Data challenges and practical aspects of machine learning-based statistical methods for the analyses of poultry data to improve food safety and production efficiency

Maurice Pitesky[1,2]*, Joseph Gendreau[1,2], Tristan Bond[1,2], and Roberto Carrasco-Medanic[1,2]

Address: [1]Department of Population Health and Reproduction, School of Veterinary Medicine, UC Davis, Cooperative Extension, Davis, CA, 95616, USA.
[2]AgriNerds, Davis, CA, 95616, USA.

ORCID information: Maurice Pitesky (orcid: 0000-0003-0084-6404)

***Correspondence:** Maurice Pitesky. Email: mepitesky@ucdavis.edu

Received: 22 May 2020
Accepted: 18 August 2020

doi: 10.1079/PAVSNNR202015049

The electronic version of this article is the definitive one. It is located here: http://www.cabi.org/cabreviews

Abstract

Leveraging data collected by commercial poultry requires a deep understanding of the data that are collected. Machine learning (ML)-based techniques are capable of "learning by finding" nonobvious associations and patterns in the data in order to create more reliable, accurate, explanatory, and predictive statistical models. This article provides practical definitions and examples of ML-based statistical approaches for the analysis of poultry production and poultry food safety-based data. In addition to summarizing the literature, two real examples of the supervised machine learning ensemble technique, random forest (RF), are provided with respect to predicting egg weights from a commercial layer farm and identifying the potential causes of a *Salmonella* outbreak from a commercial broiler facility. Specifically, as an example, for the prediction of egg weights, a training model and a test model were created, and a modification of RF was used to explore the ability to predict egg weights. Results identified multiple variables including Age, Farm Location, Body Weight, Total Eggs, Hens Housed, and House Style which were predictive of the continuous variable Egg Weight. With respect to the accuracy of the variable Egg Weight, the average error between the predicted and actual egg weight was determined to be less than 3%. With respect to broiler food safety, a relational database was constructed and a supervised RF model was developed to identify the predictors of *Salmonella* in a grow-out farm and associated broiler processing plant. Predictors of *Salmonella* that included livability, density of birds in the grow-out farm, and breeder age were identified. The task of choosing the most appropriate ML-based model(s) that accounts for the large number of variables common to the poultry industry and addresses the intricate interdependence between several production parameters and inputs while predicting multiple sequential outputs is complex. The use of ML techniques in combination with new data streams including sensors (e.g., visual and audio), IoT, and Web-scraping could offer a more comprehensive, efficient, and timely approach toward evaluating productivity, food safety, and profitability in commercial poultry.

Keywords: machine learning, artificial intelligence, data mining, knowledge discovery in databases, exploratory data analysis, supervised and unsupervised models, food safety, production efficiency, predictions

Review Methodology: The present article constitutes a detailed literature review using the online subscription-based citation indexing database, Google Scholar (keyword search terms using a Boolean search with only the "AND" modifier) terms included: poultry, agriculture, machine learning, artificial intelligence, sensors, random forest, boosting, support vector machine, data mining, and meat. References from the articles obtained by this method were evaluated for additional relevant material. The authors also checked for any upcoming studies that are not yet published and published books. In addition, the original data from several commercial broiler and layer companies were used for the ML-based technique RF.

Methods for RF analysis of broiler data: Live production and processing data were provided by a commercial broiler company over a 3-year time frame from 2013 to 2016. The data were collected from 138 premises (i.e., breeding facilities, hatcheries, grow-out farms, and processing plants) from one company. The data set contained 124 different variables with over 30,000 observations. Data analysis was conducted using the following packages from R [1]: party [2], randomForest [3], and pROC [4]. In order to perform these analyses, live production and processing data from an integrated commercial broiler company were collected. A complete and exhaustive exploratory data analysis (EDA) was performed using graphic techniques, and recursive partitioning algorithms to detect and understand patterns and select the predictors for *Salmonella* presence in live production (breeding facilities, hatcheries, and the grow-out facilities) and the processing plant.

Method for RF analysis of egg weights: Live production and processing data were provided by a commercial layer company over a 7-year time frame from 2010 to 2017. The data were collected from 220 farms representing multiple integrated-layer companies. The data set contained 43 variables with over 40,000 observations. Data analysis was conducted using the randomForest package [3] from R [1]. In order to account for temporal variability, "lag data," or data from previous time steps, were utilized to predict the production variables in a time series model. In this implementation, to be truly predictive required that the RF model predicts all inputs such that the predictive input data would be available for the model to then predict the dependent variable, egg weights. The model was developed using the average egg weight, mortality, average bird weight, bird age, total eggs produced, and the feed conversion ratio from the week prior as inputs to predict the same data points in the subsequent week.

Review text

In 2017, the artificial-intelligence (AI) program AlphaGo Zero [5] developed by Google defeated an 18-time world champion Lee Sedol in the ancient Chinese board game Go. This was the first time a computer program defeated a world champion Go player [5]. What made the accomplishment most impressive is that AlphaGo Zero won the match without learning from any human moves [6]. Specifically, AlphaGo Zero played itself (over 30 million times) before playing Mr. Sedol, as opposed to the previous efforts that involved "learning" from approximately 100,000 previously played games before playing an expert [5]. This accomplishment (beating a human without relying on current and historic human knowledge and strategy) reflects the continued application of a suite of new computer-based approaches that are classified as artificial intelligence (AI) and that are transforming multiple disciplines and activities including health care, internet searches, language translation, genomics, engineering design, and agriculture [7–11]. While these examples show the potential for these suites of AI-based techniques, it is just as important to understand the limitations and the effects those limitations will have in multiple fields including poultry production if AI-based techniques are treated as a "black box" and not questioned and hence not used appropriately. One recent example was a study by Google that the authors (and media) claimed showed that their AI system could outperform human radiologists at identifying breast cancers via mammograms and made several troubling assumptions based on false premises regarding the value of breast cancer screening [12]. Specifically, it was not recognized that different cancers act differently with respect to patient outcome and hence some forms are probably better left alone (i.e., the cure is worse than the disease for some noninvasive slow-growing tumors). To that point, the Google study assumed that the identification of cancer via a mammogram was always "good," and hence, if the Google AI technique could identify more cancers, then the ML-based diagnostics were an improvement over its human diagnostic counterpart (i.e., the radiologist). Unfortunately, as we all know biology (e.g., cancer biology, poultry husbandry, and food safety) is not so simple. Specifically, with respect to breast cancer, mammograms appear to be excellent at identifying indolent forms of cancer (i.e., cancers that do not pose a health risk), and hence, some of the diagnostic benefits of mammography are questionable [13]. Combine this with the fact that doctors are inclined to aggressively treat most cancers that can thus lead to unwarranted interventions (i.e., excess treatment) and the question of whether mammograms and AI-assisted interpretation of those mammograms are actually helpful is a more complicated debate. In other words, the combination of humans and machines has the potential to improve diagnostics sensitivity but also negatively affects the outcome due to our lack of knowledge and biases. The overarching message is that "algorithms" may have significant negative implications if we do not recognize issues such as data bias, confirmation bias, and human ignorance. Applying the lessons from AlphaGo and the Google ML mammography study to agriculture will be fundamental toward the ultimate success or failure of ML-based techniques in agriculture. In short, at this point in time, the verdict is still out on how useful these types of techniques are in the multiple fundamental areas of agriculture, including poultry production and food safety. This article is intended to provide some context on supervised and unsupervised ML-based techniques and the potential of new data streams which may (or may not) improve the data analysis.

Part I: Food safety and production challenges with data

Food safety data challenges

With respect to attribution, as per the U.S. Center of Disease Control (CDC) poultry, meat is the most common

food identified in foodborne outbreaks, illnesses, and hospitalization and the second highest in food safety-related deaths [14]. While the same study identified the most common reported factors contributing to poultry-associated foodborne outbreaks were food-handling errors (64%) and inadequate cooking (53%), it is important to acknowledge that the continued efforts to reduce the poultry contamination at the farm and processing plant are necessary. In 2011, cultures of 13% of chicken samples and 6% of ground turkey samples yielded *Salmonella* sp. and 38% of chicken samples yielded *Campylobacter* sp. [15]. With respect to the value of those data and the overall efficacy of the U.S. poultry food safety-based surveillance system led by the Food Safety Inspection Service (FSIS) branch of the United States Department of Agriculture (USDA) and the Food and Drug Administration (FDA) for broilers and layers, respectively, the following limitations currently exist:

- Testing for foodborne pathogens is strictly qualitative as opposed to quantitative. Hence, the ability to use current surveillance diagnostic tests as a risk factor for food safety is poor since the load of bacteria present is important to determine from a risk perspective. While non-culture-based methods exist for quantification [16, 17] in poultry environments, no mandated integrated national system currently exists for strain identification and corresponding quantification at the retail level.
- At the processing plant, qualitative testing is not strain or serotype specific.
- At the layer farm, only one serotype of *Salmonella* is tested (*Salmonella* Enteritidis) as per the FDAs egg safety rule (21 CFR Part 118) [18].
- While there is a general consensus that testing by itself is not the solution to food safety, there is also evidence that the amount of testing of commercial poultry is epidemiologically uneven and inadequate. Specifically, at the retail level, only 22 states are currently enrolled in the FDAs National Antimicrobial Resistance Monitoring System (NARMS) program, which surveilled the retail poultry for four bacteria: *Salmonella*, *Campylobacter*, *E. coli*, and *Enterococcus*. NARMS data indicate the presence or absence of those bacteria and antimicrobial resistance (AMR) genes in several common meats sold at the retail level, including whole chickens, chicken parts, and ground turkey. In addition to providing the surveillance of AMR in retail meat, the data can also be used to provide information at the retail level on the presence/absence of enteric bacteria.
- The current USDA-FSIS performance standard requires that between 0 and 5 carcass rinsate samples/mo. must be collected from the poultry processing plants depending on the total product volume produced [19]. Testing as noted is qualitative but not quantitative. Reductions in sample size have been found, for example, to decrease the population level sensitivity for comminuted poultry-*Campylobacter* standards [20]. In summary, when considering the total amount of poultry processed and consumed per year in the U.S., a general lack of qualitative and quantitative testing exists in order to understand the baseline levels of contamination at multiple stages of the poultry supply chain [21]. However, it should be noted that the goal of this surveillance is not only to reduce the amount of contaminated product from being shipped but also to advise FSIS on how the industry is doing in order to establish appropriate baselines [17].

From a data perspective, an additional complication is the current reality that food safety data streams are highly siloed (i.e., poorly integrated). For example, NARMS and FSIS performance standard data are not integrated with each other or even available on a real-time dashboard for companies or consumers. These data are further not linked to the CDC Foodborne Disease Outbreak Surveillance System (FDOSS), where the CDC monitors and reports the incidence and the rate of foodborne illnesses [22]. This prevents companies from easily integrating these data into their company data where they could be used by the companies to improve outcomes or better understand risk. Alternatively, many companies have parallel levels of food safety surveillance which mirror or often supersede FSIS sampling guidelines which they can integrate into their data which often includes various other relevant non-processing plant data including data associated with breeding farms, hatcheries, grow-out facilities, and processing plants [14]. While these approaches offer insights and improve analyses at the company level, this system creates inconsistencies.

Production-based data challenges

Poultry is the second most consumed meat worldwide, and the production and trade supply chains are highly integrated regionally and beyond [22]. Poultry companies collect multiple levels of data throughout the supply chain (Fig. 1). However, even for integrated companies, these data are often siloed as opposed to integrated into a data warehouse (Fig. 1), where the data can be easily available to be processed and analyzed more robustly. In parallel, private companies like Agri Stats, Inc., provide redacted data back to poultry companies on a variety of metrics related to production and economics at an industry level. The ability to utilize these types of data allows companies a practical benchmarking tool with respect to production and economics.

In addition to separate data silos for breeders, hatcheries, grow-out farms. and processing plants, additional data including pathology, titer, and nutritional data are rarely easily available and poorly integrated. This is especially a relevant topic as new data streams described below add new additional data. However, the availability of reliable and robust data sets is often expensive and rare [5].

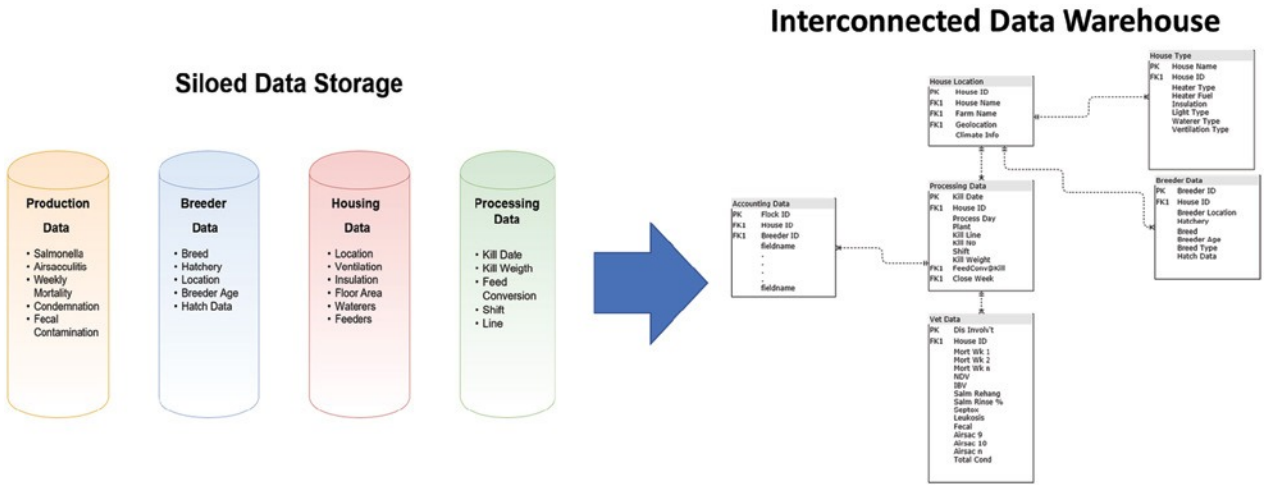

Figure 1. Example of siloed and integrated data from breeding farms, hatcheries, grow-out farms, and poultry processing plants. For integrated companies, the ability to connect all these data into a data warehouse is integral toward facilitating data analysis and identifying risk factors across the entire supply chain. Not included in the figure are data from diagnostic labs such as titer and pathology data in addition to nutritional feed formulations and publically available data including data from FoodNet (CDC).

New data streams

Although Hans Hoffmann was an artist, his quote, "The ability to simplify means to eliminate so that the necessary may speak" has significant relevance in the data sciences. As we enter the 2020s, a combination of Moore's Law with respect to computer processing power, Big Data, 5th generation mobile networks, and the Internet of Things (IoT) will facilitate the connectivity of literally billions of devices in multiple fields including agriculture [23]. Below is a short summary of several potential promising techniques that will create new data streams and new integration and analyses challenges with potential ML-based solutions.

Molecular: The big data associated with next-generation sequencing (NGS) can be incorporated into risk analysis and predictive algorithms that aid the poultry industry to anticipate the areas of weakness in the food safety system [24] with the goal of using these data to improve monitoring and anticipate vulnerabilities in food safety [25–27].

Welfare: Sensors, cameras, and microphones for acoustic monitoring represent a relatively new approach toward monitoring welfare and production in animal agriculture and poultry. Specifically, understanding how chickens behave using sensors, cameras, and microphones can offer producers a new tool in identifying disease [28, 29]. In general, these types of automated continuous approaches in poultry are primarily at the research [28], development, and prototype level [7]. However, the potential applications are highly practical. For example, automated imagery analytics based on supervised machine learning-based approaches could be used to detect morbidity, presymptomatic signs of mortality [23, 29], and ectoparasite infestation [28].

Environmental: Meteorological variables with high spatial resolution including outside temperature, precipitation, wind speed, dew points, and air quality are publically available data that can be collected remotely. Open source data include NOAAs weather database [30], the Parameter elevation Regression on Independent Slopes Model (PRISM) [31], and the weather underground [32].

Web Scraping: Web Scraping is an automated (versus manually downloading relevant material) method of extracting large amounts of data from Websites and other online sources. The types of data that can be extracted include economic [33], environmental (e.g., temperature and humidity) as described above, and a whole list of other data (e.g., crop production reports and stock-based futures). The data collected can then be used in association with other data available to run further analysis. The mechanics of Web Scraping typically involve the utilization of an application programming interface (API) which allows for the interaction between your query application and the relevant data.

While the above approaches offer interesting new data sets that may offer new levels of understanding with respect to food safety, production efficiency, and welfare, it is important to heed the words of Hans Hoffman stated at the beginning of this section. To that point, the task of extracting useful information from a database is known as Knowledge Discovery in Database (KDD) which is often used interchangeably with the term Data Mining. Both describe the process of simplification of large data sets. As these types of data sets have gotten larger, more complex, and diverse, new concepts and terms have evolved to organize these data. As the name implies, the term "Big Data" is used to describe a data set that is Big (i.e., terabytes or more). More specifically, Big Data is also described using the four V's that represent the characteristics of the data: volume (i.e., the total amount of data); velocity in which data are being produced and or

ingested into a database; variety in data types such as text, images, and sound; and veracity of the data, because when the first three V's (velocity, volume, and variety) increase, veracity typically decreases [34].

The ability to put all the data into a data warehouse that continuously ingests new data and stores the data securely via a cloud or block storage approach in a relational database system (e.g., SQL) is fundamental toward an improved data analysis. While this review does not focus on these aspects of data ingestion and organization, it is important to at least recognize this fundamental necessity of integrating all the available in order to realize the potential of the below described ML-based analyses.

Part II: Types of machine learning and relevant examples in poultry

The ability to analyze the data described above requires advanced approaches. Because there is no such thing as an algorithm that is appropriate for every situation, the following section is meant to define and provide examples of commonly used ML-based approaches in poultry and food safety.

Relevant ML-based definitions and examples

Statistics is a science that is focused on the appropriate use of the data to learn, understand, make inferences, and ultimately make predictions and/or understand causality [35]. However, in order to properly utilize statistics, various assumptions about the nature of the data need to be made. For example, in linear regression, the presence of outliers in the data are not well accounted for [36, 37]. By contrast, in random forest, which is a ML-based algorithm that makes no assumption about data distribution, the presence of outliers is accounted for [38]. This is relevant for food safety in that relevant outliers need to be "recognized" and "valued" in the analyses.

Machine learning is a subset of AI, which allows for the development of novel algorithms and statistical models without using explicit instructions. In effect, the "machine" (i.e., software) trains itself on a continual basis and uses new data to "learn." Specifically, ML is based on the development of algorithms that use statistical models and pattern recognition to let the computer learn or improve its performance without being explicitly instructed or programmed [36, 37]. From a practical perspective in poultry production, this allows producers to make predictions on variable such as egg weight (Fig. 2) and better understand historic data with respect to the presence of *Salmonella* in raw poultry (Fig. 3).

Supervised and unsupervised learning are the two major classifications of ML. Supervised learning algorithms have a target continuous or categorical variable (i.e., regression, recursive partitioning algorithms), while unsupervised learning algorithms do not have a target variable or known label(s) within a variable; they have to search for them via techniques such as cluster analysis [35]. The following is a review of various relevant ML-based methods in poultry food safety and production which represent a group of

Figure 2. Comparison of random forest prediction data against control data for average egg weight over the lifespan of a flock. The mean absolute error (MAE) is 1.547 grams and the root mean square (RMS) error is 3.172 grams. The average error between the predicted and actual egg weigh was determined to be less than 3% using data (i.e., egg weight, mortality, average bird weight, bird age, and total eggs produced) from the prior week.

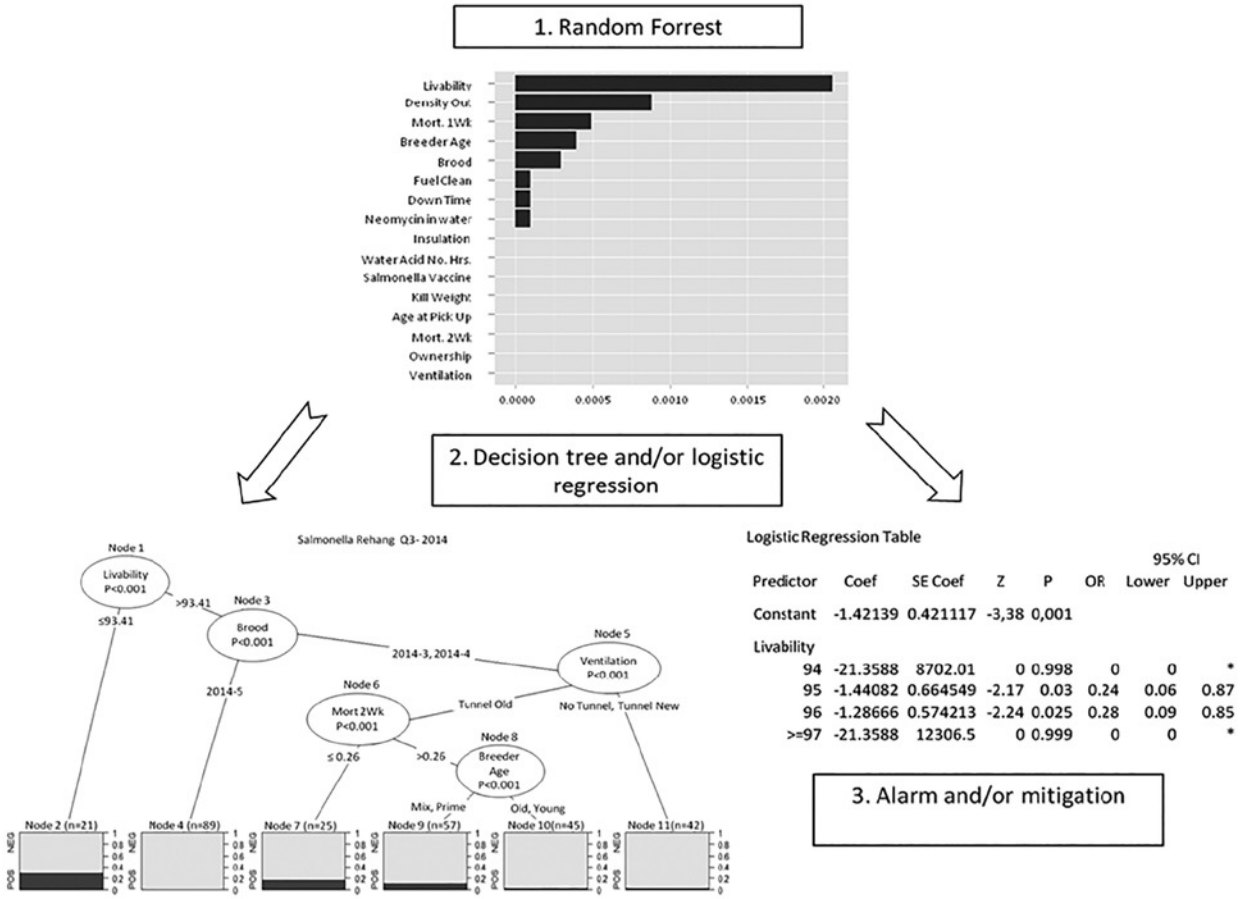

Figure 3. Description of the KDD process performed in broilers to predict the presence of historic *Salmonella* presence/absence at the re-hang line in the processing plant. The diagram shows the analysis performed over one quarter and shows the results for that particular data set. Results are not generalizable to other quarters and companies. In this KDD process, RF was used to predict the important variables from over 50 potential variables. Next, the most important variables were included in a hierarchical decision tree and standard logistic regression in order to identify the specific cutoff points in relevant variables. For example, livability below 93.41% was found to be a predictor for *Salmonella* presence at the re-hang line.

methods including partitioning, hierarchical, density based, grid based, and model based [39].

Simple and linear regression

One of the simplest and oldest versions of ML is regression. Regression describes the relationship between an independent variable or predictor and a dependent variable (i.e., response or target) to help explain the change in the response when the predictor(s) change. The result is a formula represented by a line that can be used to make predictions including interpolations (estimate that lies between two observations) and extrapolations (estimate beyond the observed data). The use of more than one independent variable to predict an outcome is called multiple regression or multivariable regression [40]. There are several types of regression methods, including logistic regression, lasso regression, ridge regression, and tree model regression [38]. With respect to poultry, the regression-based models have been used to analyze poultry growth curves under different diets [41], food quality characteristics [42], and the detection of contaminants in poultry carcasses [43].

Logistic regression is generally preferable when the outcome is a categorical dichotomous variable (e.g., *Salmonella* positive or *Salmonella* negative). Logistic regression models will describe the probability of an output given a variable or risk factor, such as the risk of a disease given some risk factor [44]. Logistic regression models are often used when studying the presence or absence of *Salmonella* and other disease where presence/absence is a useful dependent variable. [45, 46]

Compared to other ML-based techniques, regression models perform well in situations with reduced number of variables, assuming the data meet certain assumptions. However, when there are reduced number of observations which are often common in poultry food safety and production, models such as Lasso and Ridge Regression can be used that "penalize" variables that are poorly represented [36, 37].

Support vector machine

Support vector machine (SVM) is a classification algorithm widely used for the analysis of high dimensional data,

where the number of variables also called features exceeds the number of observations [47]. SVM searches for observations that lie at the edge(s) that is between the observations of different classes and different values for the target variable and use these variables to find the optimal line or hyperplane that identifies different classes for the target variable [35, 48]. In short, it categorizes high-dimensional continuous data. While it is currently primarily used in fields like bioinformatics, gene expression, and image recognition [47], SVM has been used to characterize the poultry meat via near-infrared (NIR) spectroscopy [42] and auditory sounds from chickens including gurgling sounds (e.g., rales) associated with several infectious respiratory agents in poultry [29].

Tree models

Tree models such as, decision trees, are commonly used ML-based technique, because they are easy to view, understand, and explain (Fig. 3). Specifically, a tree model can be used as a regression-based technique if the model is predicting a continuous variable or as a classification if predicting a categorical variable [48]. Generally speaking a decision tree consists of a root, branches, nodes, leaves, and rules or decisions [35]. The tree is built using a recursive partitioning algorithm that selects the best variable (root) partitioning point of an independent variable (rule or decision) that can be used to predict a dependent variable [48]. This partitioning (split) point is used to define two groups of observations (i.e., decision tree branches), those that perform better in predicting and those that do not. Next, the observation of each branch of the algorithm selects again an independent variable (node) and repeats the previously described process until no more observations are left or partitioning is no longer possible [35, 48]. Commonly the choice of the optimal splitting point is done by using a measure of dispersion such as a Gini index [35, 37]. Tree models are a widely used ML-based algorithm and have shown good performance in predicting egg weight in experiments compared with other regression model [49]. In addition, tree models have been used to predict symptoms of disease in poultry, using audio signal processing [50]. In Fig. 3, a hierarchical tree model was used following the selection of various variables identified by RF to identify the cutoff points associated with the dependent variable *Salmonella* as detected in the processing plant.

Ensemble tree models

Ensemble modes create multiple single models and combine them to a single one that performs better than each one separately. Examples of ensemble models include RF and Boosting [48]. In general, the ensemble models allow for greater confidence with respect to repeatability, effect size, and significance. Also, in general, these ensemble methods are preferential to single regression trees because they use at least two sources of randomization during construction [51–53]. Specifically in an ensemble approach, the data are randomly selected using non-parametric boot-strap based methods [54]. Bootstrap-based methods "draw with replacement" and thus allow multiple calculations (e.g., confidence intervals, means, standard errors, and regression) as opposed to just one. Those data (i.e., in-bag samples) and the data that remain are the "out of bag" samples. Next classification and regression trees are used to build a non-pruned tree with the in-bag samples where a small number of explanatory variables are randomly chosen to determine the best split at each node. Each node split is based on the Gini index [55]. The out-of-bag samples are used as an estimate of performance of each tree and provide a way to rank the importance of the explanatory variables through the calculation of the mean decrease accuracy [38, 53, 54].

Random forest

Random forest is a statistical tool that works by analyzing a random selection of variables and observations to then create a decision tree with a subset of data [48, 55]. Since it is an ensemble model, it can create hundreds of decision or classification trees and then combine the results of those trees to produce a model that is more predictive [38, 51, 53]. The results indicate the variables that were more important in predicting the dependent variable and also how efficient the model is in producing the partition. One of the advantages of this algorithm is that hundreds of variables can be included as independent variables which are not possible for other statistical techniques including regression. The model also performs well using variables of low correlation, also known as noise, which is an important asset when the goal is knowledge discovery and prediction [56]. For example, RF is commonly used to analyze the microarray data for gene selection from a large number of genes and noise [38, 57]. One practical application of this approach uses RF to understand the genotypic molecular mechanisms underlying egg shell quality during the production cycle [58].

The ability to use variables with low correlation and being able to analyze hundreds of variables simultaneously are key advantages of RF for the analysis of poultry based data. To that point, RF has been used for organ and meat quality classification [42, 59], food safety for predicting the prevalence of *Listeria* spp. in pastured poultry environments using various types of meteorological data [27], and poultry health to predict highly pathogenic avian influenza (H5N1) outbreaks [60].

We have used this approach, as can be seen in Fig. 3, as a way to reduce the number of exploratory variables from close to 100 to below 10. As Fig. 3 shows, following the use of RF, we then utilized both conditional decision trees and single logistic regression to further clarify specific cut points with respect to livability, flock, ventilation, and breeder age (Fig. 3).

In Fig. 2, we used RF to predict egg weights 1 week into the future using historic data. While the average error

between the predicted and actual egg weights was determined to be less than 3% based on the egg weight, mortality, average bird weight, bird age, and total eggs produced from the prior week, the model was not as accurate at capturing week to week variations. However, the potential for improvements in predictions exists using additional data including more data points and the consideration of additional variables including additional lag data.

Boosting

Boosting is another ensemble, step-wise tree-based model that works in a similar fashion to RF in that Boosting constructs multiple decision trees to combine them and produce a better final model [48, 61]. In Boosting, each fitted model at every step in the tree attempts to compensate for the shortcomings of the previous fitted models [61]. The Boosting algorithm is considered ideal for observations that are difficult to classify [48, 61]. Specifically, the algorithm identifies observations that are difficult to classify in the first model and "tries" to classify them in the next steps until all the observations are classified. While Boosting is ideal for helping with classification, Boosting does not perform well in situations where the data have noise [48]. Consequently, when the data have noise, it is preferable to use an algorithm that includes bagging such as RF. In poultry, Boosting has been used as model to predict the presence of *Listeria* spp. in the poultry environment [51]. In addition, Boosting has been used in food safety using genome data from nontyphoidal *Salmonella* strains collected as part of the National Antimicrobial Resistance Monitoring System (NARMS) program, to predict antimicrobial minimum inhibitor concentration (MICs) values, achieving overall accuracies of 95–96% within a ± 1 2-fold dilution factor [62].

Artificial neural networks

Artificial neural networks or neural networks are supervised learning techniques inspired by the structure of biological neural networks and try to emulate how we think the brain works [63]. They have become useful because they do not require making any assumptions about the data. Artificial neural networks can be used with categorical and continuous variables, and can use a large number of predictors [35, 37]. The method takes multiple values of the input layer and using a function called the activation function, classifies the output using a user-defined criteria [35]. The simplest form of a neural network is the perceptron or single layer perceptron (SLP) and the basis for the more advanced models like multilayer perceptron model (MLP), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) [63]. Artificial neural networks are computational models that are utilized to identify the structure in large high-dimensional datasets [64]. The technique has been used for the analysis of poultry data including egg classification, where dielectric spectroscopy and neural network were used to detect egg freshness [56]. In addition, neural networks were used among other ML-based algorithms to optimize poultry premises selection in order to improve poultry farm efficiency [65].

While this review article does not focus on the applications of machine learning in poultry genetics, it should also be mentioned that neural networks and other deep learning-based techniques are used for genomic-based linear prediction (GBLUP) and selection [66]. Specifically, methods such as genomic-based linear prediction (GBLUP), Bayesian regression, random forest, and artificial neural networks have been used for genomic selection to improve the production efficiency in livestock [67].

Practical approaches toward analysis

While there are several open-source and non-open-source statistical software platforms and packages specific to various computer languages, from a practical perspective, selecting an appropriate software analysis tool is dependent on the skill set of the organization itself. Some prepackaged programs exist that are useful for exploratory data analysis. However, statistics and ML-specific packages written for statistics-oriented programming languages allow for more in-depth insights through software written in-house or via a third party. Initially, the software can be very useful; however, over time, users often recognize the advantages of using a computer language to manage and perform multi-level tasks with several steps, ranging from data capture, data cleaning, data manipulation, variable creation, merging, sub-setting, exploratory data analysis, plotting, running multiple analyses, model selection, model testing, model evaluation, and model and deployment often in a cyclical manner. Thus, many users tend to move to a computer language to perform these jobs. Popular open-source computer languages used to perform the tasks described above include Python [68] and the R Project [1], which both have large amounts of the required packages to carry out all the steps involved in a Knowledge Discovery in Database including the ML-based techniques described in this article. However, one of the main advantages of the R Project software is that the functionality of the software can be extended by the addition of small groups of files and resources packed as a collection of functions and computer programs, called packages. Since this software has thousands of packages used in both academia and industry, it provides unique flexibility to analyze and visualize data and results. In addition, there are additional R packages that can be used to build user interfaces in order to integrate custom models with various Web services. However, it should be acknowledged that there are several elements that are involved in a software development project and not all can be done by using just one computer language. Therefore, the R Project

can be used to do most of the core jobs, especially data management, visualization, and analyses, and other programming languages can be used to perform other tasks, such as data management and server storage, integration, and Web development. While the R Project does offer flexibility to analyze and visualize data, it does require a certain level of technical expertise in both computer programming and statistics. The learning curve is generally described as steep. In addition, while R is a powerful software computer language, it takes longer to process data relative to lower level languages. In addition, the combination of large data sets and complex ML-based approaches toward analysis is becoming more common [69, 70]. Therefore, it is becoming more common to utilize R packages that utilize techniques such as parallel computing, which while not discussed in this review offer the potential to decrease processing time by executing multiple computing tasks simultaneously [69–71].

Conclusion and predicting the future

Yogi Berra, the former New York Yankees manager and player, once said, "It's tough to make predictions, especially about the future." ML and AI are the powerful methodological approaches to predict the future and understand the past. While the refinement and further development of these types of ML- and AI-based techniques are integral toward improving our understanding of poultry production and food safety, it would be naïve to think that these tools are the end-all and be-all of poultry production. Even in a "fully developed" AI world, things like human instincts, institutional knowledge, and random variability still need to be acknowledged and integrated into decision-making. This review aims to define and describe the current applications of ML-based approaches for poultry with respect to production and food safety. If ML- and AI-based approaches are going to prove to be "revolutionary," we need to make sure we use the appropriate methods for the question(s) asked and acknowledge the flaws. Unfortunately, if our knowledge about the variables associated with that outcome are incomplete, we run the risk of implementing and institutionalizing a new flawed approach that may or may not be better than the current system.

References

1. Team RC. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.
2. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. Journal of Computational and Graphical Statistics 2006;15(3):651–74.
3. Liaw A, Wiener M. Breiman and Cutler's random forests for classification and regression, R package version 4.6-12. Vienna: R Foundation for Statistical Computing; 2015.
4. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12(1):77.
5. Silver D, Huang A, Maddison C, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484–9. doi: 10.1038/nature1696168.
6. Gibney E. Self-taught AI is best yet at strategy game Go. Nature 2017;10(1):68–74.
7. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. npj Digital Medicine 2018;1:40. doi: 10.1038/s41746-018-0048-y.
8. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics 2015;16(6):321–32.
9. Li H, Chen H. Human vs. AI: an assessment of the translation quality between translators and machine translation. International Journal of Translation, Interpretation, and Applied Linguistics (IJTIAL). 2019;1(1):43–54.
10. Sharpe C, Wiest T, Wang P, Seepersad CC. A comparative evaluation of supervised machine learning classification techniques for engineering design applications. Journal of Mechanical Design 2019;141(12):121404. doi: 10.1115/1.4044524.
11. Marvin HJ, Janssen EM, Bouzembrak Y, Hendriksen PJ, Staats M. Big data in food safety: an overview. Critical Reviews in Food Science and Nutrition 2017;57(11):2286–95.
12. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94.
13. Gøtzsche PC. Mammography screening is harmful and should be abandoned. Journal of the Royal Society of Medicine 2015;108(9):341–5.
14. Chai S, Cole D, Nisler A, Mahon BE. Poultry: the most common food in outbreaks with known pathogens, United States, 1998–2012. Epidemiology and Infection 2017;145(2):316–25.
15. FDA-NARMS. NARMS retail meat surveillance laboratory protocol 2020. Available from: URL: https://www.fda.gov/media/93332/download
16. Lee SH, Jung BY, Rayamahji N, Lee HS, Jeon WJ, Choi KS, et al. A multiplex real-time PCR for differential detection and quantification of *Salmonella* spp., *Salmonella enterica* serovar Typhimurium and Enteritidis in meats. Journal of Veterinary Science 2009;10(1):43–51.
17. Blevins RE, *et al*. Historical, current, and future prospects for food safety in poultry product processing systems. In: Food and feed safety systems and analysis. USA: Academic Press, 2018. p. 323–45.
18. Administration FAD. Prevention of *Salmonella enteritidis* in shell eggs during production, storage and transportation. Final rule. Federal Register. 2009;(130):33029.
19. FSIS. Food Safety and Inspection Service 2020 Annual Sampling Program Plan 2020. Available from: URL: https://www.fsis.usda.gov/wps/wcm/connect/e8c5ea4e-5c48-452d-b6e3-b21ccc769cf6/fsis-annual-sampling-plan-fy2020.pdf?MOD=AJPERES
20. Ebel ED, Williams MS, Amann DM. Quantifying the effects of reducing sample size on 2-class attributes sampling plans: implications for United States poultry performance standards. Food Control 2020;111:107068.

21. Ebel ED, Williams MS. Assessing the effectiveness of revised performance standards for *Salmonella* contamination of comminuted poultry. Microbial Risk Analysis 2019:100076.
22. Rama EN, Singh M. Regulations in poultry meat processing. In: Food safety in poultry meat production. Cham: Springer; 2019. p. 293–301.
23. Balachandar S, Chinnaiyan R, editors. Internet of Things based reliable real-time disease monitoring of poultry farming imagery analytics. In: International Conference on Computer Networks, Big data and IoT. Cham: Springer; 2018. p. 615–620.
24. Feye KM, Ricke SC. Establishment of a standardized 16S rDNA library preparation to enable analysis of microbiome in poultry processing using illumina MiSeq platform. In: Foodborne bacterial pathogens. Springer; 2019. p. 213–27.
25. Oakley BB, Morales CA, Line J, Berrang ME, Meinersmann RJ, Tillman GE, et al. The poultry-associated microbiome: network analysis and farm-to-fork characterizations. PLOS One. 2013;8(2):e57190.
26. Ricke SC, et al. Unraveling food production microbiomes: concepts and future directions. In: Food and feed safety systems and analysis. USA: Academic Press, 2018. p. 347–74.
27. Kim SA, Park SH, Lee SI, Owens CM, Ricke SC. Assessment of chicken carcass microbiome responses during processing in the presence of commercial antimicrobials using a next generation sequencing approach. Scientific Reports 2017;7(1):1–14.
28. Abdoli A, Murillo AC, Gerry AC, Keogh EJ. Time series classification: lessons learned in the (literal) field while studying chicken behavior. arXiv preprint arXiv:191205913. 2019.
29. Rizwan M, *et al*. Identifying rale sounds in chickens using audio signals for early disease detection in poultry. In: 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP). USA: IEEE; 2016. p. 55–9.
30. NOAA-NCDC. Climate Data Online 2020. Available from: URL: https://www.ncdc.noaa.gov/cdo-web/ [accessed on: 2020 May 16]
31. Daly C. Descriptions of PRISM spatial climate datasets for the conterminous United States. Corvallis, OR: PRISM Climate Group, Oregon State University; 2013. (PRISM Doc., 14 p).
32. Weather Underground 2020. Available from: URL: https://www.wunderground.com/
33. Hillen J. Web scraping for food price research. British Food Journal 2019;121(12):3350–61. doi: 10.1108/BFJ-02-2019-008.
34. Frampton M. Complete guide to open source big data stack. USA; Apress, 2018.
35. Ramasubramanian K, Singh A. Machine learning using R. New Delhi, India: Apress; 2017.
36. Ghatak A. Machine learning with R. Singapore: Springer; 2017.
37. Swamynathan M. Mastering machine learning with python in six steps: a practical implementation guide to predictive data analytics using python. India: Apress; 2019.
38. Breiman L. Random forests. Machine Learning 2001;45(1):5–32.
39. Majumdar J, Naraseeyappa S, Ankalaki S. Analysis of agriculture data using data mining techniques: application of big data. Journal of Big Data 2017;4(1):20.
40. Haroon D. Python machine learning case studies. USA; Apress; 2017.
41. Adeola O, Ileleji K. Comparison of two diet types in the determination of metabolizable energy content of corn distillers dried grains with solubles for broiler chickens by the regression method. Poultry Science 2009;88(3):579–85.
42. Santana EJ, Geronimo BC, Mastelini SM, Carvalho RH, Barbin DF, Ida EI, et al. Predicting poultry meat characteristics using an enhanced multi-target regression method. Biosystems Engineering 2018;171:193–204.
43. Lawrence KC, Windham WR, Park B, Heitschmidt GW, Smith DP, Feldner P. Partial least squares regression of hyperspectral images for contaminant detection on poultry carcasses. Journal of Near Infrared Spectroscopy 2006;14(4):223–30.
44. Kleinbaum DG, Klein M. Introduction to logistic regression. In: Logistic regression. New York, NY: Springer; 2010. p. 1–39.
45. Ashgar SS. Campylobacteriosis in Makkah City, Saudi Arabia. The Egyptian Journal of Medical Microbiology 2013;38(1234):1–4.
46. St. Amand JA, Otto SJ, Cassis R, Annett Christianson CB. Antimicrobial resistance of *Salmonella enterica* serovar Heidelberg isolated from poultry in Alberta. Avian Pathology 2013;42(4):379–86.
47. Carugo O, Eisenhaber F. Data mining techniques for the life sciences. Vol. 609. USA: Humana Press; 2010.
48. Williams G. Data mining with Rattle and R: the art of excavating data for knowledge discovery. New York/ Dordrecht/Heidelberg/ London: Springer Science & Business Media; 2011.
49. Orhan H, Eyduran E, Tatliyer A, Saygici H. Prediction of egg weight from egg quality characteristics via ridge regression and regression tree methods. Revista Brasileira de Zootecnia 2016;45(7):380–5.
50. Carroll BT, Anderson DV, Daley W, Harbert S, Britton DF, Jackwood MW. Detecting symptoms of diseases in poultry through audio signal processing. In: 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE; 2014. p. 1132–35. Available from: URL: https://www.semanticscholar.org/paper/Detecting-symptoms-of-diseases-in-poultry-through-Carroll-Anderson/461a977e4d0dd918e4fb35beffd9c30e5e1f9988
51. Golden CE, Rothrock Jr MJ, Mishra A. Comparison between random forest and gradient boosting machine methods for predicting *Listeria* spp. prevalence in the environment of pastured poultry farms. Food Research International. 2019;122:47–55.
52. Philibert A, Desprez-Loustau ML, Fabre B, Frey P, Halkett F, Husson C, *et al*. Predicting invasion success of forest pathogenic fungi from species traits. Journal of Applied Ecology 2011;48(6):1381–90.
53. Liaw A, Wiener M. Classification and regression by randomforest. R News 2002;2(3):18–22.
54. Efron B, Tibshirani RJ. An introduction to the bootstrap. Boca Raton, FL: CRC Press; 1994.

55. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC Press; 1984.

56. Soltani M, Omid M. Detection of poultry egg freshness by dielectric spectroscopy and machine learning techniques. LWT- Food Science and Technology 2015;62(2):1034–42.

57. Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7(1):3.

58. Ramzan F, Klees S, Schmitt AO, Cavero D, Gültas M. Identification of age-specific and common key regulatory mechanisms governing eggshell strength in chicken using random forests. Genes 2020;11(4):464.

59. Philipsen MP, et al. RGB-D segmentation of poultry entrails. In: International Conference on Articulated Motion and Deformable Objects. Cham: Springer; 2016. p. 168–74.

60. Kane MJ, Price N, Scotch M, Rabinowitz P. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics 2014;15(1):276.

61. Friedman JH. Stochastic gradient boosting. Computational Statistics & Data Analysis 2002;38(4):367–78.

62. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, et al. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. Journal of Clinical Microbiology 2019;57(2):e01260–18.

63. Beysolow II T. Introduction to deep learning using R: a step-by-step guide to learning and implementing deep learning models using R. New York, NY: Apress; 2017.

64. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.

65. Rahimi I, Behmanesh R. Improve poultry farm efficiency in Iran: using combination neural networks, decision trees, and data envelopment analysis (DEA). 2012.

66. Zingaretti LM, Gezan SA, Ferrão LFV, Osorio LF, Monfort A, Muñoz PR, et al. Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. Frontiers in Plant Science 2020;11:25.

67. Hayes BJ, Lewin HA, Goddard ME. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. Trends in Genetics 2013;29(4):206–14.

68. Van Rossum G. Python programming LInguage. In: USENIX Annual Technical Conference. Wilmington: Python Software Foundation; 2007.

69. Schmidberger M, Morgan M, Eddelbuettel D, Yu H, Tierney L, Mansmann U. State-of-the-art in Parallel Computing with R. Journal of Statistical Software. 2009;47(1):31.

70. Rossini AJ, Tierney L, Li N. Simple parallel statistical computing in R. Journal of Computational and Graphical Statistics 2007;16(2):399–420.

71. McCallum E, Weston S. Parallel R. Sebastopol, CA: O'Reilly Media, Inc.; 2011.