An ensemble learning approach to identify pastured poultry farm practice variables and soil constituents that promote *Salmonella* prevalence

Nisha Pillai a,†, Moses B. Ayoolab, y, Bindu Nanduric, Michael J. Rothrock Jr d, Mahalingam Ramkumar a,*

a Department of Computer Science and Engineering, Mississippi State University, MS 39762, USA
b Geosystems Research Institute, Mississippi State University, MS 39762, USA
c Department of Comparative Biomedical Sciences, Mississippi State University, MS 39762, USA
d Egg Safety and Quality Research Unit, USDA-ARS U.S. National Poultry Research Center, GA 30605, USA

ARTICLE INFO

Keywords:
Pastured poultry farm
*Salmonella*
Ensemble learning
Food safety

ABSTRACT

Animal sourced foods including contaminated poultry meat and eggs contribute to human non-typhoidal salmonellosis, a foodborne zoonosis. Prevalence of *Salmonella* in pastured poultry production systems can lead to contamination of the final product. Identification of farm practices that affect *Salmonella* prevalence is critical for implementing control measures to ensure the safety of these products. In this study, we developed predictive models based predominantly on deep learning approaches to identify key pre-harvest management variables (using soil and feces samples) in pastured poultry farms that contribute to *Salmonella* prevalence. Our ensemble approach utilizing five different machine learning techniques predicts that physicochemical parameters of the soil and feces (elements such as sodium (Na), zinc (Zn), potassium (K), copper (Cu)), electrical conductivity (EC), the number of years that the farms have been in use, and flock size significantly influence pre-harvest *Salmonella* prevalence. Egg source, feed type, breed, and manganese (Mn) levels in the soil/feces are other important variables identified to contribute to *Salmonella* prevalence on larger (>3 flocks reared per year) farms, while pasture feed and soil carbon-to-nitrogen ratio are predicted to be important for smaller/hobby (<3 flocks reared per year) farms. Predictive models such as the ones described here are important for developing science-based control measures for *Salmonella* to reduce the environmental, animal, and public health impacts from these types of poultry production systems.

1. Introduction

*Salmonella*, a rod-shaped, motile, gram-negative bacterium, is among the five topmost important foodborne pathogens in the United States (https://www.cdc.gov/foodsafety/foodborne-germs.html). *Salmonella* is the causative agent of salmonellosis and gastroenteritis which manifest as diarrhea, vomiting, fever, abdominal pain, and cramps. While most foodborne infections get resolved without treatment, there are also numerous cases where hospitalization is required. According to the Centers for Disease Control and Prevention (CDC), non-typhoidal *Salmonella* is the etiology of 1.35 million infections, 26,500 hospitalizations and 420 deaths in the United States every year (https://www.cdc.gov/Salmonella/index.html). The USDA estimates the cost impact of *Salmonella* to be over $3 billion annually (Hoffmann, 2015).

Poultry is a known reservoir of *Salmonella* spp., and consumption of contaminated poultry products is a predominant source of human infections. *Salmonella* contamination in poultry production systems is a worldwide major public health concern as millions of people depend on poultry products for their daily nutritional source. Additionally, occasional recall of poultry products due to *Salmonella* contamination also has a negative impact on both farm businesses and the nation’s economy as a whole. Pastured poultry, where chickens are raised outdoor in movable pens and given access to fresh pasture on a daily basis, is becoming more popular in recent years as an alternative to conventional poultry production. This is due to perceived better ethical standards of raising livestock, and additional nutritional value such as increased levels of Omega 3, vitamin A and E content (Hilimire, 2012).

It is important to understand the impact of diverse environmental/management variables on the colonization and prevalence of *Salmonella*...
in pastured poultry for developing control measures to ensure safety and enhance productivity. A recent study (Hwang et al., 2020a) using machine learning approaches to understand the impact of meteorological variables on *Salmonella* prevalence has identified soil humidity, wind gusts, and average temperature as influential factors. Another study (Hwang et al., 2020b) focusing on farm management practices identified the number of years of operation, and specific food sources as influential variables associated with *Salmonella* prevalence.

The objective of this study is to expand the scope of previous studies that utilized farm management practice variables to predict *Salmonella* prevalence using RandomForest algorithm (Hwang et al., 2020b) by including the composition of the soil on which the broilers are raised. We used a combination of traditional and deep machine learning techniques with bagging approach, to predict *Salmonella* prevalence based on the physicochemical properties of feces and soil, and different farm management related variables. Deep learning methods have been proven effective both with low-dimensional and high-dimensional data i.e., the number of observations is significantly lower than the number of features (Clarke et al., 2008; Konietschke et al., 2021), and with noisy and complex data. Combining multiple uncorrelated machine learning models is likely to handle underfitting and overfitting, better than a single model. The variance in the prediction error is reduced by ensemble approaches in comparison to other learning techniques such as logistic regression or support vector machines. Five algorithms (RandomForest, XGBoost, Multi-layer perceptron, Generative adversarial network (GAN), Auto-Encoder) were carefully selected to increase generalizability by employing diverse approaches to improve performance. The RandomForest approach is a well-known bagging-based ensemble approach that minimizes variance and is useful in case of outliers. XGBoost, on the other hand, is a boosting-based ensemble approach that provides better performance, considering the limitations of the previous layer. Given that the available data is imbalanced (has more negative samples than positive samples), and complex, deep neural network layers may overfit. Incorporating a simple multi-layer perceptron into the selection of features/management practices alleviates this problem. In addition, we added two well studied generative modeling techniques. Supervised classification using GANs differentiates positives and negatives based on similarity, while unsupervised auto-encoders learn a meaningful representation of the features and utilize the representation to do classification. A majority voting from these algorithms facilitates a diverse and powerful way for decision-making and is used in this study.

2. Materials and methods

2.1. Sample collection, pathogen isolation and physicochemical analysis

The data utilized in this study was described earlier (Hwang et al., 2020b). A brief description of the materials and methods for quick reference is included in this section. Five independent samples of feces and soil were collected from 11 pastured farms at i) the time of placing the broilers on pasture, ii) halfway through the study period, and iii) around the time of processing. *Salmonella* spp. were cultured and isolated from both feces and soil samples as previously described (Hwang et al., 2020a). Briefly, as a pre-enrichment step, the stomach homogenates remained in the filtered stomacher bags and were incubated overnight at 35 °C. Two different enrichments broths were used to isolate *Salmonella* from these environmental samples: tetrathionate (TT; Becton-Dickinson) broth and brilliant green sulfa with novobiocin (BGS; Becton Dickinson) agar and xylose lysine tergitol-4 (XLT-4; Becton Dickinson) agar. These plates were incubated overnight at 35 °C, and on each plate, three *Salmonella*-like colonies per subsample were selected and confirmed using triple sugar iron agar (TSI; Becton-Dickinson) and lysine iron agar fermentation (LIA; Becton-Dickinson) using an incubation period of 18–24 h at 35 °C. Final confirmation of suspect TSI/LIA isolates was performed using *Salmonella* polyvalent O antisera agglutination (Becton-Dickinson), using manufacturer’s specifications. Physicochemical analysis was done as previously described (Rothrock et al., 2019). The moisture content of the fecal and soil samples was determined by drying overnight at 65 °C and calculating the difference between the wet and dried weights of the soil/feces. Fecal and soil pH and electrical conductivity (EC) were determined using an Orion Versa Star Advanced Electrochemistry Meter (Thermo Fisher Scientific, Waltham, MA, USA) using a 1:5 dilution in distilled water. Fecal and soil samples were submitted to the University of Georgia Soils Testing Laboratory for determining the elemental composition.

Thirty-one distinct farm variables and management practice variables associated with the feces and soil that were inputs for RandomForest algorithm (Hwang et al., 2020b) were used with an ensemble of five different machine learning approaches. Furthermore, the following twenty-four constituents/properties of poultry feces and pastured soil were included as additional input variables for our models: acidity/alkalinity (pH), electrical conductivity (EC), moisture, total carbon (TotalC), total nitrogen (TotalN), carbon-to-nitrogen ratio (C:N), aluminum (A), boron (B), calcium (Ca), cadmium (Cd), chromium (Cr), copper (Cu), iron (Fe), potassium (K), magnesium (Mg), manganese (Mn), molybdenum (Mo), sodium (Na), nickel (Ni), phosphorous (P), lead (Pb), sulphur (S), silicon (Si), and zinc (Zn). Altogether, we used a total of fifty-five variables from individual feces and soil samples as input for our model training, evaluation, and prediction.

2.2. Influential variable prediction

The combined performance of a large number of relatively uncorrelated machine learning models will typically surpass the performance of any of the individual models. Consequently, we developed five different predictive models and analyzed the most influential farm practice variables and physicochemical properties, with the goal of developing actionable recommendations to farmers to reduce *Salmonella* prevalence.

In the rest of this paper, X denotes a feature vector generated from the farm practice variables. In common models, X includes both feces and soil samples, while in others it is restricted to either feces or soil.

a) RandomForest: RandomForest (Ho, 1995) is a powerful algorithm comprised of an ensemble of individual decision trees that work together. Decision tree algorithms are highly dependent on the data they are trained on. A change in the training dataset can have a significant impact on the tree structure. At the same time, RandomForest builds several decision trees from the random samples selected from the dataset with replacement and predicts the outcome based on majority votes. We used 100 trees for the RandomForest and conducted 100 iterations using Scikit-learn library.

b) eXtreme gradient boosting: XGBoost (Chen and Guestrin, 2016) is a decision-tree-based ensemble technique that incrementally improves performance by sequentially adding new models to correct previous models’ errors. It minimizes the weak prediction loss of current models using a gradient descent algorithm when adding new models. We used XGBoost gradient boosting library (Chen and Guestrin, 2016) to build trees with a depth of 6, and 45 iterations for training.

c) Multi-layer perceptron: An MLP (Haykin, 1994) is a feed-forward artificial neural network model which is capable of distinguishing data that cannot be linearly separated. It consists of multiple layers of neural network nodes with a non-linear activation function, and each node is connected to the next layer of nodes with certain weights. Based on the amount of error in the output of a node, the connection weights are adjusted at the learning stage (back propagation). Our model includes a hidden layer with 30 neural network units and a binary classification output layer. Our kernel weights are initialized using he uniform (He et al., 2015) method, where random samples are drawn from a uniform distribution within [-lim, limit]. Here the limit is the square root of (6/(number of input units)). And we use Adam
optimizer (Kingma and Ba, 2014) for stochastic gradient descent to train a model that handles sparse gradients on noisy problems and ReLU (Rectified Linear Unit) (Agarap, 2018) as an activation function.

d) **Stacked auto encoder**: AutoEncoder (Liou et al., 2014) consists of stacked neural network layers that reconstruct the original inputs by taking an unlabeled dataset as an input. It consists of an input layer, multiple hidden layers, a latent or representative layer, and an output layer. The networks are learned using the back propagation algorithm by minimizing the reconstruction error, which measures the differences between the original and reconstructed inputs. The latent layer expects to construct a meaningful representative embedding of the unlabeled data. An autoencoder should be well-balanced i.e, be sensitive enough to the inputs to reconstruct the inputs and insensitive enough to the inputs to have a generalized model. Our encoder and decoder include a hidden layer with 70 neural network units and a latent layer with 30 hidden units. Our kernel weights are initialized using he uniform, Adam optimizer is used for stochastic gradient descent, and ReLU is used as an activation function.

e) **Generative adversarial network**: Generative adversarial networks, or GANs for short, are deep learning methods for generative modeling. Generative modeling involves discovering and learning regularities and patterns in data via unsupervised learning techniques. GAN considers the problem as a supervised learning problem with two submodels: a generator model and a discriminator model. In our *Salmonella* architecture, the generator model tries to generate new samples from the negatives, and the discriminator model tries to distinguish between positives and negatives. Both models are trained together to achieve a generic model that can distinguish between positives and negatives. Using backpropagation, we calculate the gradient of category loss to update the weights of the network and reduce the overall loss of this generic deep neural network. Prediction network uses the combination of positive and negative layer outputs to predict a binary classification result $P(Y=1|Z)$ (*Salmonella* is present when $Y=1$). We use a hidden layer of 70 units for generator and latent layer, 50 units for discriminator layer, and 30 units for latent embedding. We use ReLU activation function, Adam optimizer, and he uniform kernel initializer for the best results.

f) **SHapley additive exPlanations**: We use SHAP (Lundberg and Lee, 2017) to explain the model behavior by computing the effect of the individual features on the model output. It compares the model’s output with and without a specific feature to determine the relative importance of that feature. Feature values are plotted as a function of model output in our dependency plots, and positive SHAP values indicate a higher level of importance for the model than negative values.

g) **Approach**: Algorithm 1 summarizes the overall approach of the feature selection. We select the influential features based on the SHAP prediction. Table 1 shows the important feature variables and their ranks in respective algorithm prediction. The top five features are shown in blue and the top ten in brown.

**Algorithm 1**: Ensemble approach to find influential variables in *Salmonella* prediction. We use RandomForest (RT), eXtreme Gradient Boosting (XGBoost), multi-layer perceptron (MLP), auto-encoder (ENC), and generative adversarial network (GAN) as learning algorithms.

**Data**: Farm dataset with practice variables and soil constituents

**Result**: Most influential variables in the *Salmonella* prediction

for sample type in feces, soil do
  for algorithm in RT, XG, MLP, ENC, GAN do
    model ← Model(data)
    topFeatures ← SHAP (model)
  end
  mostInfluential ← topFeatures agreed by atleast 3 models
end
2.3. Salmonella pathogen prediction

Our study compared three different deep learning algorithms to build *Salmonella* pathogen detection models (Section II-B). In comparison to multi-layer perceptron and auto-encoder models, the generative adversarial network performed better in classifying *Salmonella*. As a preprocessing step, we use unit normalization techniques to rescale and standardize our farm input variables to a notionally general scale. Following that, we apply a random sampling technique to balance our positive and negative sample counts. The resulting features are then used as input to the generative adversarial network (GAN).

a) **Unit Normalization**: We scale every sample to a unit length vector that shrinks/stretches the input feature vector to a unit sphere (Eq. (1)). By normalizing each sample, the vector is scaled to the unit norm regardless of how the samples are distributed.

\[
x = \frac{x}{||x||}
\]

b) **Random Oversampling**: We used random oversampling (Lemaitre et al., 2017) techniques to balance the imbalanced input sample dataset. Random oversampling selects the examples from minority class in random order and add them to the training set to increase the sample size. This sampling technique iterates the sampling until the samples of the minority class equal the samples of the majority class.

3. Results & discussion

*Salmonella* was found in 123 samples out of 720 fecal samples (17%), but only in 12.9% (93/717) of soil samples (Table 2). We used 5-fold cross-validation for all the models to train, evaluate, and predict the data sampled from the dataset. In 5-fold cross-validation, the training set is split into 5 smaller sets. Our model is trained on 4 folds and the resulting model is subsequently validated on the remaining data. The process is repeated 5 times, and then we average the performance. To analyze the effects of the most influential variables in the prediction models, we generated feature importance plots using SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) libraries with the most models, we generated feature importance plots using SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) libraries with the most

### Table 2. Number of feces and soil poultry farm samples/datasets that are positive or negative for the presence of *Salmonella*, used in this study for generating predictive models using machine learning. The common model includes samples from both feces and soil.

|                | Positive Samples | Negative Samples | Total |
|----------------|------------------|-------------------|-------|
| Feces          | 123              | 597               | 720   |
| Soil           | 93               | 624               | 717   |
| Common Model   | 216              | 1221              | 1437  |

An imbalanced dataset with around 85% of valid negative samples could be a contributing factor to the high false negatives. Overall, our learning models are able to predict *Salmonella* prevalence with more than 85% F1-Score prediction confidence.

We also plotted a receiver operating characteristic curve (ROC curve) that shows the performance of our binary classification model at all classification thresholds (Figure 1). The curve is plotted using the parameters true positive rate and false positive rate. Our models perform with an AUC of at least 89%, a higher prediction accuracy than previous report of 83% with a Random Forest model (Hwang et al., 2020b).

3.2. Critical farm practice variables

Elemental metals are naturally occurring constituents of soil where they usually exist at concentrations less than 1000 mg per kilogram of soil as trace elements (mg/Kg or part per million, ppm). Contamination from fertilizer, pesticides, wastewater and organic manure are some of the numerous sources that could alter the level of these elements (Wuana and Okieimen, 2011). Macronutrients such as N, P, K, Ca, Mg and S, and micronutrients namely Cu, Fe, Mn, and Zn are predominant in poultry settings (Ihnat and Fernandes, 1996).

Results from soil model indicate that Na levels in the soil >20 ppm reduce *Salmonella* prevalence (Figure 2). Crippen et al. showed that poultry soil Na in conventional poultry farms ranges between 1200 and 3500 ppm (Crippen et al., 2016). Data analysis from the current study shows that about 302 samples, representing 21% of total samples, with detectable Na had <20 ppm. Increased levels of Na in the form of sodium chloride has been shown to be toxic and inhibitory to the growth of *Salmonella* and its capability to form biofilms (Illdis et al., 2018). Sodium could act as acidifier to inhibit *Salmonella* growth and the use of sodium bisulfate salt is known to reduce *Salmonella* below the infectious dose (Ditoe et al., 2019).

Zn is a microelement reported to be in the lower range of concentration of 10–75 ppm in poultry feces and 20 ppm in other agricultural soils (Crippen et al., 2016; Alengebawy et al., 2021). Our prediction algorithms suggest that Zn >20 ppm in soil will mitigate *Salmonella* prevalence. In vitro analysis of the impact of Zn in the form of zinc sulfate on *Salmonella* shows high sensitivity and complete inhibition of the pathogen at lower concentration of 0.8 mg/mL compared to other pathogens (Surjawidjaja et al., 2004). Zinc could directly act as an antimicrobial as its minimum inhibitory concentration on different strains of *Salmonella* which has been recently determined to be between 400 and 800 mg/L (Mustafa et al., 2021). Our model predicted high levels of K, greater than 5000 ppm in fecal sample, to inhibit the growth and prevalence of *Salmonella*. Potassium as a macronutrient has been reported to be in range 3000–11,000 ppm in poultry soil and feces (Crippen et al., 2016). Similar to Zn and Na, K has antimicrobial activity that has been shown to effectively reduce *Salmonella* colony counts (Batista et al., 2019).

Our model predicts that increased levels of Na, Zn and K will help reduce *Salmonella* prevalence on pastured poultry farms. Based only on feces model, we predict that reduced levels of Cu will reduce *Salmonella* prevalence (Figure 3). Bervanger et al. (2018) estimated the broiler requirement of Cu as an essential micronutrient to be 12.5 ppm. Our feces model predicts equivalent levels of Cu or lower to be suitable for reducing *Salmonella* prevalence. Copper is an important cofactor for enzymes such as laccase, superoxide dismutase, oxidases and dehydrogenase, and contributes to *Salmonella* in vivo fitness (Pezza et al., 2016). Therefore, possible reduction of Cu content in broiler feed could potentially help reduce *Salmonella* viability and prevalence. Soil electrical conductivity (EC) is a measure of soil salinity (Azeez and Van Averbeke, 2012) and an indicator of nutrient levels (Heiniger et al., 2003). Low EC corresponds to low salinity/nutrients and vice versa. EC was identified to be an important factor that regulates *Salmonella* prevalence only in our soil model. Soil EC ≥100 mS/m is recommended to reduce incidence of *Salmonella*, however, 401 out of 717 (representing 56%) of our soil...
samples with EC measurements fall below this value. Crippen et al. (2016) reported normal EC range of soil from conventional poultry farms to be between 400 and 1000 mS/m, which supports our prediction.

As has been noted in prior work from this study (Hwang et al., 2020b), farms operating for a longer time (YearsFarming) possibly have the potential to harbor and perpetuate the spread of foodborne pathogens within poultry and beyond. This could be a result of pathogens such as *Salmonella* developing cross resistance to common cleaning agents, or capability to form biofilms and attach themselves to different surfaces in the pre-harvest poultry environment. Based on our model, three years is the maximum length of time using normal day-to-day operation sanitation regimes before a more intensive, whole operation sanitation is required to mitigate *Salmonella* prevalence.

### 3.3. Production scale

Monitoring the average number of birds on-farm in a year could help in reducing *Salmonella* in the soil and we recommend flock size greater than 1000, if feasible based on the resources available to the pastured poultry operation and state laws/regulations. Contrarily, according to Eguale (2018), a larger flock size is associated with a greater prevalence of *Salmonella* in fecal samples. To reduce the risk of pathogen

### Table 3.
The classification results of *Salmonella* prevalence predictive models. Models are trained with feces/soil physico-chemical properties and farm practice variables.

| Sample Type | Model | Electrical Conductivity (EC) | Magnesium(Mg) | Phosphorus(Ph) | Sodium(Na) | Potassium(K) | Zinc(Zn) | Copper(Cu) | Calcium(Ca) | Flocking Years | GMO feed base feed (BrGMOFree) | Free feed (PaGMOFree) | Flock present on farm | Average Number of Birds on Farm in a year | Soy Free broad feed (BrBFeedFree) |
|-------------|-------|------------------------------|---------------|---------------|------------|--------------|----------|------------|-------------|----------------|-------------------------------|-----------------|--------------------------|--------------------------------|--------------------------------|
| Feces       | RT    | 1                            | 5             | 7             | 8          | 9            | 4        | 5          | 6           | 3              | 14                            |                 | 5                        | 1                         | 1                             |
|             | XG    | 1                            | 3             | 12            | 2          | 7            | 5        | 6          | 12          | 15                          | 6               | 5                        | 1                         | 1                             |
|             | ENC   | 1                            | 3             | 3             | 3          | 3            | 3        | 3          | 8           | 8              | 8                             |                 | 8                        | 8                         | 8                             |
|             | GAN   | 4                            | 1             | 1             | 6          | 1            | 9        | 1          | 6           | 6              | 8                             |                 | 8                        | 8                         | 8                             |
| Soil        | RT    | 1                            | 5             | 6             | 6          | 3            | 7        | 6          | 1           | 8              | 8                             |                 | 8                        | 8                         | 8                             |
|             | XG    | 1                            | 3             | 7             | 8          | 3            | 7        | 6          | 3           | 8              | 8                             |                 | 8                        | 8                         | 8                             |
|             | MLP   | 3                            | 4             | 4             | 5          | 4            | 6        | 6          | 6           | 3              | 6                             |                 | 6                        | 6                         | 6                             |
|             | ENC   | 2                            | 4             | 4             | 5          | 4            | 3        | 6          | 3           | 6              | 6                             |                 | 6                        | 6                         | 6                             |
|             | GAN   | 4                            | 2             | 3             | 5          | 6            | 12       | 9          | 10          | 11             | 11                            |                 | 11                       | 11                        | 11                            |

### Figure 1.
Receiver operating characteristic curve (ROC curve) for *Salmonella* predictive models. The prediction accuracy for *Salmonella* prevalence in feces, soil and common models is at least 89%.

### Figure 2.
Relevant variables for predicting *Salmonella* prevalence from Soil Sample Model. The model suggests avoiding very low values for electrical conductivity (EC), sodium (Na), and flock size to reduce *Salmonella* prevalence in soil.

### Figure 3.
Relevant variables for predicting *Salmonella* prevalence from Feces Sample Model. For feed variables (BrGMOFree, PaGMOFree) 0 and 1 on the x-axis indicate the presence or absence of GMO, respectively. The model recommends low copper (Cu) and pasture and brood feed with GMO to reduce *Salmonella* prevalence.
transmission and the risk of human infection from poultry contact, Tobin et al. (2015) also advocate limiting the flock size.

Considering that flock size is one of the most influential features for Salmonella predictive model, and given the contradictory literature, it is reasonable to divide our samples into smaller hobbyist farms and larger professional farms. We consider the farm units rearing more than three flocks a year to be professional while those raising fewer broilers to be hobby. The hobby farms all had flocks of 50 or less birds, while the professional farms all had flocks with higher bird densities (typically 75 or more). Also, the professional farms sold their birds to local restaurants as well as the community at large, while the hobby farms only sold to community members directly. We performed independent ensemble-based analysis on samples from professional (Table 4) and hobby (Table 5) farms to identify the importance of different input variables in these cases.

1) Recommendations for Professional Farmers: In addition to variables identified in previous section contributing to Salmonella prevalence in general poultry settings, we specifically looked at factors in professional poultry farms (Tables 6 and 7). Our predictive models recommend lowering Mn content of the feed less than 100 ppm to reduce Salmonella in feces (Figure 4). Mn content of poultry feces has been reported to be between 7 and 55 ppm in conventional poultry farm settings. Mn is known to be essential in protecting against nitrosative (Yousuf et al., 2020) and oxidative stress (Rishi et al., 2010) and ultimately against host neutrophil killing (Díaz-Ochoa et al., 2016).

Our soil model suggests that the source of broiler eggs (EggSource) can contribute to Salmonella prevalence (Figure 5). Contamination of the surface and internal content of the egg with Salmonella has been previously reported as the major source of Salmonella outbreak in the United States (Gantois et al., 2009). Whiley and Ross (2015) identified a combination of many factors such as vaccination, stress, frequency of cleaning, and animal feed type to be responsible for Salmonella contamination of the eggs and recommend pasteurization and/or irradiation for decontamination (Whiley and Ross, 2015).

The feces model indicates that differences in the feed based on the top three protein sources (BroodFeed) can impact Salmonella prevalence. Despite low moisture content in the animal feed to discourage pathogen growth, recent analysis shows that about 6 percent of all commercial poultry feeds tested across the United States are contaminated with Salmonella (Shariat et al., 2021). A similar study reported 4 percent Salmonella contamination in pig feed in Vietnam (Minh et al., 2020). These studies suggest that animal feed producers need to adopt stringent measures to mitigate the prevalence of Salmonella and other pathogens. Alternatively, professional farmers could purchase poultry feed from producers who report no or negligible pathogen contamination of their products.

Our feces model also suggests that the breed of the broilers can impact Salmonella prevalence. The slow growing Freedom Ranger breed appears more susceptible to infection. Although, Cornish Cross breed is favored by most commercial poultry farms due to fast growth rate and heavy breast mass, they are known to be at a higher risk for poor mobility (Riber et al., 2018). This could possibly limit the Cornish Cross movement range, make them less healthy and more susceptible to infection.

2) Recommendations for Hobby Farmers: In the hobby farms with just one or two flocks a year, the feces model (Table 6) suggests that lowering the carbon-to-nitrogen ratio in the feed (C:N ≤ 15) will reduce Salmonella carriage than the fast growing Cornish Cross. Although, Cornish Cross breed is favored by most commercial farms due to fast growth rate and heavy breast mass, they are known to be at a higher risk for poor mobility (Riber et al., 2018). This could possibly limit the Cornish Cross movement range, make them less healthy and more susceptible to infection.

Table 4. The most influential variables in the Salmonella prevalence predictive model for Professional (>3 flocks reared) pastured poultry farms dataset. The top five ranks are shown in blue and the top ten in brown.

| Sample Type | Model | Manganese (ppm) | Phosphorous (ppm) | Sulfur (ppm) | Calcium (ppm) | Protein (ppm) | Cyanogenic (ppm) | Egg Source | Poultry Feed | Flock Size | Age of flock in days | Sex of flock | Brood breed (broiler/fee) | Age of birds in days | Loss (My) | Breed of birds from freedom range (broiler/free) | Goals present as ferns |
|-------------|-------|-----------------|-------------------|--------------|---------------|---------------|------------------|-------------|--------------|-----------|-----------------------|-------------|------------------------|---------------------|-----------|-------------------------------|-------------------|
| Professional | RT    | 11              | 7                 | 4            | 1             | 18            | 52               | 1           | 2            | 2         | 6                      | 1            | 11                     | 8                   | 10        | 5                           | 10                |
|              | MG    | 7               | 3                 | 4            | 3             | 28            | 21               | 1           | 2            | 2         | 6                      | 1            | 11                     | 8                   | 10        | 5                           | 10                |
|              | ENC   | 3               | 5                 | 1            | 1             | 9             | 1                | 1           | 2            | 2         | 6                      | 1            | 11                     | 8                   | 10        | 5                           | 10                |
|              | GAN   | 3               | 5                 | 1            | 1             | 16            | 9                | 2           | 2            | 2         | 6                      | 1            | 11                     | 8                   | 10        | 5                           | 10                |

Table 5. The most influential variables in the Salmonella prevalence predictive model for Hobby (<3 flocks reared) pastured poultry farm samples. The top five ranks are shown in blue and the top ten in brown.

| Sample Type | Model | Manganese (ppm) | Phosphorous (ppm) | Sulfur (ppm) | Calcium (ppm) | Protein (ppm) | Cyanogenic (ppm) | Egg Source | Poultry Feed | Flock Size | Age of flock in days | Sex of flock | Brood breed (broiler/fee) | Age of birds in days | Loss (My) | Breed of birds from freedom range (broiler/free) | Goals present as ferns |
|-------------|-------|-----------------|-------------------|--------------|---------------|---------------|------------------|-------------|--------------|-----------|-----------------------|-------------|------------------------|---------------------|-----------|-------------------------------|-------------------|
| Hobby       | RT    | 4               | 12                | 8            | 2             | 8             | 7                | 4           | 7            | 7         | 6                      | 3            | 2                      | 2                   | 6         | 16                          | 14                |
|              | MG    | 3               | 7                 | 4            | 1             | 8             | 7                | 4           | 7            | 7         | 6                      | 3            | 2                      | 2                   | 6         | 16                          | 14                |
|              | ENC   | 1               | 4                 | 4            | 3             | 7             | 4                | 4           | 7            | 7         | 6                      | 3            | 2                      | 2                   | 6         | 16                          | 14                |
|              | GAN   | 1               | 4                 | 4            | 3             | 7             | 4                | 4           | 7            | 7         | 6                      | 3            | 2                      | 2                   | 6         | 16                          | 14                |

N. Pillai et al. Heliyon 8 (2022) e11331
sources of protein in the pasture feed can affect *Salmonella* prevalence. Similar to the suggestion above to ensure pathogen-free brood feed for professional farms, possible withdrawal of certain pasture feed or thorough screening of feed for *Salmonella* is recommended for hobby farms to reduce *Salmonella* in feces. Based on the soil model (Table 7), we recommend increasing the Na content to 10 ppm and maintain Mn levels between 40 and 100 ppm in hobby farms. The recommended levels of Na and Mn in the soil content will be beneficial in limiting the presence of *Salmonella*. Soil model also suggests that maintaining a flock size of \( \leq 25 \) eggs if practical for a given poultry operation, will reduce *Salmonella* (Figure 7).

In conclusion, pre-harvest samples (soil and feces) collected from 11 Southeastern United States pastured poultry farms were used to identify farm practice variables that could predict *Salmonella* prevalence. While we were able to test for 55 variables in our models, we acknowledge that this does not represent an exhaustive account of all possible variables. There are potentially other management, environmental, or physicochemical variables that we were not able to gather data for in our experimental design but may potentially be predictive of *Salmonella* presence in pastured poultry systems. But these results do demonstrate that if the data is available for these other variables, they could be included in these types of ensemble machine learning model approaches. Utilizing an ensemble-based approach that combines RandomForest and XGBoost with three deep learning techniques, we provide management recommendations to reduce *Salmonella*, that will ultimately contribute to food safety and protect the environment. We identified optimal levels of elements in poultry pre-harvest environment to mitigate *Salmonella* presence. Low levels of sodium (20 ppm), zinc (20 ppm), potassium...
(5000 ppm), and electrical conductivity (EC, 100 mS/cm), as well as higher levels of copper (10 ppm) promote Salmonella prevalence. Since influence of flock size was counterintuitive (lower Salmonella prevalence for both very small and very large flock sizes), there was a need to build independent models for small (hobbyist, <3 flocks) farms and large (professional, ≥3 flocks) farms. While Mn, egg source, brood feed and broiler breed were important predictive variables for large farms, pasture feed, soil C:N, and Na were important for small farms. While these analyses are based on over 2000 samples from 41 flocks from 11 farms, collected over 4 years, the management practices on these types of farms are not nearly as uniform as in conventional commercial poultry production in the US. Therefore, these models may not be applicable to all pastured poultry operations across the US. But these models do highlight the utility of using ensemble machine learning methods to develop predictive tools for pastured poultry farmers to reduce food safety issues within their operations. These results will help farmers to develop science-based control measures to reduce Salmonella prevalence in pastured poultry, a sustainable agriculture technique.

Declarations

Author contribution statement

Mahalingam Ramkumar: Conceived and designed the experiments; Analyzed and interpreted the data.
Nisha Pillai, Moses B. Ayoola: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.
Bindu Nanduri: Conceived and designed the experiments.
Michael J. Rothrock Jr.: Contributed reagents, materials, analysis tools and data.

Funding statement

This work was supported by USDA NACA Project [58-0200-0-002].

Data availability statement

Data will be made available on request.

Declaration of interest’s statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgement

We thank the Agricultural Research Service, USDA CRIS Project “Reduction of Invasive Salmonella enterica in Poultry through Genomics, Phenomics and Field Investigations of Small Multi-Species Farm Environments” #6040-32000-011-00-D for providing the data that was used in this study.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Manz, D., Monga, R., Moore, S., Murray, D., Olah, G., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.
Agarap, A.F., 2018. Deep Learning Using Rectified Linear Units (Relu) arXiv e-prints, pages arxiv-1803.
Akratos, C.S., Tekirlepoulou, A.G., Vasiadou, I.A., Vayenas, D.V., 2017. Gocomposting of olive mill waste for the production of soil amendments. In: Olive Mill Waste. Elsevier, pp. 161–182.
Alengbeawoy, A., Abdelkhaled, S.T., Qureshi, S.R., Wang, M.-Q., 2021. Heavy metals and pesticides toxicity in agricultural soil and plants: ecological risks and human health implications. Toxics 9 (3), 42.
Azeem, J., Van Averbeke, W., 2012. Dynamics of soil ph and electrical conductivity with the application of three animal manures. Commun. Soil Sci. Plant Anal. 43, 865–874.
Battista, A.F.F., dos Santos, A.R., da Silva, A.F., Trevisman, D.A.C., Ribeiro, L.H., Campanerut-Sa, P.A.Z., de Abreu Filho, B.A., Junior, M.M., Milcha, J.M.G., 2019. Inhibition of salmonella enterica serovar typhimurium by combined carvacrol and potassium sorbate in vitro and in tomato paste. LWT 100, 92–98.
Berwanger, E., Vieira, S., Angel, C., Kindlein, L., Mayer, A., Ebbing, M., Lopes, M., 2018. Copper requirements of broiler breeder hens. Poultry Sci. 97 (8), 2785–2797.
Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16. ACM, pp. 785–794.
Chollet, F., et al., 2015. Keras.
Clarke, R., Ressom, H.W., Wang, A., Xuan, J., Liu, M.C., Gehan, E.A., Wang, Y., 2008. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat. Rev. Cancer 8 (1), 37–49.
Crippen, T.L., Sheffield, C.L., Byrd, J.A., Esquivel, J.F., Beier, R.C., Yeater, K., 2016. Poultry litter and the environment: physiochemical properties of litter and soil during successive flock rotations and after remote site deposition. Sci. Total Environ. 553, 650–661.
Diaz-Ochoa, V., Del, M., Lee, C.S., Klauw, S., Behnsen, J., Liu, J.Z., Chinn, N., Nuccio, S., 2012. Dynamics of soil ph and electrical conductivity with flock size in hobby farming. LWT 127, 109423.
Eugale, T., 2018. Non-typoidal salmonid sequestrs in poultry farms in central Ethiopia: prevalence and antimicrobial resistance. BMC Vet. Res. 14 (1), 1–8.
Gantois, I., Ducatelle, R., Pasmans, F., Haesebrouck, F., Gantois, I., Ducatelle, R., Pasmans, F., Haesebrouck, F., Humphrey, T.J., Van Creveld, S., 2015. Keras.
Heiniger, R.W., McBride, R.G., Clay, D.E., 2003. Using soil electrical conductivity to improve nutrient management. Agron. J. 95 (3), 508–516.
Heiniger, R.W., McBride, R.G., Clay, D.E., 2003. Using soil electrical conductivity to improve nutrient management. Agron. J. 95 (3), 508–516.
Hwang, D., Rothrock Jr., M.J., Pang, H., Guo, M., Mishra, A., 2020a. Predicting salmonella prevalence associated with meteorological factors in pastured poultry. Cell Host Microbe 19 (6), 814–825.
Hwang, D., Rothrock Jr., M.J., Pang, H., Guo, M., Mishra, A., 2020a. Predicting salmonella prevalence associated with meteorological factors in pastured poultry. Cell Host Microbe 19 (6), 814–825.
Hwang, D., Rothrock Jr., M.J., Pang, H., Guo, M., Mishra, A., 2020a. Predicting salmonella prevalence associated with meteorological factors in pastured poultry. Cell Host Microbe 19 (6), 814–825.
Hwang, D., Rothrock Jr., M.J., Pang, H., Guo, M., Mishra, A., 2020a. Predicting salmonella prevalence associated with meteorological factors in pastured poultry. Cell Host Microbe 19 (6), 814–825.
Hwang, D., Rothrock Jr., M.J., Pang, H., Guo, M., Mishra, A., 2020a. Predicting salmonella prevalence associated with meteorological factors in pastured poultry. Cell Host Microbe 19 (6), 814–825.
Hwang, D., Rothrock Jr., M.J., Pang, H., Guo, M., Mishra, A., 2020a. Predicting salmonella prevalence associated with meteorological factors in pastured poultry. Cell Host Microbe 19 (6), 814–825.
Iliadis, I., Daskalopoulou, A., Simoes, M., Giaouris, E., 2018. Integrated combined effects of temperature, ph and sodium chloride concentration on biofilm formation by...
salmonella enterica ser. enteritidis and typhimurium under low nutrient food-related conditions. Food Res. Int. 107, 10–18.

Kingham, B.P., Ba, J., 2014. Adam: a method for stochastic optimization. In: 3rd International Conference for Learning Representations.

Konietzsche, F., Schwab, K., Pauly, M., 2021. Small sample sizes: a big data problem in high-dimensional data analysis. Stat. Methods Med. Res. 30 (3), 687–701.

Lemaitre, G., Nogueira, F., Artidai, C.K., 2017. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. 18 (17), 1–5.

Liou, C.-Y., Cheng, W.-C., Liou, J.-W., Liou, D.-R., 2014. Autoencoder for words. Neurocomputing 139, 84–96.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Adv. Neural Inf. Process. Syst., vol. 30. Curran Associates, Inc., pp. 4765–4774.

Minh, D.K., Hounmanou, Y.M.G., Mai, H.B.T., Olsen, J.E., Dalsgaard, A., 2020. Prevalence and genomic characterization of salmonella weltevreden in commercial pig feed. Vet. Microbiol. 246, 108725.

Mustafa, G.R., Zhao, K., He, X., Chen, S., Liu, S., Mustafa, A., He, L., Yang, Y., Yu, X., Penttinen, P., et al., 2021. Heavy metal resistance in salmonella typhimurium and its association with disinfectant and antibiotic resistance. Front. Microbiol. 2120.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pezza, A., Postel, L.B., Lopez, C., Soncini, F.C., 2016. Compartment and signal-specific codependence in the transcriptional control of salmonella periplasmic copper homeostasis. Proc. Natl. Acad. Sci. USA 113 (41), 11573–11578.

Riber, A., Van de Weerd, H., De Jong, I., Steenfeldt, S., 2018. Review of environmental enrichment for broiler chickens. Poultry Sci. 97 (2), 378–396.

Rishi, P., Jindal, N., Bhurrhan, S., Tiwari, R.P., 2010. Salmonella–macrophage interactions upon manganese supplementation. Biol. Trace Elem. Res. 133 (1), 110–119.

Rothrock Jr., M.J., Locatelli, A., Feye, K.M., Caudill, A.J., Guard, J., Hiett, K., Ricke, S.C., 2019. A microbiomic analysis of a pasture-raised broiler flock elucidates foodborne pathogen ecology along the farm-to-fork continuum. Front. Vet. Sci. 6, 266.

Shariat, N.W., Feye, K.M., Richards, A.K., Bosher, R., Flores, Z., Rubinelli, P.M., Olson, E.G., Ricke, S.C., 2021. Incidence of salmonella serovars isolated from commercial animal feed mills in the United States and serovar identification using crispr analysis. J. Appl. Microbiol. 130 (6), 2141–2146.

Surjawidjaja, J.E., Hidayat, A., Lesmana, M., 2004. Growth inhibition of enteric pathogens by zinc sulfate: an in vitro study. Med. Princ. Pract. 13 (5), 286–289.

Tobin, M.R., Goldshear, J.L., Price, L.B., Graham, J.P., Leibler, J.H., 2015. A framework to reduce infectious disease risk from urban poultry in the United States. Publ. Health Rep. 130 (4), 380–391.

Whiley, H., Ross, K., 2015. Salmonella and eggs: from production to plate. Int. J. Environ. Res. Publ. Health 12 (3), 2543–2556.

Wuana, R.A., Okieimen, F.E., 2011. Heavy metals in contaminated soils: a review of sources, chemistry, risks and best available strategies for remediation. Int. Sch. Res. Notices 2011.

Yousuf, S., Karlinsey, J.E., Neville, S.L., McDevitt, C.A., Libby, S.J., Fang, F.C., Frawley, E.R., 2020. Manganese import protects salmonella enterica serovar typhimurium against nitrosative stress. Metallomics 12 (11), 1791–1801.