International Journal of Statistics and Probability; Vol. 14, No. 1; March 2025
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Comparisons of Supervised Artificial Neural Networks With Population-Based Statistical Probability Models in Moderate Sized Samples

M. Brimacombe

Correspondence: M. Brimacombe, CT Children's and Department of Pediatrics, University of Connecticut School of Medicine, USA

Received: February 26, 2025 Accepted: March 28, 2025 Online Published: March 31, 2025

doi:10.5539/ijsp.v14n1p58 URL: https://doi.org/10.5539/ijsp.v14n1p58

Abstract

Some of the basic issues affecting the comparison of population based statistical models and data-centric artificial neural network machine learning models are reviewed in moderate sized data samples. Comparisons of artificial neural networks and population-based models should consider and reflect both the data-centric and probability based nature of the models being compared. This is examined in a series of examples. Some guidelines for developing useful comparative settings are given. Improving the understandability of machine learning methods is an important goal.

Keywords: artificial neural networks, comparartive inference, logistic regression, probability models, statistical inference

1. Introduction

The use of machine learning and related artificial intelligence (A.I.) data-analytic approaches in the development of predictive models in medical and scientific research has recently grown substantially, providing an alternative, to some extent, to the pre-experimental perspective of population-based probability models and related statistically designed experiments (Sidey-Gibbons J. and Sidey-Gibbons C., 2019). This data-centric emphasis has engendered much discussion on a related series of topics: the broadening of data analytic tools, the importance of designed experiments and studies, prediction versus understanding, the nature and relevance of probability and population-based inference, the related use of mathematical models in science, and the interpretation of supervised machine-learning approaches (Donoho, 2017; Nielsen, 2015; Casella and Berger, 2002; Brimacombe, 2019).

Comparing the various approaches to developing predictive models has become an area of research interest (Spreafico, Hazewinkel, Van de Sande, 2024; Kontopoulou, Panagopoulos, Kakkos, Matsopoulos, 2023).

2. Statistical Approaches

Statistical approaches to modeling data have a history that begins before the onset of the computer as a tool of research. The processing of large amounts of data into information from a statistical perspective has typically required the defining of a model, often a probability model, within which information can be interpreted. In the 1920's the statistician and population geneticist R.A. Fisher, following the earlier work of the mathematician Laplace and population geneticist K. Pearson, defined the statistical likelihood function and extended the idea of sufficient statistics to guide the processing of raw data into information that can be used to estimate and test the parameters of a probability based model (Casella and Berger, 2002), often viewed as population characteristics. The statistics employed were often aggregate measures, for example sample means, variances and proportions and these had sampling distributions. Specific parameter values could then be tested using Fisher's p-value concept, and estimated using confidence intervals. Constructs such as the ANOVA table, least squares geometry and the likelihood function helped define and popularize the population-based approach

and its measures of significance across a wide variety of scientific and medical areas of research (Brimacombe, 2019; Brimacombe, 2024).

The frequentist approach to statistics and probability focuses on pre-experimental design and the avoidance of systematic long-run bias in statistical estimates, as well as pre-experimentally defined probability concepts and models based on frequentist probability ideas for a population from which the sample dataset will be drawn. Type I error, p-values, power and sample size justifications, the design of experiments, predictive models and 95% confidence intervals all reflect this frequentist or repeated sampling perspective.

The Bayesian statistical perspective is also defined in the context of population-based probability models (Box and Tiao, 1973). In this setting a parametric probability model is also assumed, but the probability is placed directly on the parameter space describing the belief of the researcher in relation to possible model parameter values, which are not directly observed. Bayes theorem allows researchers to update these beliefs using the statistical likelihood function (to summarize model-data information) and related probability calculations which are formally conditioned on the observed data. The approach requires a baseline assessment of existing belief regarding possible model parameter values, a prior density, as part of the model building process. There is often a need for extensive numerical integration to obtain marginal inferential elements for specific model parameters, usually in the form of 95% credible regions, posterior odds ratios and Bayes factors (Congden, 2003). Both the standard frequentist approach and the related Bayesian-likelihood approaches focus on population-based inference with model parameters viewed as representing population characteristics.

Note that recent use of resampling methods via the bootstrap (Efron and Hastie, 2021). has increased the ability of frequentist methods to more formally incorporate observed data into statistical inference descriptions and procedures.

Early on in the application of applied statistics, data-based methods such as linear discriminant functions, eigenvector based principal components, classification trees, cluster analysis and other data-centric aspects of statistical scientific study were developed, but often viewed as secondary and "hypothesis generating" (Donoho, 2017), not particularly useful in regard to formal inferences, which were to be based on pre-experimentally defined population models, study designs and probabilities.

3. Machine Learning Approaches

With the computer now a primary tool and central element of research, along with access to very large amounts of data, data scientists, statistician, computer scientists and engineers have begun developing primarily data-centric approaches to the analysis of patterns in data, for example machine learning and related complex algorithmic methods of inference. This has lead to the application of computer based methodologies and algorithms to very large databases, for example, datasets generated by genomic data chips (Dennise, Dalma-Weiszhausz, Warrington, Tanimoto, Miyada, 2006) and internet data flows (Zhang et al., 2023) which provide, instantaneously, thousands of variables and hundreds of thousands of measurements. This type of data may not reflective of a carefully designed experiment or study protocol (though it can be), and the opportunity for misleading data and thus misleading inferences can be an issue (Ching et al., 2018).

Under the various names of data mining, BigData, A.I., data science and machine learning, these data-centric methods have been developed for the fitting of flexible models with algorithms based entirely on the large databases themselves as the point of reference, not necessarily related to population-based probability models and a reference population. Indeed, when comparing statistical and machine learning models, it is useful to note that statistical approaches typically use probability models to generalize predictive inferences to a population, using the information in a given sample. Machine learning models measure the accuracy of predictions by comparing the agreement of predictive classifications defined for two randomly defined portions of the observed dataset.

A commonly used machine learning approach is supervised artificial neural networks (ANNs). These use multi-nodal networks with many hidden layers as a basic model or template to achieve predictions regarding the response of interest in relation to a set of explanatory variables (supervised learning). While statistical models are often parsimonious in the

number of parameters to be considered in a model, closely related to the number of variables and correlations among the variables, ANN models use stochastic gradient and related back-fitting algorithms with a nonlinear iterative structure and involve far greater numbers of secondary fitting related parameters in the model fitting process (Higham and Higham, 2019). This, along with the iterative nature of the fitting process and the multi-nodal form of the underlying model, tends to give such models a black-box aspect to their interpretation. Such models however apply to a very flexible and wide set of data types (Krizhevsky, Sutskever, Hinton, 2012).

Initially these methods did not work very well, but over time, using back propagation fitting methods along with stochastic gradient based approaches (Hardt, Recht, Singer, 2016) the convergence of these approaches improved. There remain issues regarding general convergence and stability in the case of neural networks (Colbrook, Antunb, Hansen, 2022), especially in regard to predictive, supervised ANNs.

The stability of these data-centric methods is not simple to assess. The most directly interpretable approach is cross validation of the model, though this has challenges (Krstajic, Buturovic, Leahy, Thomas, 2014). The basic ANN model validation process is based on randomly splitting the dataset into training and testing components with the model fitted on the training data subset and subsequently validated on the testing data subset. The final underlying model or black box is inferentially specific to the dataset at hand and typically offers limited insight into the contribution of individual variables.

The data-centric nature of this approach is not something entirely new, for example, in medical research. Some clinical variables themselves are defined in such a data-centric manner. For example, the Beck Depression Index (Elovanio et al., 2020). measure of depression must be standardized to the particular clinic being used as the basis for clinical subjects in the study. In a similar manner, machine learning based results in medicine can be heavily dependent on the particular hospital or clinic population from which the dataset is being drawn (Morgan et al., 2019).

4. Comparative Inference

Discussions of how to compare or integrate machine learning based methods in relation to more standard statistical approaches to data analysis have arisen as data-centric methods have been widely disseminated through data science courses, multiple departments and widely reported applications (Efron, 2024; Bi, Goodman, Kaminsky, Lessler, 2019).

Statistical methods employ probability models, defined pre-experimentally, to model and interpret data, often in tandem with formally designed experiments. The two basic approaches reflect a pre-experimental focus on expected behavior of statistical estimation and testing procedures (frequentist statistics) or a focus on learning behavior and the use of likelihood based statistics and the observed data to update beliefs regarding model parameters (Bayesian statistics). These are distinct perspectives in relation to probability.

Machine learning, as noted above, often provides an assessment of predictive accuracy by comparing the agreement of predictive classifications defined for two randomly defined portions of the observed dataset, training the model on the training portion of the data and predicting the response of interest in the testing portion of the dataset. While this is nominally predictive, it is really measuring agreement of the basic data patterns of interest between the two randomly defined portions of the data. This is not a formal probability based analysis, more a measure of agreement. As well, the black box interpretation of machine learning generally limits this type of information available for individual parameters (a type of information available in most statistical approaches) due to the highly nonlinear nature of the model and the related nonlinear, iterative nature of the training/fitting procedure (Higham and Higham, 2019).

Thus comparing the accuracy and related variability among statistical and machine learning based approaches can be challenging. In the context of ANN models, the use of the sigmoidal (logistic) function to map the output to a (0, 1) scale allows for application of basic assessments similar to those found in logistic regression procedures. These include the ROC curve, AUC, sensitivity, specificity, false positives, overall classification accuracy.

To adequately compare machine learning to frequentist methods, it is also useful to randomly replicate the training/validation

sampling process for machine learning approaches several times, reporting the range or median values of the model output values obtained. Median values are reported for ten repetitions of each fitting process.

Here, using a series of well-known, small to moderate sized datasets that show challenging levels of random variation, the level of accuracy in estimation and predictive classification was examined to see if the comparative accuracy alters as the analysis moves from statistical probability-based model to machine-learning data-centric approaches, where training portions of the database were used to predict patterns in the remainder or testing component of the database.

Using logistic regression as a primary context, the stability of the population based models in frequentist and data-conditional Bayesian settings was assessed by comparing overall mean squared error, and for individual parameters, confidence interval lengths and Bayesian credible interval lengths. Once the stability of the statistical procedures was observed, frequentist logistic regression was then directly compared to supervised ANN models using diagnostic classification errors including ROC and AUC. Underlying randomness in the ANN assessment procedure was provided by carrying out the training/testing random split ten times and reporting median and range values. STATA (version 16) and R (version 4.1.2) were used for all calculations and figures.

In some previous comparative studies, often in the form of simple comparisons between fitted logistic regression models and supervised machine learning methods (Bisaso, Karungi, Kiragga, Mukonzo, Castelnuovo, 2018), the data-centric approach has been shown to be slightly better in terms of classification error measures. But often these comparisons do not adequately include consideration of sample-to-sample variation, reflect very large sample sizes, or incur selection bias in relation to probability-based logistic regression models. The logistic regression in question is often evaluated using the training data to initially fit the model, and the testing data used to validate the model. This may seem to provide a fairer basis for comparison with ANN models, but logistic regression models are primarily population-based probability models with assumptions reflecting this context. Placing logistic regression models (generalized linear models) in the setting of observed data (training and testing) should only be seen as a type of sensitivity assessment. Statistical models are built to reflect averaged results over a set of possible future outcomes, modeled using probability based model structures. They would be built differently if they were only to be defined and applied in a data-centric setting. Standard logistic regression models are reported here.

Note that sources of variation typically occur at the individual variable and subject level and this is the basis of probability based modeling in the frequentist statistical setting. The ANN model takes the set of data values and variables (features) as given and randomness is generated by the random training/testing split of the dataset.

Recent work applying and comparing ANN models with logistic regression models in modeling viral infections, including COVID-19-related conditions, can be found in (Ansari and Baker, 2021; Haleem, Javaid, Khan, 2019) and the references therein.

5. Model Structure

In the 1920's and 1930's modern statistical linear models and associated ANOVA decompositions were developed to allow for parameter driven probability and population based models and the interpretation of individual parameter and associated variables. Exponential families were especially emphasized since the number of sufficient statistics that could be defined within a likelihood function setting could be matched to the number of unknown parameters in the underlying probability model (Pitman, 1936).

Machine learning data-centric models do not reflect such structures. For example, ANN models are deliberately based on multi-nodal networks of variables with many parameters, many of which do not correspond to individual variables. This highly recursive and nonlinear structure gives no simple linear equation as an output with individual parameters and variables being tested for relevance. This approach mimics, at some level, the perceived multi-nodal structure of the brain (Higham and Higham, 2019), with some adherence to the logistic function based structure that underlies the output of logistic regression models. Various weighting schemes provide a means by which many additional parameters can be

placed in the model, far more than the number of variables considered. A typical output is given in Figure 1.

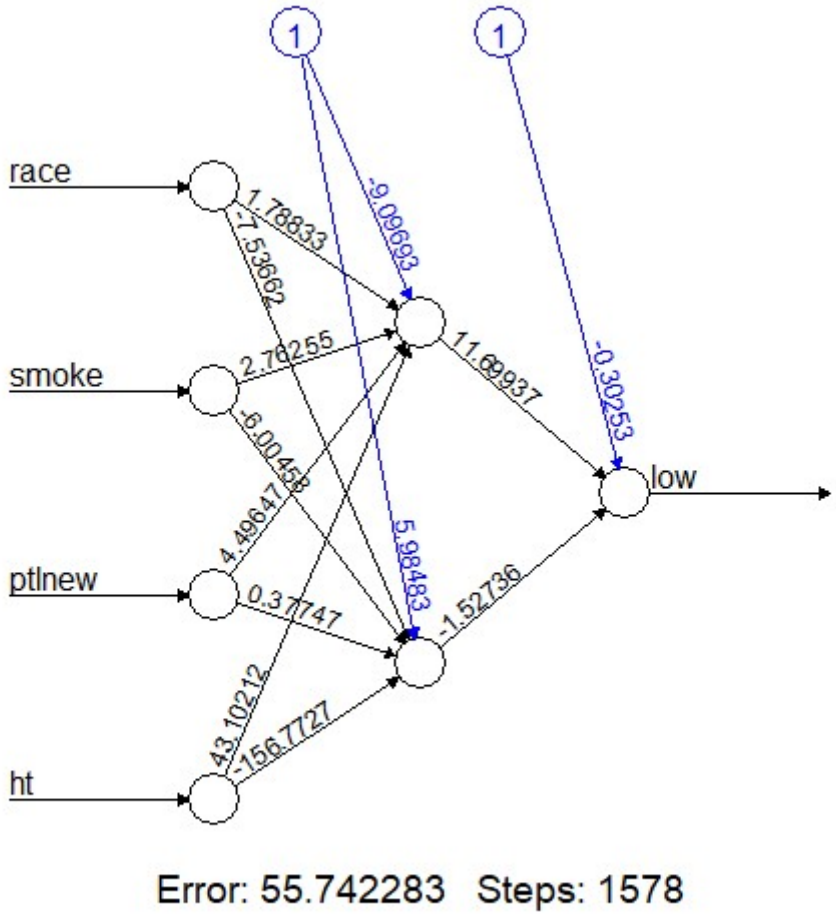


Figure 1. Articial Neural Network for Predicting Low Birthweight

It may be questionable to assume that the relationship between, say the onset of a disease and the length of exposure to a risk factor has anything to do with the multi-nodal structure of the human brain. But such a flexible network structure has become a standard template in many data-centric supervised models of scientific or medical responses.

The stability and convergence of such models is difficult to assess in general (Alain and Bengio, 2017). Many issues remain open to study as the tools required to assess highly nonlinear recursive functions are limited. For example, simple nonlinear recursive functions underlie and generate the complexity of mathematical chaos based models (Bassingthwaighte, Liebovitch, West, 1994). Machine learning nonlinear recursive environments are levels of complexity beyond this, though once the aggregation of information begins, in some settings there is often a low dimensional projection of information to be expected (Udell and Townsend, 2019). In some data settings linear and generalized linear statistical models and associated ANOVA decomposition may provide useful baselines of pattern recognition from which to interpret the results given by machine learning models in specific research settings.

6. Searching for Information

For researchers wrestling with understanding information and data patterns in large databases, there are now several paradigms and approaches and tools for data analysis. The various perspectives and procedures mentioned above can be used to help understand the same database and set of variables. Each provides a slightly different perspective on the

information in the data and how it is to be interpreted. These then must also be interpreted with regard to the medicine or science in question.

ANN models are typically focused on achieving a high level of predictive classification accuracy, with limited attribution of the accuracy or level of variation to the specific variables or features in the dataset (Kernbach and Staartjes, 2022). Statistical methods, developed from a more linear perspective, often use the ANOVA framework to specifically examine the amount of variation in the data explained by specific variables and related parameters in the model (MacKinnon, Luecken, 2011).

The developmental process for a modern quantitative model may now include: initial data summary and frequentist likelihood or Bayesian statistical analysis, non-parametric or bootstrap frequentist analysis, SVD based pre-processing of data variables and dimension reduction, related unsupervised or cluster analysis, recursive partitioning and random forests for the development of classification trees, neural network and random forest classification tree summary models based for prediction models (supervised learning). More deeply algorithmic approaches are also available in the deep-learning context where adequate logical and data structures can be initialized in the model and stochastic gradient maximization applied (Schmidhuber, 2015).

7. Brief Methods Review

The standard methods of statistical analysis and predictive model building, and the perspectives they represent, are each briefly reviewed here. The summarized probability based models they support can be seen as the essence of the information that can be drawn from the processed data values.

7.1 Frequentist Linear and Generalized Linear Models

The linear model is the mainstay of most initial modeling of a dataset. If the variable measurements are continuous, then the simple equation $y = X\beta + \varepsilon$ can be fit to the data using the least squares method of estimation, where X is an n by p matrix comprised of p columns of variable measurements, β is a (p by 1) column vector of unknown coefficients to be estimated, ε is an (n by 1) vector of independent random errors and y is the (n by 1) vector of response values. The errors are also assumed to be independently distributed $\varepsilon_i \sim N(0, \sigma^2)$.

The least squares criteria is given by:

$$\min_{\beta}(y - X\beta)'(y - X\beta) \tag{1}$$

The estimated least squares coefficient values are given by $b = (X'X)^{-1}X'y$ and have a normal distribution: $N(\beta, \sigma^2(X'X)^{-1})$, which can be used to estimate or test hypotheses in regard to the various parameters making up the β vector. The least squares estimate b can be shown to be an unbiased and minimum variance estimator for β. Observed errors (residuals) $(y - Xb)$ can be plotted and used to assess assumptions. The fitted or predicted values for the estimated model are given by $\widehat{y} = Xb$ (Kutner, Nachtsheim, Neter, Li, 2005).

Generalized linear models extend the linear model format to settings where the error terms follow a wider set of possible distributions; those comprising the exponential family of distributions. These include the binomial distribution (logistic regression), Poisson, Gamma, Beta and others (Dobson and Barnett, 2008). Typically this requires a transformation of the response scale, for example use of the logit function in logistic regression. Large sample maximum likelihood-based theory underlies estimation and testing and ANOVA type decompositions can be used to estimate or test hypotheses regarding individual or subsets of the parameter vector.

The likelihood function in general is given by:

$$L(p; x) = c\prod_{i=1}^{n} f(x_i; p) \tag{2}$$

Where c is a constant, p is the parameter of interest reflecting the assumption of an independent and identically distributed probability structure.

Logistic regression models provide a nonlinear model based on the logistic curve that models the probability (p) of an occurrence of a specific outcome (yes/no) for each subject in the study:

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)} \tag{3}$$

The logistic curve structure for p is then embedded in a likelihood function based on the binomial distribution:

$$L(p; x) = cp^x(1 - p)^{n-x} \tag{4}$$

and used to generate the fitted logistic regression likelihood function. Maximum likelihood based estimates for the β_i and their related sampling properties are obtained via large sample results. Various classification errors can be assessed using the fitted model. These include ROC/AUC, sensitivity, specificity, false positives, false negatives, and overall classification accuracy. These standard approaches to developing predictive frequentist statistical models have been used for the past 80 years with their strengths and weaknesses well known (Hosmer and Lemeshow, 1989).

7.2 Bayesian Perspective

The Bayesian approach to statistical inference uses the same likelihood-oriented model-data combination used in frequentist statistical methods to define a statistical model and process information (Congden, 2003). The observed data however is formally conditioned upon and probability transferred to the parameter space via use of Bayes theorem. Selection of a prior density, $p(\beta)$, reflecting existing beliefs in regard to possible values of the parameter β is necessary. The resulting posterior density is given by:

$$p(\beta|data) = k \cdot p(\beta) \cdot L(\beta|data) \tag{5}$$

where k is a constant of integration, and $L(\beta|data)$ the likelihood function. Numerical integration to obtain marginal distributions for each element of β provides estimation of individual parameters in the sense of updating prior information for each of them. Bayes factors and posterior odds ratios can be used to assess more global hypotheses and overall model fit, but here accuracy is measured simply as the length of credible intervals for model parameters of interest and measures of standard error. Non-informative priors for individual model parameters in the form of highly dispersed independent normal densities, standard for logistic regression, are used here.

7.3 Artificial Neural Networks

A common machine learning approach to relating a response y and a data matrix X is the ANN model. These are methods of classification that are essentially nonlinear recursive network functions defined loosely in a manner replicating a simplified nodal model of the human brain (Higham and Higham, 2019). In regards to supervised modeling, these use a multiple hidden layer model structure, applying back-propagation methods and multiple replications to fit the model. The sigmoidal function serves as an activation function in the hidden and output layers, rescaling outputs to $(0, 1)$. The back-propagation algorithm applies various learning rates and momentum values. The learning process is typically discontinued when the average error (mean square error) in the training set decreases to 0.00001.

In terms of overall model structure and key aspects (see Figure 1), (i) the response and explanatory variables are the input, (ii) the number of hidden layers a key determinant of the complexity of the model and related number of fitting parameters, and (iii) the output is a classification value in $(0, 1)$ and its interpretation. These components provide the overall information context for the neural network approach.

Within each layer of nodes, the model is fit using the stochastic gradient algorithm and back-propagation and the resulting weights and bias values re-weighted onto a $(0, 1)$ scale via use of the logistic or sigmoid function. This is then used as input to the next hidden layer level and the fitting process repeated giving an updated sigmoidal function as output. The convolution process here gives a highly nonlinear overall fitting process and the number of hidden layers reflect the degree of nonlinearity on a general scale.

The predictive accuracy of the algorithm and model may be assessed using $50/50$ randomized training versus testing/validation resamples with the training sample used to build a model with which to predict the testing sample response. As the output for each subject is a yes/no classification value, similar to logistic regression, ROC, AUC and other measures of classification accuracy and error can be generated and compared to frequentist and other methods.

Given the sample size of the examples considered here, and the stability of the data, two hidden layers are used as a basic reference in the ANN throughout. As with all ANNs, the use of high numbers of hidden layers raises the issue of serious over-fitting and the model connecting the response to any pattern, even if essentially random, in the data. The real-world examples chosen display patterns that are only mildly significant, challenging to model with moderate goodness-of-fit and significant levels of variation.

8. Training Artificial Neural Networks

The multi-nodal model used by ANNs is iteratively fit to the observed training data. Once the model has been developed (trained) it can be fit to the testing component of the data. The fit to the training data uses a cost minimization format to obtain a sequence of fitted models that hopefully converge due to the use of least squares type criteria, backward propagation fitting methods and related stochastic gradient algorithms (Higham and Higham, 2019).

The ANN model classifies a binary output y_i in relation to a set of explanatory variables $X = [x_1, ..., x_n]$. As with logistic regression models, the logistic or sigmoidal function is used to link continuous or discrete input to an output lying between 0 and 1 at each nodal value. This can be written:

$$\eta(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

for $x > 0$. The structure of the ANN model sets up layers of neurons, each of which takes on values based on the $\eta(x)$ function. This set of $\eta_i(x)$ values can be seen as generating a vector of outcomes of nodal values for each hidden layer, h_j, $i = 1, .., m$, as the fitting algorithm is applied. Write this as $a_i' = (a_1, ..., a_n)$.

The fitting procedure is iterative, and results of each layer are passed to the next layer, after being adjusted by selected weights (W) and biases (b) at each stage. The number of these parameters at each stage far exceeds the number of explanatory variables at a given stage (n) as they include all parameters previously calculated at each hidden layer node. These are updated as the fitting procedure proceeds.

This can be expressed as:

$$\eta(Wa + b) \tag{7}$$

and the i^{th} nodal value a_i for the next layer is given by:

$$= \eta(\sum_j w_{ij} a_j + b_i) \tag{8}$$

These iterative calculations, for example in the third hidden layer of the *ANN* model, are of the form:

$$\eta(W_3[\eta(W_2 a + b_2)] + b_3) \tag{9}$$

To fit the model to the observed data, $y(x_i)$, a cost function can be defined as a measure of distance to be minimized between fitted model and observed outcome:

$$Cost(W_1, W_2, W_3, b_1, b_2, b_3) = \sum_i \|y(x_i) - \eta(W_3[\eta(W_2 a + b_2)] + b_3)\|^2 \tag{10}$$

This is a function of the parameters with the data $y(x)$ known. The training of the network is the selecting of parameter values to minimize the $Cost(\cdot)$ function. As the fitting or "training" procedure increases the number of hidden layers, earlier parameters are modified via back propagation with convergence attained by applying stochastic gradient descent methods (Golas et al., 2018).

9. Datasets

Several standard datasets drawn from the medical literature were used here to generate comparative assessments. Note that all reported datasets should be viewed carefully and have a risk of bias. If the study design was not carried out as per study protocol or the sampling was not representative of the population of interest, bias may result. If there is bias in a dataset it cannot be adjusted for after the fact. Bias is best dealt with, if possible, by the application of randomized design principles when carrying out the study (Hosmer and Lemeshow, 1989).

Here, frequentist, Bayesian and machine learning approaches are applied to a common set of examples, and all three will similarly affected by the presence of bias in the data. The focus here is on the comparative aspect of the three approaches in the context of the actual datasets.

9.1 Risk Factors for Low Birth Weight

This data was collected at the Baystate Medical Center, Springfield, Massachusetts during 1986 to examine factors contributing to increased risk of low birth weight infants (yes/no). Data was recorded for 189 women of whom 59 had low birth weight infants. Initial variables included: low: an indicator (y/n) of low birth weight (¡ 2.5kg), age: mother's age in years at the time of birth, lwt: mother's weight (in pounds) at the time of the last menstrual period, race (white y/n), smoke: indicator of mothers smoking status (y/n), ptlnew: previous premature labors (y/n), ht: indicator of hypertension (y/n), and ui: indicator of uterine irritability (y/n). Data was obtained from Andrews D.F. and Herzberg A.M. (1985). See Figure 1 for ANN representation of the best fitting set of variables.

9.2 The Process of Calcium Oxalate Crystals Formation

The data set is based on 79 specimens analyzed to assess if certain physical characteristics were related to formation of calcium oxalate crystals, the response of interest. Variables included; r: an indicator of the presence of calcium oxalate crystals (yes/no), gravity, ph, osmolarity (proportional to the concentration of molecules in solution), conductivity (proportional to the concentration of charged ions in solution), urea concentration in millimoles per litre, and calcium concentration. Data was obtained from Andrews D.F. and Herzberg A.M. (1985).

9.3 Craniometry Measurements

A set of related cranial measurements drawn from a set of human skull measurements denoted here as x_1, x_2, x_3, x_4 was assessed on each of $n = 350$ individuals. The response of interest is whether they belong to a specific lineage group (yes/no). Data was based on the cranial dataset in Howells W.W. (1996).

9.4 Simulated Neurological Treatment Dataset

A clinical trial examining the usefulness of an intervention to relieve neurologic related pain using osteopathic and natural remedies was conducted on a sample of $n = 350$ patients. Pain reduction was the response of interest assessed as a yes/no variable. Two clinical measurements and a level of biomarker expression were assessed for each subject and used to predict pain reduction (yes/no).

All datasets are available upon request.

10. Results

For each dataset considered, the comparison here compares probability-based frequentist models with (i) Bayesian probability models conditioned on the observed data, as a means of assessing the stability of probability-based modeling, and (ii) entirely data-centric ANN models, examining changes in accuracy of estimation in fitted models and accuracy in classification with the number of hidden layers set at 2 and 12, using training and testing data components. Measures of sample variation and related accuracy of estimation for frequentist and Bayesian models are measured by confidence and credible interval lengths. A Type I error of 0.05 was used for frequentist singificance.

Frequentist logistic regression is then directly compared to ANN models by comparing accuracy of classification by examining sensitivity, specificity, ROC curve and AUC diagnostic measures of goodness-of-fit and accuracy. For the ANNs, this is repeated ten times to account for data variation in the randomly generated training and testing components, with median values reported.

Table 1. Frequentist and Bayesian Fitted Models

(Variables significant in the frequentist model shown, Bayes prior density $N(0, 10000)$)

	Frequentist			**Bayes**		
	OR	**SE**	**95% Conf.Int.(length)**	**OR**	**SE**	**95% HPD Int.(length)**
Low Birthweight						
Race	1.7	0.35	[1.15,2.57] (1.42)	1.8	0.39	[1.11,2.57] (1.46)
Smoking	2.7	1.02	[1.28,5.67] (4.39)	3.0	1.22	[1.25,5.68] (4.43)
Ptlnew	2.2	0.75	[1.13,4.31] (3.18)	2.4	0.89	[0.96,4.15] (3.19)
Ht	3.6	2.27	[1.05,12.36] (11.31)	4.6	3.53	[0.66,11.23] (10.57)
Crystal Formation						
Osml	1.0	0.12	[1.01,1.06] (0.05)	1.0	0.01	[1.02, 1.07] (0.05)
Cond	0.5	0.10	[0.35,0.75] (0.40)	0.5	0.10	[0.28,0.67] (0.39)
Urea	1.0	0.01	[0.93,0.98] (0.05)	1.0	0.01	[0.93,0.98] (0.05)
Calcium	2.2	0.48	[1.39,3.35] (1.96)	2.4	0.62	[1.44,3.62] (2.18)
Craniometry						
x_1	0.9	0.04	[0.79,0.95] (0.16)	0.87	0.04	[0.79,0.95] (0.16)
x_3	1.1	0.05	[1.03,1.23] (0.20)	1.13	0.05	[1.04,1.23] (0.19)
Neurology						
Clin1	1.1	0.02	[1.01, 1.09] (0.08)	1.1	0.02	[1.01, 1.10] (0.09)
Clin2	1.6	0.31	[1.09, 2.35] (1.26)	1.7	0.36	[1.11, 2.45] (1.34)
Biomarker	1.3	0.10	[1.11, 1.49] (0.38)	1.3	0.10	[1.13, 1.52] (0.39)

(11)

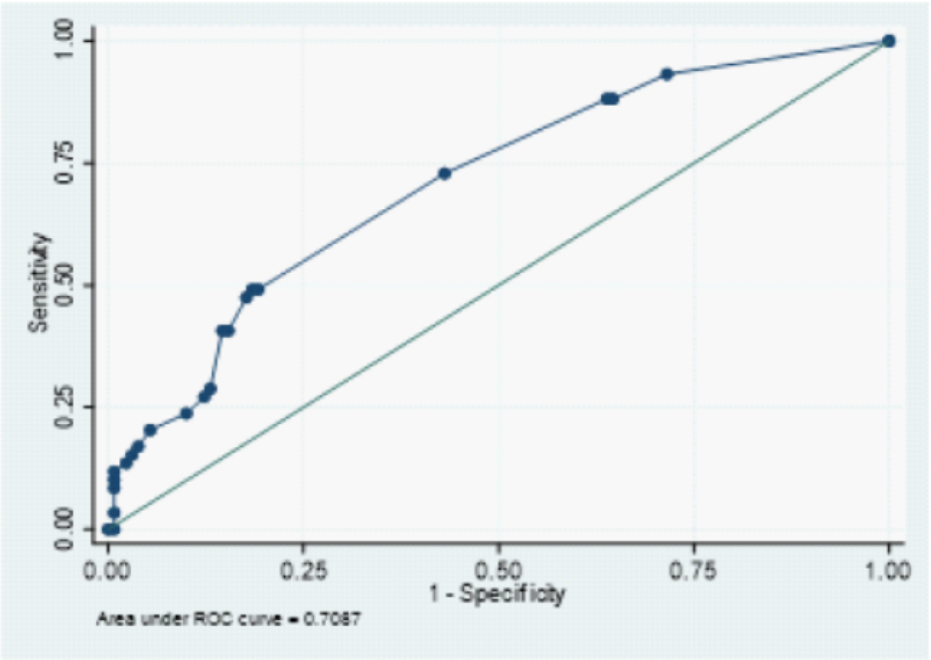
The frequentist and Bayesian measures of accuracy here are very similar across the various datasets. This is to be expected given the choice of a non-informative prior. See Table 1. The conditional aspect of the Bayesian analysis does not have a great impact here and the small number of parameters limits possible shrinkage effects due to Monte Carlo based numerical integration. These support a fairly stable and moderately predictive statistical analysis.

Table 2. Logistic regression versus Neural Network (hidden layers = 2, 12)

Variables	Sensitivity	Specificity	False +	False -	Overall True	AUC
Logistic Regression (p = .5)						
Low Birthweight	23.7%	90.0%	10.0%	76.3%	69.3%	70.9%
Crystals	69.7%	86.4%	13.6%	30.3%	79.2%	86.6%
Craniometry	20.0%	97.5%	2.5%	80.0%	82.0%	73.7%
Neurology	14.3%	98.9%	1.2%	85.7%	92.6%	86.7%
ANN (h = 2, p = .5)						
Low Birthweight	44.0%	84.9%	15.3%	56.0%	72.6%	65.1%
Crystals	63.2%	86.4%	13.6%	36.8%	70.7%	76.8%
Craniometry	27.8%	92.0%	8.0%	72.2%	75.0%	67.2%
Neurology	33.3%	93.6%	6.4%	66.6%	89.3%	72.4%
ANN (h = 12, p = .5)						
Low Birthweight	52.0%	66.1%	33.9%	48.0%	61.9%	56.1%
Crystals	80.0%	44.0%	56.0%	20.0%	60.0%	67.4%
Craniometry	22.2%	92.0%	8.0%	77.8%	73.5%	63.8%
Neurology	40.0%	97.3%	2.8%	60.0%	90.5%	70.5%

(12)

On average, overall classification accuracy and AUC values for the ANN based classification results are similar or less accurate than the standard frequentist logistic regression result across the various datasets, using a p = 0.5 threshold cutoff. The sensitivity of the ANN model is slightly higher across all datasets, reflecting perhaps the data-centric nature of the model, but often with lower specificity. The effect of increasing hidden layers and thus ANN model complexity is negligible compared to variation in the complex ANN models across the various datasets. The higher number of hidden layers gives mostly lower overall classification accuracy and AUC values. See Table 2.



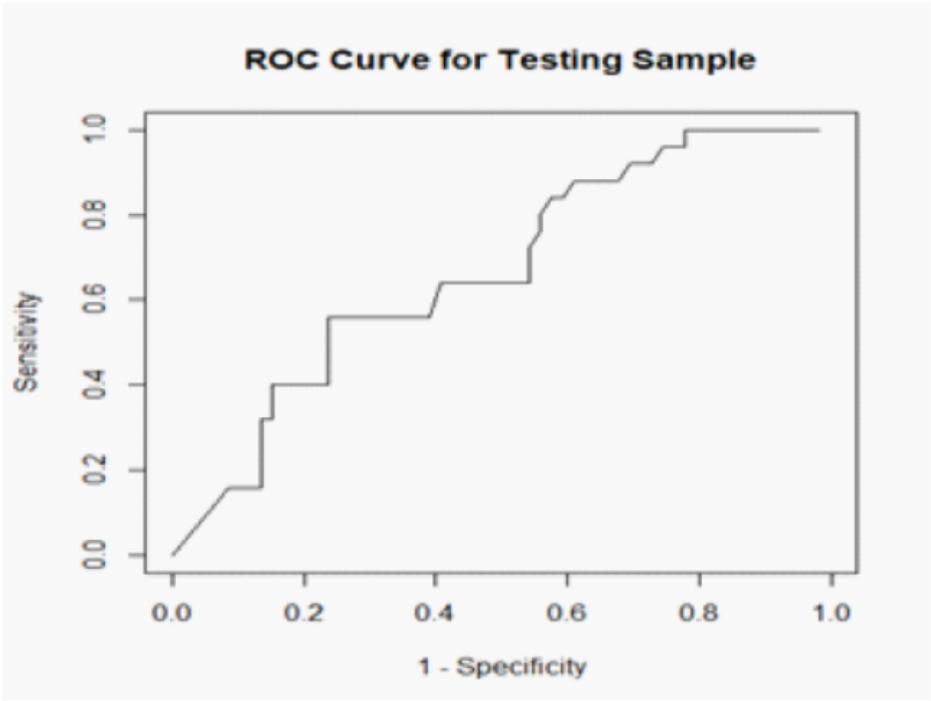


Figure 2a. Low Birthweight Example

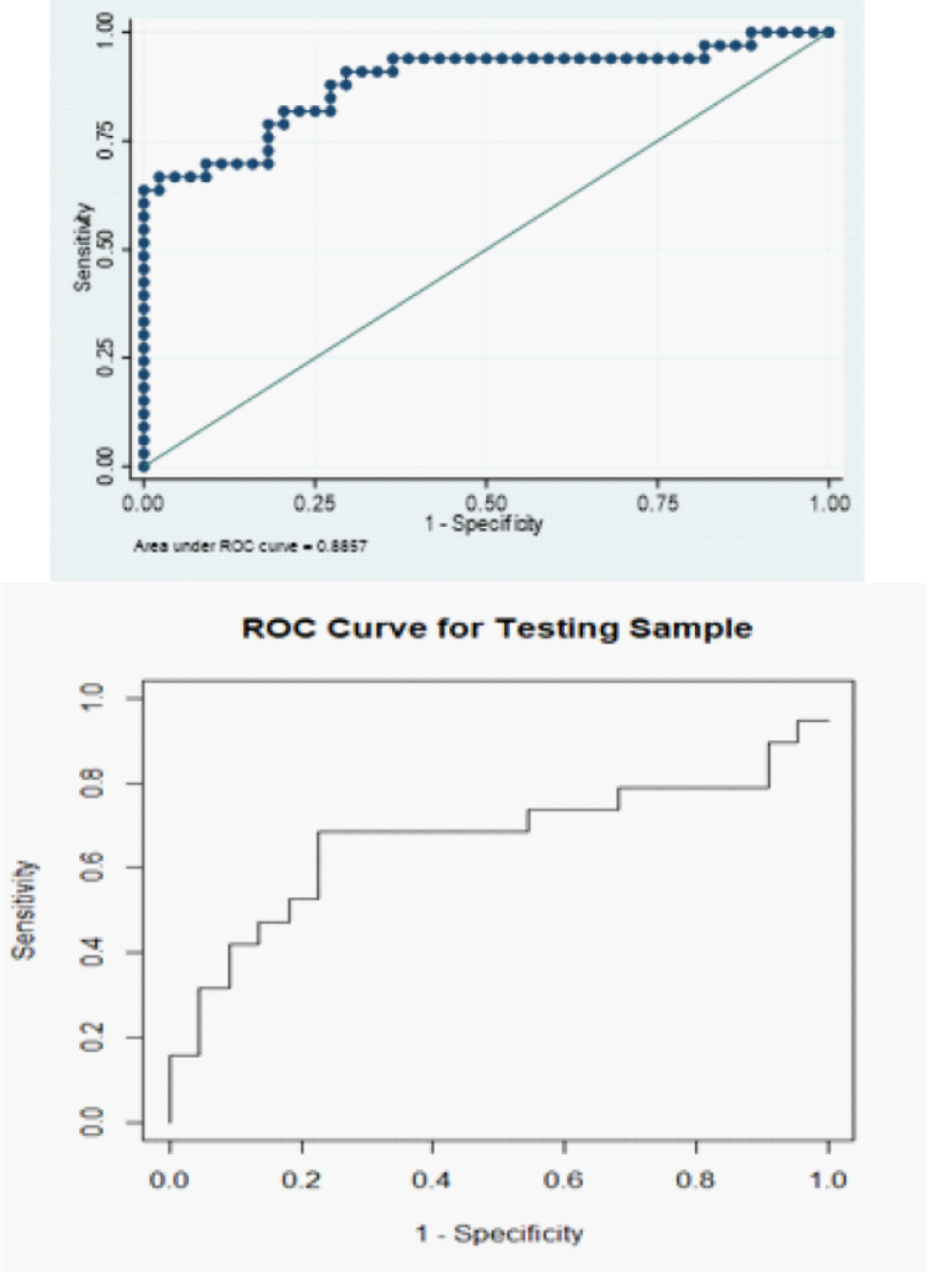
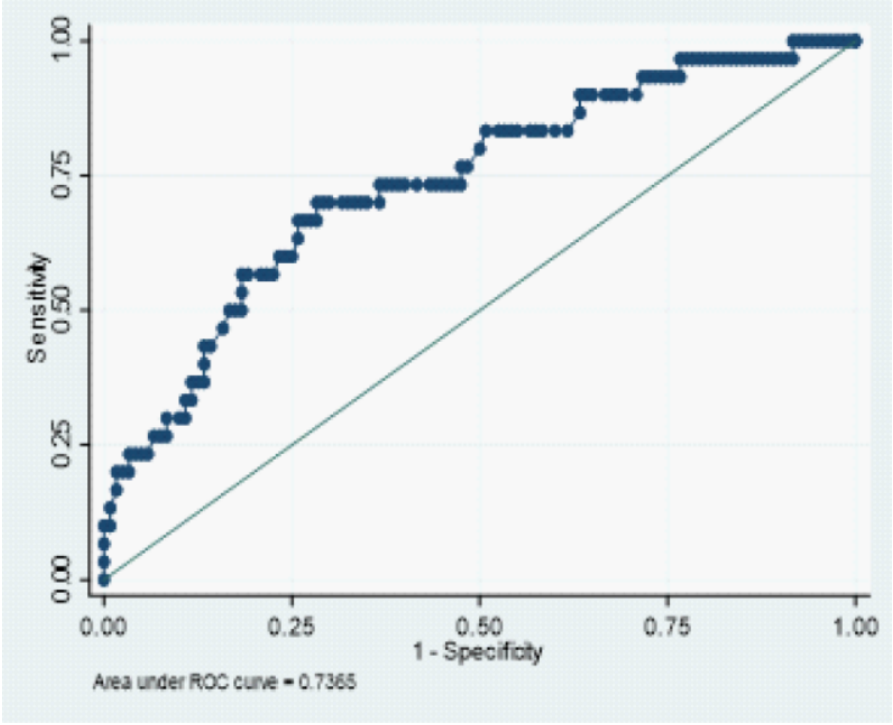


Figure 2b. Crystals Example



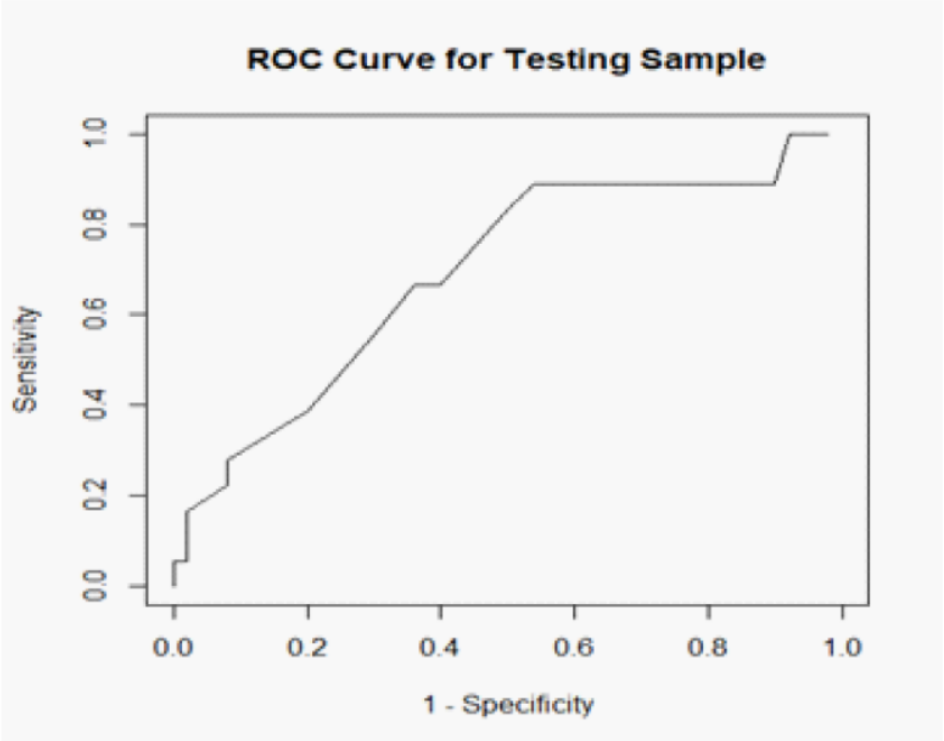
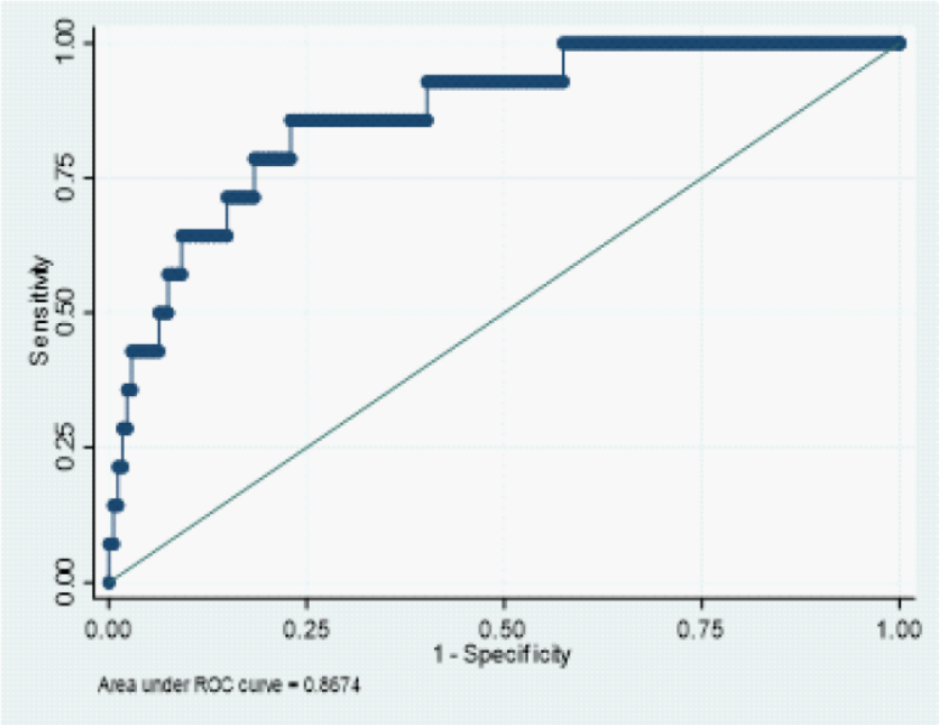


Figure 2c. Skulls Example



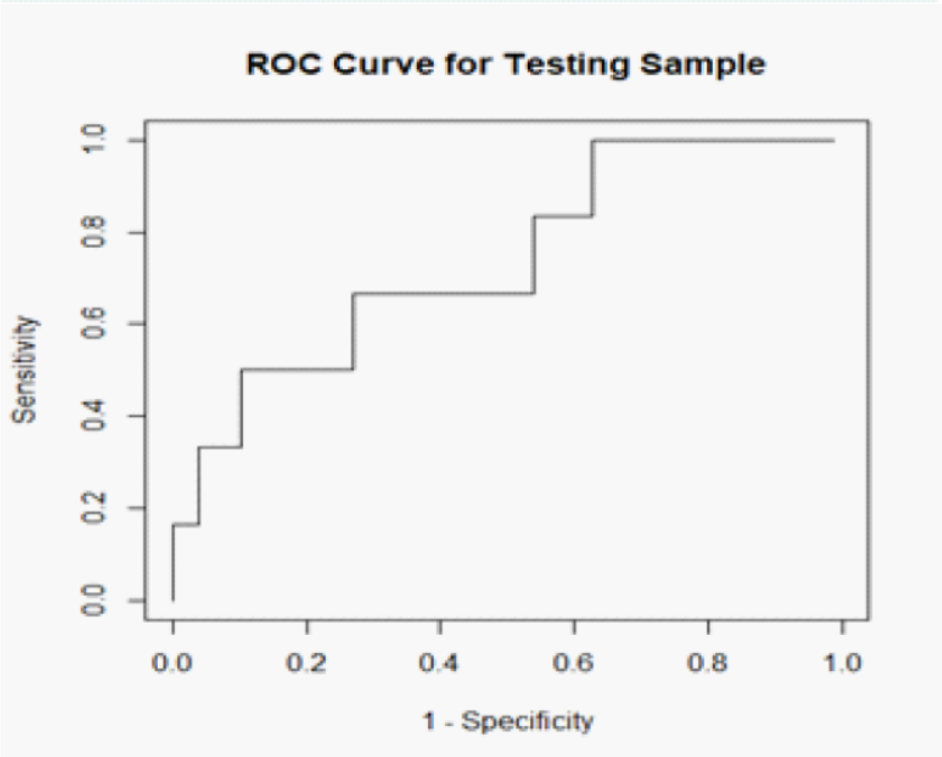


Figure 2d: Neurology Example

Figures 2. ROC Curves for Logistic Regression and ANN (Hidden layers = 2)

The ROC curves (see Figure 2) and related AUC values assessing overall model fit, show limited predictive accuracy on the part of the ANN model (h = 2). This was not improved upon by using higher levels of complexity in the ANN (h = 12) model.

11. Discussion

There is a serious discussion ongoing in many areas of science and medicine regarding the use of mathematical and probability models versus data-centric machine learning, BigData and AI models. But there are few guidelines on how to compare such modeling approaches, where such comparisons are appropriate. Probability based methods, frequentist and Bayesian, represent different probability based perspectives on the model and data being studied. Machine learning models are not probability based and do not generalize to the population unless the dataset itself is representative or very large, representing most of the population.

Advocates of machine learning methods in large samples often cite improved fit as a basis for selecting ML methods over standard statistical methods, often due to the high number of fitted parameters, but the improved goodness-of-fit in some settings may be simply be a function of the greater data-centric focus of these approaches and the selection bias this perspective imposes on probability based methods. In smaller datsets this may not be the case. The highly flexible nature of neural networks may require very large sample sizes to converge with high degrees of accuracy (Colbrook, Antunb, Hansen, 2022; Schmidhuber, 2015) and be stable.

Statistical approaches employ population based probability models to generalize the sample to the population. If they are mostly linear, an ANOVA type breakdown often gives the relative importance of each variable. ANN and machine learning models in general are highly nonlinear, deliberately flexible to model more of the randomness in the observed data. But these models only assess whether the observed pattern in one portion of the data (training) is similar to the pattern in the rest of the data (testing). This is more a measure of agreement than prediction. Statistical models are justified by repeated sampling and modeling what will happen on average in a population-based setting. These comparisons reflect two different contexts in which the data is being analyzed, and comparisons of the methods should reflect this difference.

Very large datasets and highly complex responses and related samples are potentially the domain of neural networks, if a general theory of convergence can be obtained. Smaller studies, where carefully designed and randomized experiments are conducted with well-defined basic responses, may remain an area of useful application of probability-based statistical models. As noted in the moderate sample size examples considered here, population models reflect average concepts, with fitted parameters often having long-term probability interpretation and stability.

As well, randomness enters into these modeling approaches somewhat differently, leading to very different definitions of probability being applied. This should focus comparisons on very basic measures of variation and accuracy. Logistic regression models reported should include population based results.

Both ANN and logistic regression models yield, as an output for each subject, a value in $(0, 1)$. These give a predicted classification value of Yes or a No according to whether the value is greater than a chosen cutoff, typically $p = 0.5$. In logistic regression applications where a population based perspective is taken, if the response of interest is rare, the cutoff may be lowered to 0.3 or 0.2. This sometimes gives better model sensitivity and specificity. If comparisons are to be made between a logistic regression model and an ANN model, the cutoff should be set at the same level.

Comparisons of data-centric versus statistical probability should reflect an awareness of the differing study design contexts underlying the application of the methods. Interpretation of these approaches in the context of medical research should more carefully consider these issues. Probability based adjustments such as weighting and random effects interpretations are standard elements of frequentist assessment and interpretation of large scale survey data. These may need to be re-developed or incorporated directly into the context of data-centric machine learning methods in order to appropriately interpret the dataset itself and related data-centric inferences (Brimacombe, 2024).

Note that all the methods considered here are subject to potential bias in the data when the data is not collected in an unbiased and representative manner. There remains an important place for the further development of quality randomized methods of data collection for machine learning applications.

In many settings, a trained, supervised ANN model may provide highly accurate predictive classification. See for example (Hogan, Brimacombe, Mosha, Flores, 2021). Note the data-centric nature of both the approach and evaluation (the particular training/testing split of the data). This is not really prediction, but more assessment of agreement between different portions of the observed data.

More hidden layers in the neural network model and related free parameters make it easier to fit the neural network model to the data, but they also make it easier to fit any response to the explanatory data elements (Petzschner, 2024). This may even give a fitted model that fits the overall data well, but yields poor predictions. In the datasets considered here, using greater numbers of hidden layers did not improve the basic level of fit as measured by the set of classification diagnostics (Checkroud et al., 2024). The random splitting of the data gave respective sample sizes that were relatively small. This effect is typically overcome in larger data sets with large numbers of hidden layers, but a general theory for convergence is not currently available (Colbrook, Antunb, Hansen, 2022).

In the past, the simplicity of ANOVA and standard least squares based analysis was supported with many citing Occam's Razor and other principles favoring simplicity and interpretability. This template, for many years, has been the approach supporting advances in science, medicine and most research fields.

The newer modeling approach underlying machine learning embraces more complex models, emphasizing predictive accuracy over understandability. The development of an integrated set of interpretable, data-based approaches to classification and supervised modeling would be helpful to more easily comprehend ANN and other machine-learning based results and understand data patterns in the presence of random selection and sampling-based variation.

References

Alain, G., & Bengio, Y. (2017). Understanding intermediate layers using linear classifier probes. *International Conference on Learning Representations.* Andrews, D. F., & Herzberg, A. M. (1985). *Data: A collection of problems from many fields for the student and research worker.* Springer-Verlag.

Ansari, R. M., & Baker, P. (2021). Identifying the predictors of COVID-19 infection outcomes and development of prediction models. *Journal of Infection and Public Health, 14,* 751-756. https://doi.org/10.1016/j.jiph.2021.03.007

Bassingthwaighte, J. B., Liebovitch, L. S., & West, B. J. (1994). *Fractal physiology.* Oxford University Press.

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American Journal of Epidemiology, 188*(12), 2222-2239. https://doi.org/10.1093/aje/kwz189

Bisaso, K. R., Karungi, S. A., Kiragga, A., Mukonzo, J. K., & Castelnuovo, B. (2018). A comparative study of logistic regression-based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. *BMC Medical Informatics and Decision Making, 18*, 77. https://doi.org/10.1186/s12911-018-0654-2

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Addison Wesley Publishing Co.

Brimacombe, M. (2019). *Likelihood methods in biology and ecology: A modern approach to statistics.* CRC Press.

Brimacombe, M. (2024). Data flow-based strategies to improve the interpretation and understanding of machine learning models. *Bioengineering, 11*(12), 1189. https://doi.org/10.3390/bioengineering11121189

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury, CA.

Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., ... Paulus, M. (2024). Illusory generalizability of clinical prediction models. *Science, 383*(6679), 164-167. https://doi.org/10.1126/science.adg8538

Ching, T., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface, 15*(141), 20170387. https://doi.org/10.1098/rsif.2017.0387

Colbrook, M. J., Antun, V., & Hansen, A. C. (2022). The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smales 18th problem. *Proceedings of the National Academy of Sciences, 119*(12), e2107151119. https://doi.org/10.1073/pnas.2107151119

Congdon, P. (2003). *Applied Bayesian modelling*. John Wiley and Sons.

Dennise, D., Dalma-Weiszhausz, D. D., Warrington, J., Tanimoto, E. Y., & Miyada, C. G. (2006). The affymetrix GeneChip platform: An overview. *Methods in Enzymology, 410,* 3-28. https://doi.org/10.1016/S0076-6879(06)10001-4

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models* (3rd ed.). Chapman & Hall/CRC Texts in Statistical Science.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745-766. https://doi.org/10.1080/10618600.2017.1384734

Efron, B. (2024). Machine learning and the JamesCStein estimator. *Japanese Journal of Statistics and Data Science, 7*, 257-266. https://doi.org/10.1007/s42081-023-00209-y

Efron, B., & Hastie, T. (2021). *Computer age statistical inference: Algorithms, evidence, and data science.* Cambridge University Press.

Elovainio, M., Hakulinen, C., Pulkki-R?back, L., Aalto, A., Virtanen, M., Partonen, T., & Suvisaari, J. (2020). General Health Questionnaire (GHQ-12), Beck Depression Inventory (BDI-6), and Mental Health Index (MHI-5): Psychometric and predictive properties in a Finnish population-based sample. *Journal of Psychiatric Research, 126,* 112973. https://doi.org/10.1016/j.jpsychires.2020.112973

Golas, S. B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., ... Jethwani, K. (2018). A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: A retrospective analysis of electronic medical records data. *BMC Medical Informatics and Decision Making, 18*(1), 44. https://doi.org/10.1186/s12911-018-0620-z

Haleem, A., Javaid, M., & Khan, I. H. (2019). Current status and applications of artificial intelligence (AI) in the medical field: An overview. *Current Medicine Research and Practice, 9,* 231-237. https://doi.org/10.1016/j.cmrp.2019.08.001

Hardt, M., Recht, B., & Singer, Y. (2016). *Train faster, generalize better: Stability of stochastic gradient descent.* In Proceedings of the 33rd International Conference on Machine Learning (pp. 1225-1234).

Higham, C. F., & Higham, D. J. (2019). Deep learning: An introduction for applied mathematicians. *SIAM Review, 61*(4), 860-891. https://doi.org/10.1137/18M1165748

Hogan, A. H., Brimacombe, M., Mosha, M., & Flores, G. (2021). Comparing artificial intelligence and traditional methods to identify factors associated with pediatric asthma readmission. *Academic Pediatrics. Advance online publication*. https://doi.org/10.1016/j.acap.2021.07.015

Hosmer, D. W., & Lemeshow, S. (1989). Applied logistic regression. John Wiley.

Howells, W. W. (1996). Howells craniometric data on the internet. *American Journal of Physical Anthropology, 101,* 441-442. https://doi.org/10.1002/(SICI)1096-8644(199612)101:4¡441::AID-AJPA1¿3.0.CO;2-2

Kernbach, J. M., & Staartjes, V. E. (2022). Foundations of machine learning-based clinical prediction modeling: Part IIłGeneralization and overfitting. In V. E. Staartjes, L. Regli, & C. Serra (Eds.), *Machine learning in clinical neuroscience* (pp. xx-xx). Acta Neurochirurgica Supplement. Springer. https://doi.org/10.1007/978-3-030-85292-4

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023). A review of ARIMA vs. machine learning approaches for time series forecasting in data-driven networks. *Future Internet, 15*(8), 255. https://doi.org/10.3390/fi15080255

Krizhevsky, A., Sutskever, I.,& Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 1097-1105). Curran Associates, Inc.

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics, 6,* 10. https://doi.org/10.1186/1758-2946-6-10

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill Irwin.

MacKinnon, D. P., & Luecken, L. J. (2011). Statistical analysis for identifying mediating variables in public health dentistry interventions. *Journal of Public Health Dentistry, 71*(Suppl. 1), S37-S46. https://doi.org/10.1111/j.1752-7325.2011.00252.x

Morgan, D. J., Bame, B., Zimand, P., et al. (2019). Assessment of machine learning vs standard prediction rules for predicting hospital readmissions. *JAMA Network Open, 2*(3), e190348. https://doi.org/10.1001/jamanetworkopen.2019.0348

Nielsen, M. (2015). Neural networks and deep learning. Determination Press.

Petzschner, F. H. (2024). Practical challenges for precision medicine: The prediction of individual treatment responses with machine learning faces hurdles. *Science, 383*(6679), 149-150. https://doi.org/10.1126/science.adm9218

Pitman, E. (1936). Sufficient statistics and intrinsic accuracy. *Mathematical Proceedings of the Cambridge Philosophical Society, 32*(4), 567-579. https://doi.org/10.1017/S0305004100019307

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61,* 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

Sidey-Gibbons, J., & Sidey-Gibbons, C. (2019). Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology, 19*, 64. https://doi.org/10.1186/s12874-019-0681-4

Spreafico, M., Hazewinkel, A.-D., & Van de Sande, M. A. J. (2024). Machine learning versus Cox models for predicting overall survival in patients with osteosarcoma: A retrospective analysis of the EURAMOS-1 clinical trial data. *Cancers, 16*(16), 2880. https://doi.org/10.3390/cancers16162880

Udell, M., & Townsend, A. (2019). Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science, 1*(1), 144-160. https://doi.org/10.1137/18M1183480

Zhang, J., Morley, J., Gallifant, J., Oddy, C., Teo, J. T., Ashrafian, H., Delaney, B., & Darzi, A. (2023). Mapping and evaluating national data flows: Transparency, privacy, and guiding infrastructural transformation. *The Lancet Digital Health, 5*(10), e737-e748. https://doi.org/10.1016/S2589-7500(23)00157-7

Acknowledgments

Not applicable.

Authors contributions

The author is responsible for all work.

Funding

No grant funding was used to supported the work.

Competing interests

The author has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Canadian Center of Science and Education. The journals policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.