilayer perceptron

	Glycerol	Rice straw	Jackfruit rind
Correctly classified instances	93.34	91.12	97.78
Cohen's Kappa statistic	0.9167	0.8889	0.9722
Root mean squared error	0.2252	0.2103	0.1424
Precision	0.944	0.918	0.980
Recall	0.933	0.911	0.978
ROC area	0.978	0.967	0.994

The design model summary **(Table 2)** for glycerol, rice straw and jackfruit rind showed significant results with correctly classified instances as 93.34%, 91.12% and 97.78% respectively (Supplementary Figure S2). The Cohen's kappa statistic value also showed strong (0.8-0.9) to almost perfect (above 0.9) results as 0.9167, 0.8859 and 0.9722 respectively. For a model to be considered a good fit, the values of root mean squared error should be less than 0.5 [8]. For the three carbon sources, the values of root mean error were 0.2252, 0.2103 and 0.1423, respectively. Hence, the classifier model developed for the three carbon sources is significant, robust measure for inter-rater reliability of attributes, and is able to predict the data accurately. To investigate the overall accuracy of the model, detailed accuracy by class is observed using precision, recall and ROC area value [9]. For the model, to predict accurately ROC area value 0.7-0.8 is considered acceptable, 0.8-0.9 is excellent and greater than 0.9 is considered outstanding [10]. The results for detailed accuracy by class for the three carbon sources show significant results for multilayer perceptron (Supplementary Table S2 - S4). The weighted average value of ROC area for glycerol, rice straw and jackfruit rind is 0.978, 0.967 and 0.994, respectively. The high value of ROC area (> 0.9) represents the model for the prediction of classes in CSL concentration with variable pyruvate concentration from the dataset obtained experimentally. Hence, it can be considered appropriate in terms of accuracy for the respective classes in the dataset. Similarly, weighted average results for precision and recall for the three carbon sources were 0.944 and 0.933, 0.918 and 0.911, and 0.980 and 0.978, respectively. Thus, the model shows an accreditation of multilayer perceptron in practical applications for full factorial experimental design. The training set modelled multilayer perceptron presented the confusion matrix with high accuracy and low error (Supplementary Table S5 - S7).

Model validation. The learning curve (training set size vs error rate) represents the rate of performance of the model when a percentage of data is removed from the dataset [11]. In contrast, ROC curve does not show learning instead it shows performance of the overall model [12]. Thus, to understand the ability of a model to learn the effect of attributes of dataset, learning curve is an essential constraint. For a model to be a good fit, the goal of the learning algorithm is to exist between an under fit and over fit. A good fit is identified by a training and validation loss that decreases to a point of stability with a minimal gap between the two final loss values [13]. The datasets from the three carbon sources is used in investigating the learning curve for the classifier and is shown in **Figure 1.**

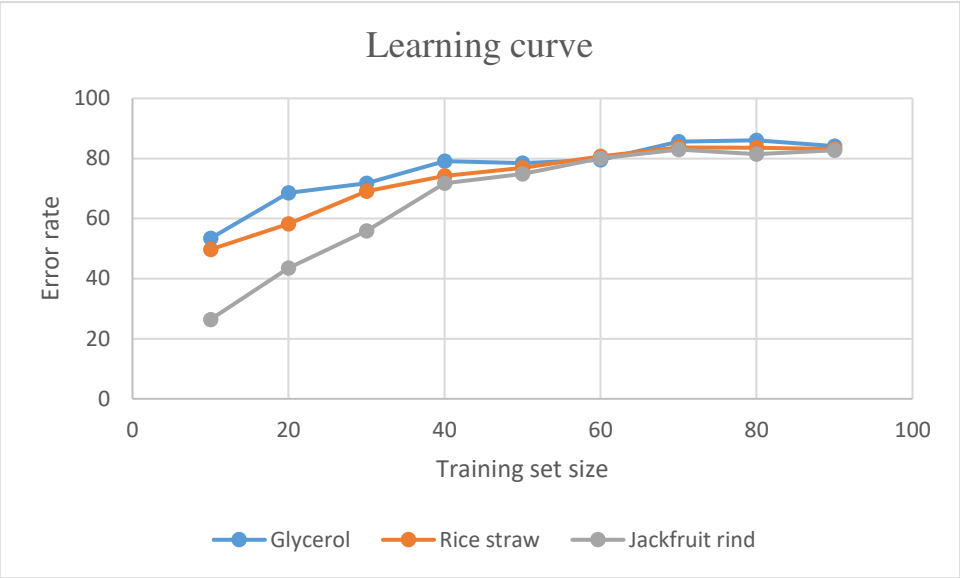

Figure 1. Learning curve for the three carbon sources

In **Figure 1**, it is shown that the learning curve converges for the three carbon sources and the datasets are linearly separable. Thus, we can say that the one hidden layer is appropriate in the multilayer perceptron classifier and the model will not benefit from adding more data. Moreover, the error rate increases with the data removed from the dataset and it shows the invariability for model developed in multilayer perceptron. Thus, with the point of stability in the final two loss values, the model is considered to be a good fit. Moreover, continued training of a good fit model will likely lead to an over fit [14]. **Figure 2(a)** shows the margin curve for the three carbon sources. The margin is defined as the difference between probability predicted for the actual class and the highest probability predicted for other classes [15]. A margin of value 1 means that the correct class was predicted with 100% confidence and a margin of value -1 means an incorrect class was predicted with similar confidence. As per the results, the classifier is confident in predicting the true classes. The central values of margin curve for glycerol, rice straw and jackfruit rind are -0.021, -0.047, and -0.051, respectively. **Figure 2(b)** represents the classifier errors plotted with class variable against the predicted class variable. The graph shows instances incorrectly classified with different colours for different class levels. **Figure 2(c)** represents the cost curve of classifier for the expected cost in three carbon sources. The cost curve plots the expected cost of using classifier against the probability cost function, which is a partisan type of $P(+)$ and is at the same extremes: zero when $P(+) = 0$ and one when $P(+) = 1$ [16]. It is denoted by $C[+|-]$ the cost of predicting + when the instance is actually −, and the contrary by $C[-|+]$. The axis are described as

$$Normalised\ expected\ cost\ =\ fn\ \times\ Pc(+) + fp \times\ (1 - Pc(+))$$

$$Probability\ cost\ function\ =\ Pc(+) = \frac{P(+)C[-|+]}{P(+)C[-|+] + P(-)C[+|-]}$$

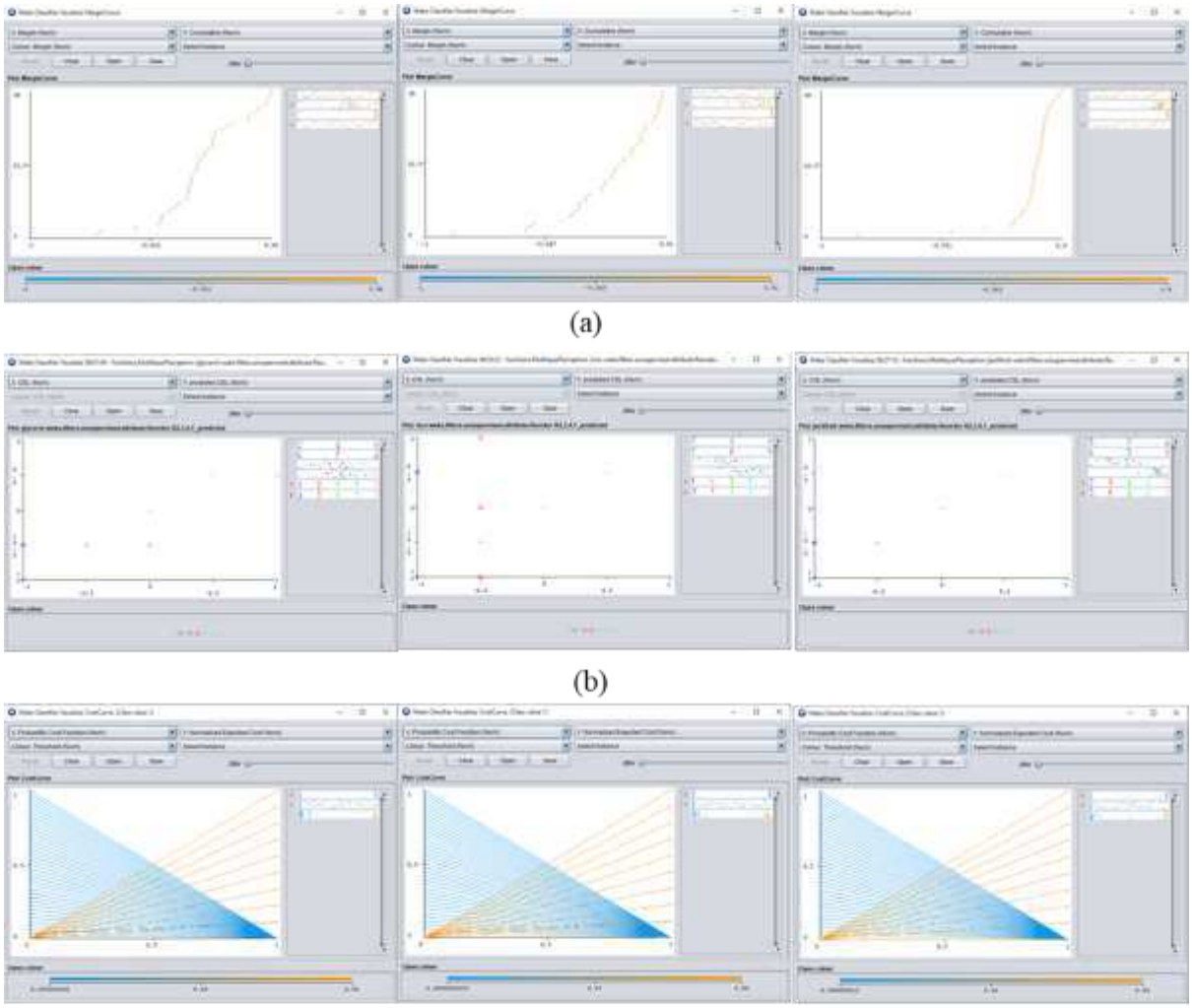

Figure 2 (a) Margin curve (b) Classifier errors (c) Cost curve; in glycerol, rice straw and jackfruit rind (from left to right)

The cost curve plots error rate as a function of $p(+)$. Error rate is the y-axis in the plot, $p(+)$ is the x-axis. The extreme values on the x-axis represent the situations where all the iterations to which the classifier will be applied are in the same class. If, $x = p(+) = 0$ means that all these iterations are negative, and if $x = p(+) = 1$ means they are all positive. When $x = 0$ a classifier's overall error rate is simply its error rate on the negatives, since no positive iterations will be presented to the classifier. When x = 1 its overall error rate is its error rate on the positives. Joining the two points by a straight line plots its overall error rate as a function of $p(+)$. This presents a single classifier which corresponds to straight line and shows how the performance varies as class distribution changes. The cost curve are the plots with expected error against the probability of one of the classes. The diagonals show the performance of two extreme classifiers, The first predicting '+' giving an expected error of one if the dataset contains no '+' instances and zero if all instances are '+'. The second predicts '-' which represents the contrary (always wrong) performance of classifier. A good classifier have low error rates with diagonals as close to the bottom as possible. If P (+) is closer to the horizontal axis then the predictor has outperformed the classifier which always predicts '-' and vice versa.

Methods

Classification. The full factorial design based glycerol dataset for comparison of classifiers was built in Microsoft Excel in the ARFF (Attribute-Relation File Format) as glycerol.arff file. The glycerol dataset was directly loaded in the pre-process tab of WEKA tool. The ARFF datasets for the two organic carbon sources (rice straw and jackfruit rind) were built and loaded independently in the pre-process tab. CSL concentration (nominal attribute) was pre-processed as the class variable for the classifier prediction of the numeric dataset (pyruvate concentration). The selection of CSL concentration as the class variable because, it acts as the replacement of inorganic nitrogen macro-nutrient in the media formulation. The processing conditions such as incubation time and incubation temperature are easily optimised in the fermenter. Moreover, the input variable in the fermenter such as inflow rate can be modelled as a class variable in real world problems associated with product yield. The response variables are results observed as the significant/non-significant values associated with the variations in input variables. Moreover, any changes in the independent variable provides a causal effect relationship on dependent variables [17]. The converse may also be true, changes in responses can be monitored with changes in input variables and any combination thereof.

Mathematically,

$$f(a,b,c) = d;\ f(a,b,d) = c;\ f(a,c,d) = b;\ f(b,c,d) = a$$

Thus, utilisation of input variables as class variables can help in classification model development in the laboratory experiments. Moreover, the number of experimental runs are generally limited, and the utilisation of full factorial design is crucial in this regard.

Comparison of classifiers. Classification uses the application of standard algorithms to clearly describe, distinguish discrete classes for the model and to predict each iteration in the dataset under experimental conditions [18]. However, the classification algorithms used to solve various problems is extremely diverse. Hence, it is imperative to study in order to deduce the most suitable classifier algorithm for application in the target problems [19]. In this study, all possible classifiers were investigated in terms of classifier accuracy using knowledge flow environment of WEKA tool. The statistical measures (highest correctly classified instances, ROC area, etc.) were used in the screening and selection of classifier for model development (Supplementary Figure S3). The classifier selected at this step was applied to the other carbon sources for classification of CSL concentration.

Model validation. After screening of classifiers, the classifier with highest accuracy i.e. multilayer perceptron was used to develop model for the three carbon sources and the standard statistical measures are represented. The linear separability can be investigated for the classification of data using the learning curve (Supplementary Figure S4). If the classification of data is plotted on a single line and the classes can be separated by a single point it is said to be linearly separable [20]. The learning curve for the screened classifier was built on percent incorrect with confidence level less than 5%. Finally, the classifier was validated using margin curve, classifier errors and cost curve visualisations for the three datasets.

Author contributions statement

M.P.; Investigation/Writing Original draft, review, editing. T.J.; Review, Editing. P.K.O.; Review, Editing, Supervision. T.P.S.; Review& Editing, Supervision. A.K.; Review& Editing, Supervision. H.P.; Review, Editing. S.C.; Review, Editing.

Additional information

Supplementary Information: The article contains supplementary material.

Competing interests

The authors declare no competing interests

References

1. Pant, M., Joshi, T., Omre, P. K., Awasthi, P. & Chauhan, L. Development of novel media formulation using rice straw and CSL for pyruvate production. *Biomass Convers. Biorefinery* (2021) doi:10.1007/s13399-021-02057-w.

2. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. *PLoS One* **14**, e0224365 (2019).

3. Cowley, L. E., Farewell, D. M., Maguire, S. & Kemp, A. M. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagnostic Progn. Res.* **3**, 1–23 (2019).

4. Costela, F. M. & Castro-Torres, J. J. Risk prediction model using eye movements during simulated driving with logistic regressions and neural networks. *Transp. Res. Part F Traffic Psychol. Behav.* **74**, 511–521 (2020).

5. Bhojani, S. H. & Bhatt, N. Wheat crop yield prediction using new activation functions in neural network. *Neural Comput. Appl.* **32**, 13941–13951 (2020).

6. Zhou, H., Myrzashova, R. & Zheng, R. Diabetes prediction model based on an enhanced deep neural network. *Eurasip J. Wirel. Commun. Netw.* **2020**, 1–13 (2020).

7. Leema N., Nehemiah, K. H., Elgin Christo V. R. & Kannan A. Evaluation of Parameter Settings for Training Neural Networks Using Backpropagation Algorithms. *Int. J. Oper. Res. Inf. Syst.* **11**, 62–85 (2020).

8. Bakker, K., Whan, K., Knap, W. & Schmeits, M. Comparison of statistical post-processing methods for probabilistic NWP forecasts of solar radiation. *Sol. Energy* **191**, 138–150 (2019).

9. Deepak, S. & Ameer, P. M. Brain tumor classification using deep CNN features via transfer learning. *Comput. Biol. Med.* **111**, 103345 (2019).

10. Nini, A. *et al.* Preoperative clinical and radiographic predictors of major vascular surgery in patients with testicular cancer undergoing post-chemotherapy residual tumor resection (PC-RPLND). *World J. Urol.* 1–6 (2021) doi:10.1007/s00345-021-03870-8.

11. Chen, I. Y., Johansson, F. D. & Sontag, D. Why is my classifier discriminatory? *Adv. Neural Inf. Process. Syst.* **2018-Decem**, 3539–3550 (2018).

12. Wehbe, R. M. *et al.* DeepCOVID-XR: An artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large U.S. Clinical data set. *Radiology* **299**, E167–E176 (2021).

13. Smith, L. N. A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay. *arXiv Prepr. arXiv1803.09820* (2018).

14. Feng, X., Tustison, N. J., Patel, S. H. & Meyer, C. H. Brain Tumor Segmentation Using an Ensemble of 3D U-Nets and Overall Survival Prediction Using Radiomic Features. *Front. Comput. Neurosci.* **14**, 25 (2020).

15. Moon, J., Kim, J., Shin, Y. & Hwang, S. Confidence-Aware learning for deep neural networks. in *37th International Conference on Machine Learning, ICML 2020* vols PartF168147-9 6990–7000 (PMLR, 2020).

16. Bertozzi, A. L., Luo, X., Stuart, A. M. & Zygalakis, K. C. Uncertainty quantification in graph-based classification of high dimensional data. *SIAM-ASA J. Uncertain. Quantif.* **6**, 568–595 (2018).

17. Benitez, J., Henseler, J., Castillo, A. & Schuberth, F. How to perform and report an impactful analysis using partial least squares: Guidelines for confirmatory and explanatory IS research. *Inf. Manag.* **57**, 103168 (2020).

18. Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M. & Moore, J. H. Benchmarking relief-based feature selection methods for bioinformatics data mining. *J. Biomed. Inform.* **85**, 168–188 (2018).

19. Aggarwal, C. C. & Reddy, C. K. Data clustering. *Algorithms Appl. Chapman&Hall/CRC Data Min. Knowl. Discov. Ser. Londra* (2014).

20. Cohen, U., Chung, S. Y., Lee, D. D. & Sompolinsky, H. Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* **11**, 1–13 (2020).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.docx