Applications of machine learning in animal and veterinary public health surveillance

J. Guitian* (1), M. Arnold (2), Y. Chang (3) & E.L. Snary (2)

(1) World Organisation for Animal Health Collaborating Centre for Risk Analysis and Modelling, c/o Veterinary Epidemiology, Economics and Public Health Group, Department of Pathobiology and Population Sciences, Royal Veterinary College, Hawkshead Lane, North Mymms, Hatfield, Hertfordshire, AL9 7TA, United Kingdom

(2) World Organisation for Animal Health Collaborating Centre for Risk Analysis and Modelling, c/o Department of Epidemiological Sciences, Animal and Plant Health Agency, Woodham Lane, Addlestone, Surrey, KT15 3NB, United Kingdom

(3) Department of Comparative Biomedical Sciences, Royal Veterinary College, Royal College Street, London, NW1 0TU, United Kingdom

*Corresponding author: jguitian@rvc.ac.uk

Summary

Machine learning (ML) is an approach to artificial intelligence characterised by the use of algorithms that improve their performance at a certain task (e.g. classification or prediction) from data itself and without being explicitly and fully instructed on how to achieve it. Surveillance systems for animal and zoonotic diseases depend upon effective completion of a broad range of tasks, some of them amenable to ML algorithms. As in other fields, the use of ML in animal and veterinary public health surveillance has greatly expanded in recent years. ML algorithms are being used to accomplish tasks that became attainable only with the advent of large datasets, new methods for their analysis and increased computing capacity. Examples include the identification of an underlying structure in large volumes of data from an ongoing stream of abattoir condemnation records, the use of deep learning to identify lesions in digital images obtained during

slaughtering or the mining of free text in electronic health records from veterinary practices for purpose of sentinel surveillance. However, ML is also being applied to tasks that had usually been tackled with traditional statistical data analysis. Statistical models have extensively been used to infer relationships between predictors and disease to inform risk-based surveillance and increasingly, ML algorithms are being used for prediction and forecasting of animal diseases in support of more targeted and efficient surveillance. While ML and inferential statistics can accomplish similar tasks, they have different strengths making one or the other more or less appropriate in a given context.

Keywords

Animal health – Infectious disease – Machine learning – Surveillance – Veterinary public health.

What is machine learning?

The advancement in computing technology and power and the explosion of data generation and storage capability in the last decades have seen the increased use of machine learning (ML) in many areas. ML is a collection of methods built upon statistics, mathematics and computer science that enable automated pattern discovery and model building at scale. Many introductory articles describing the various ML techniques have been produced targeting researchers and scientists in different fields [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. We do not intend to reproduce those efforts but aim to put the ML methods in context of their techniques and purposes in comparison to traditional statistical data analysis and to present ML solutions to specific surveillance tasks that cannot effectively be addressed by traditional statistical data analysis. In this section we will contrast unsupervised ML with the use of descriptive statistics, and supervised ML with the use of statistical modelling (inferential statistics) to highlight the similarity in the approaches they use and the differences in purposes.

Descriptive statistics are numerical and graphical summaries that describe the basic characteristics of the data. For example, Pearson's correlations and multi-way contingency tables describe association

between continuous variables and between categorical variables while, similarity measures such as Euclidean distance or Manhattan distance summarise likeness between observations. Although it is possible to comprehend these descriptive statistics in smaller settings, the information can quickly become difficult to synthesise with increased number of variables and observations. Unsupervised ML techniques basically explore and process further these descriptive statistics to discover hidden patterns and groupings in the data and to extract useful features from the data. The main tasks unsupervised ML are used for are dimension reduction, clustering and association rule mining. Dimension reduction techniques such as principal components analysis are frequently used to reveal hidden patterns in high-dimensional inter-related data [12]. They are used, for example, to summarise a large number of correlated bioclimatic variables or to assist visualising population structure in genetic variation [13, 14]. Specifically, matrix decomposition techniques are applied to Pearson's correlation matrix or to the raw data matrix to transform those correlated variables into a new set of uncorrelated components. The output eigenvalues and eigenvectors from the matrix decomposition and the calculated components from the analysis can then be presented in tables or figures to visualise any hidden patterns in the data. Likewise, clustering techniques utilise similarity measures from descriptive statistics to group individuals that share similar characteristics [15, 16]. Unsupervised ML can also identify interesting relations between variables and networks in large databases by exploring those frequent if-then conditional patterns in the data [17, 18]. Big data visualisation is another area of unsupervised ML that has expanded those traditional graphical techniques with enhanced computer systems to collect and process raw data, and to present the information graphically in a way that we can gain insights, for example, using a dendrogram heatmap to depict gene expression data or using cartography to visualise geospatial spread of infectious diseases [19, 20]. Figure 1 presents examples of data visualisation generated by unsupervised ML algorithms for association rules mining, clustering and dimension reduction.

Inferential statistics allow us to make inferences about the population using sample data. Statistical models make assumptions (not necessary

causal) about the data generation process and these distributional assumptions are incorporated in the parameter estimation process. Linear regression or logistic regression are two classical examples of statistical models that infer the relationships between predictors and the outcome based on estimated regression coefficients and their 95% confidence intervals. If we wish to use statistical models for prediction, 95% prediction intervals are commonly used to account for both uncertainty in estimated parameters and random variation in future observations. In contrast, the primary purpose of supervised ML techniques is making classification, diagnostics, prediction or forecasting on unknown data, and the models' performance are judged by their estimation errors. Most of the methods used in classical statistical modelling (i.e. methods that make specific assumptions about joint or conditional probability distributions of the data) can also be utilised in supervised ML, as many supervised ML methods use data directly to find generalisable predictive patterns. Supervised ML methods also exploit regularisation algorithms to reduce overfitting (models that fitted too closely to the data compromising its classification or prediction ability in unseen data) and choose optimisation algorithms to improve prediction accuracy; these mathematical processes are seldom carried out in classical statistical modelling.

Inference and prediction are complementary. The strength of statistical models is on inference, and they incorporate our knowledge of the data generation process. On the other hand, supervised ML emphasises choosing the best predictive algorithms, and the models will have various degrees of interpretability and explainability. In the next sections we will look at the scope of utilising ML in animal and veterinary public health surveillance and their specific applications.

Machine learning in animal and veterinary public health

The scope of artificial intelligence (AI) in the context of public health has recently been reviewed by Schwalbe and Wahl [23], who identified four categories of AI-driven health interventions:

1) diagnosis

2) mortality and morbidity risk assessment

3) disease outbreak prediction and surveillance

4) health policy and planning.

It is possible to identify recent contributions of ML to animal and veterinary public health that broadly fall within these categories, as well as others that would not clearly fit within any of them.

As in healthcare medical applications, signal processing methods in combination with ML can be used to enhance the performance of diagnostic or classification systems in animals or herds. Promising results have been obtained for example when convolutional neural networks were used to recognise and quantify specific lesions on digital images captured during routine slaughtering of pigs [24]. Improvements to diagnostic performance by applying ML are not limited to imaging data, classification tree analysis has been shown to be able to enhance the sensitivity of the classification regime on which the eradication programme for bovine tuberculosis in the United Kingdom (UK) is based [25]. Decision trees are a method of supervised learning that can be used for regression or classification tasks. They consist of a tree-like structure where each node represents a single input feature and, for numeric features, their split value. The final nodes after which no further splits take place are referred to as the leaves of the tree and represent the output that is used to classify or predict. Identification of the best feature and threshold value to split the data is carried out in order to generate the most homogeneous sub-nodes with respect to the outcome of the tree. Decision trees are one of the most widely used ML methods and the key component of other algorithms such as random forests.

With regard to the second domain of application, ML has been used, for example, to attempt to predict cases of lameness in dairy cows based on milk production and conformation traits [26]. The predictive performance of the classifiers built in this study was suboptimal, but as

acknowledged by the authors, it could possibly be improved by expanding the spectrum of data with which the models were trained. Indeed, this study illustrates how the capacity of ML algorithms to accurately predict presentation of a multifactorial condition, such as lameness in dairy cattle, relies on them being trained on data that captures the wide array of disease determinants. The ability of ML algorithms to generate real-time risk predictions based on a broad range of risk factors was the motivation for the use of ML to expand conventional risk prediction approaches and generate daily predictions for highly pathogenic avian influenza risk for poultry farms in the Republic of Korea [27], an application that falls within the third category of AI-driven interventions listed above.

As for applications for health policy and planning, we are not aware of the use of ML algorithms to support allocation of resources for animal disease surveillance in the same way they have been used in public health resource allocation [28]. On the other hand, ML has been used to generate information in order to support animal health surveillance planning and outbreak response. In a recent example, to address the lack of comprehensive and accurate poultry population data in the United States of America (USA), Patyk *et al.* developed an automated machine learning process to locate commercial poultry operations and predict their size and type in the USA. The authors used a supervised ML algorithm to detect poultry operations from aerial imagery [29].

In recent years, there has been a rapid expansion in the application of ML to very diverse challenges in animal health, some of which do not entirely fall within the above areas of application, which mostly refer to the use of supervised algorithms for purpose of prediction or classification. Unsupervised ML methods have been used, for example, to discover underlying structure in poultry condemnation data to uncover potential indicators for broiler chicken health and welfare surveillance (cluster detection and association rule mining) [21, 30] and to classify cattle herd types to inform control and surveillance of endemic diseases (dimension reduction) [31]. Supervised ML methods for regression/classification have also been applied to animal and veterinary public health challenges beyond the four domains identified

by Schwalbe and Wahl [23], a recent example being the identification of carnivore and bat species not recognised as reservoirs of rabies with trait profiles suggesting their capacity to be or become reservoirs [32].

An important emerging area of application of ML algorithms in the context of animal health surveillance is the analysis and extraction of information from clinical records for the purpose of syndromic surveillance [33]. Recent studies have shown the potential of applying machine learning algorithms to automate mining of free-text data in clinical and post-mortem reports; an application that can greatly facilitate the adoption of animal health syndromic surveillance [34, 35, 36]. At farm level, precision technologies are providing farmers and veterinarians with large amounts of data the analysis of which can greatly support health and production management. Machine learning algorithms are central to the analysis of such data and, as for text data, they can enable their used for the purpose of syndromic surveillance [37, 38].

In summary, due to their diversity and versatility, ML algorithms are being applied to an increasing range of tasks in animal and veterinary public health. In addition to broad domains of application analogous to those recognised in the field of global health, more specific uses of ML to address particular tasks continue to emerge. In the following section, we present examples of the application of selected machine learning algorithms to carry out specific tasks of relevance in the context of veterinary public health surveillance. Although some of the examples fall within the four categories listed above, they are not intended to mirror them.

Examples of application in animal and veterinary public health surveillance

Use of machine learning to maximise probability of pathogen detection

As described above, ML can be applied for different purposes within the context of animal and veterinary public health. In the area of surveillance it can be used to determine the likelihood of pathogen

detection. This allows researchers to prioritise samples or cases that have the highest probability of being positive, ensuring resources and laboratory capacities are focused on these samples and to assist in the design of any future programmes of surveillance. Such approaches have been applied to animal disease, but also food borne disease [39] and plant diseases [40] and make the most of the metadata associated with the biological samples or cases, such as geographical location, type/age of host, etc.

As an example, Walsh *et al.* [41] used gradient boosted trees, which is an extension of a classification tree (Table I) to determine the likelihood that avian influenza virus (AIV) will be isolated from an individual wild bird sample. Many categorical features were included in the algorithm, including sex, age, type of bird, latitude, longitude and polymerase chain reaction (PCR) cycle threshold-values (Ct-values). The dataset came from active surveillance of wild birds in the USA, with particular focus on the migratory flyways. The samples tested were cloacal and oropharyngeal swabs and in total 24,243 were available for analysis, 90% of which were used within the training dataset. An important aim of the study by Walsh *et al.* was to assess the predictability of the PCR Ct-values as a cut-off of 35 is traditionally used, as per designed for commercial poultry, and its validity was uncertain for wild bird samples. Results showed that the PCR Ct-value was indeed the most important feature in order to predict AIV isolation from a sample, followed by location (latitude and longitude) and, as the others were not influential, only these three features were used in the final model. Importantly, the results showed that AIV could be isolated from 16% of the samples that would have been classified negative via the traditional method. The findings of this study are of direct relevance to the AIV wild bird surveillance programme, which can be enhanced with this knowledge.

Use of machine learning for early detection of emerging animal and foodborne diseases

ML algorithms are central to platforms that have been developed in recent years for the purpose of early detection of emerging animal

diseases such as the platform for automated extraction of animal disease information from the web, commonly known as PADI-web, biosurveillance system [42]. This platform collects daily news articles in 16 different languages from the web and uses machine learning algorithms to classify news as relevant or not and sentences within the news based on the type of event and type of information.

Methods for automatic identification of topics in text objects such as electronic clinical records have also proved to be applicable for tracking of cases in the event of an outbreak. Following anecdotal evidence of an increase in the number of dogs seen in UK veterinary practices with an acute vomiting syndrome, application of latent Dirichlet allocation to the clinical free text component of electronic health records from veterinary practices part of a sentinel surveillance system, facilitated identification of the potential cause of the outbreak [43].

ML algorithms have been used in a study aiming to explore the potential of combining genomic and epidemiological metadata for purpose of foodborne disease surveillance. Although the study was based on simulated data, it demonstrated that, given sufficient data, ML can be used to develop predictive models of human infection that could enhance the predictive power of current early detection algorithms or help tracing the source of a foodborne outbreak [44]. The ML approach used in this study – boosted decision trees – makes use of two types of algorithms: decision trees and boosting, an ensemble method that fits sequentially a series of 'weak' learners such as decision trees producing a final strong learner as a weighted combination of the individual ones.

Use of machine learning for forecasting disease occurrence

One area where ML has been used in the field of animal health surveillance is in the development of models that make predictions about which farms are more likely to become infected with specific pathogens, based on previous case data and a set of potential risk factors. For example, ML models have been applied to porcine epidemic diarrhoea virus (PEDV) to predict future PEDV trends in Ontario, Canada [45]. In this study, three different ML approaches

(random forest, neural networks and classification trees) were used to identify climatic factors that were predictive of the size of the outbreak (small, medium, large) by training the ML algorithms on observed data from 161 farms over 171 weekly time points. The best performing approach, random forest, had a 68% accuracy based on the testing set of data. Another study on PEDV focused instead on applying ML algorithms (random forest, support vector machines, and gradient boosting machines) to weekly farm incidence data from 332 sow farms to train predictive models and to identify the key factors that are associated with PEDV occurrence on individual farms [46]. They identified the relative importance of a number of factors in determining the risk of infection, including animal movements into nearby farms, local hog density, environmental factors, and landscape structure. The best performing model showed a greater than 80% accuracy in predicting whether a farm would become infected within a one-week period, based on the testing data.

A further example is the application of ML methods to identify predictive factors for farms becoming positive for bovine tuberculosis (bTB) in different risk areas in Great Britain [47, 48]. This work used a very extensive data set of potential herd-level predictors including demographic herd characteristics and bTB-related variables, cattle movements, badger density, land class data and a range of climatic variables; these were used along with the outcome variable, which was whether a herd was a bTB incident in 2016 (out of approximately 38,000 herds). The initial data set had 141 predictors, so several steps were carried out to reduce the number of predictors to improve the speed and performance of the algorithms (classification and regression trees, multinomial logistic regression, random forest and regularised logistic regression). The best performing models had an overall accuracy of approximately 80% at predicting which farms would become bTB positive. The main output from the models was the set of key risk factors that are predictive of a farm becoming bTB positive, as these then enable the targeting of controls. There were differences in which risk factors were in the list of the ten most important variables between the ML approaches, but also which factors were common to multiple methods. Analysis of these most important factors help to

identify farm-level variables that could be used to select farms where enhanced surveillance or control measures could be targeted with important benefits for cost-effective control.

Other potential applications of machine learning in animal health surveillance

ML methods have been applied successfully to questions related to the susceptibility of hosts to disease. For example Becker *et al.* use ML to identify which bat species are potential reservoir hosts of betacoronaviruses [49]. They used host-virus data from 710 host species and 359 virus genera from GenBank integrated with a mammal phylogenetic super tree and ecological traits of bat species. An ensemble of eight models was applied to the same data set, which were either network based (4), trait-based (3) or a hybrid approach (1). The analysis showed that the ecological trait models predicted well the novel hosts and can now be used to inform surveillance and, in particular, to optimise wildlife sampling for undiscovered viruses. Another example of application of ML to generate insights into potential reservoirs of disease is the study by Wardeh *et al.* [50]. Predictions of associations between known viruses and potential reservoirs of disease (zoonotic and non-zoonotic), were obtained with an ensemble of six models using a large data set of mammal-pathogen interactions. The results highlighted that current knowledge is likely to heavily underestimate the number of existing associations, particularly in wild and semi-domesticated mammals.

An application of ML in the context of disease surveillance that deserves special mention is the exploration of genome sequencing data. The characteristics of these data (large, complex and hiding patterns that would be challenging to determine via other means) make ML methodologies ideal for their analysis. Furthermore, sequencing data is now becoming more readily available due to reduction in cost and the increase in through-put within veterinary and public health institutes. Examples of tasks relevant for design and implementation of infectious disease surveillance, which have been successfully accomplished by applying ML to whole genome sequencing data include source

attribution [51], assessment of pathogenicity [52], prediction of antibiotic resistance phenotypes [53] and prediction of clinical outcomes [54].

The above examples mostly deal with aspects of pathogen/disease detection, their patterns of distribution in the population and forecasting their future occurrence. Undoubtedly, a better understanding of pathogen/disease occurrence can inform and enhance surveillance systems for infectious diseases. However, another key element in the design and implementation of surveillance systems for animal diseases is knowledge about the structure of the animal population. To that end, it has been shown that national level databases could potentially be harvested to provide insights into livestock distribution and holding type which are critical for surveillance when up-to-date census data are not available. For example, a database containing environmental, climatic and demographic variables was utilised to develop 'species distribution models' focusing on estimating farm animal population numbers [55]. In this study, the authors built three ML models using a commercial livestock database and 22 environmental and socio-economic predictors. Boosted regression trees and random forest had comparable performance (86–92% for livestock unit prediction and 47–60% for cattle prediction) while the K-nearest neighbour model performed poorly in all simulations. While it is valuable to develop supervised ML models that predict livestock distributions using environmental and socio-economic predictors, it is not always possible to develop such prediction algorithms when the 'ground truth' is not available. K-mean cluster analysis, an unsupervised ML technique, has been used to classify holdings into different groups based on a set of proxy indicators [56]. External validation of the clustering results (matched to other data sources on pig population data where the true holding type was known) indicated it had good agreement (79%) and was comparable to the agreement of clustering based on a manual model of 78% (using expert opinion to develop rules that classify holdings based on movement characteristics). The authors concluded that although the manual model provided accurate results, it was time consuming, and the developed unsupervised ML model example can provide useful degree of accuracy for surveillance and risk assessment.

In Table I, we provide a brief description of methods commonly used in the field of animal and veterinary public health that have been referred to in this section. The table also provides introductory papers for further detailed reading on each method. To guide those readers interested in exploring further the use of ML algorithms to accomplish specific tasks, Table II provides examples and references of application of specific ML methods.

Practical recommendations

Consider training a meta-model

Each ML method will have different strengths and weaknesses, and it is difficult to know *a priori* which approach will work best for any given problem. Users can also apply a stacking ensemble approach, where multiple methods are applied in parallel and the final model is a weighted combination of the predictions from all the models.

Consider the transparency of the approach

A disadvantage of ML algorithms when compared to statistical methods such as regression is the limited 'interpretable' information they provide beyond their immediate task (e.g. classifying observations). For example, neural networks have been found to be very effective at making predictions where there are complex non-linear relationships between variables, but they might be unsuitable for identifying individual risk factors from which to target farms for surveillance or control measures.

Consider whether the main objective is to explain or to predict

As the main focus of ML algorithms is on prediction rather than explanation, there can be differences in the variables that are included in the final predictive model between ML and classical statistics. While explanation is not the primary aim of ML methods, some of the factors that are found to be important for prediction by ML algorithms could be the target of further investigation and could shed light of causative explanations, even where they were not found to be statistically

significant by classical statistics approaches. If the intention is to build a ML model in order to predict disease occurrence, the algorithm should ideally be trained on data that capture the array of disease determinants. This is particularly important when trying to predict the occurrence of multifactorial conditions.

Consider the balance of domain expertise and machine learning expertise

For example, Sperschneider [40] suggests that 5% of the time will be spent training the model but 95% selecting the most appropriate features, which needs biological/epidemiological expertise. Likewise, ML expertise is needed to ensure that the correct model is applied for the available data, and that overfitting is avoided.

Conclusions

Surveillance in animal and veterinary public health involves diverse interconnected tasks in pursue of the broad objectives of enabling early detection of animal and zoonotic disease threats, providing assurance of freedom from hazards in animals and their products and evaluating controls for endemic diseases. The rapid expansion of ML in recent years has seen their application to many different tasks in relation to animal and veterinary public health surveillance. In some cases, ML algorithms are used to achieve tasks that only became relevant and within reach with the advent of big data, the development of new methods for data analysis and increased computing power. But ML methods are also being used to address technical needs of surveillance systems that have usually being answered by traditional statistical data analysis, such as the identification of samples, animals or herds to be targeted based on *a priori* risk. Both groups of applications are expanding rapidly, and while many are still limited to case studies illustrating their potential, it can be expected that in the coming years they become increasingly embedded into national and international surveillance systems. Despite their potential to enhance surveillance efforts in animal and veterinary public health, ML algorithms are not a replacement for traditional statistical analysis. Their successful integration as part of surveillance systems should take into

consideration the value of combining multiple methods for the same or similar tasks and the relative importance of transparency vs. (sometimes moderate) increases in predictive performance. For this, an adequate balance between domain expertise and ML expertise is essential.

References

[1] Badillo S., Banfai B., Birzele F., Davydov I.I., Hutchinson L., Kam-Thong T., Siebourg-Polster J., Steiert B. & Zhang J.D. (2020). – An introduction to machine learning. *Clin. Pharmacol. Ther.*, **107** (4), 871–885. https://doi.org/10.1002/cpt.1796

[2] Roth J.A., Battegay M., Juchler F., Vogt J.E. & Widmer A.F. (2018). – Introduction to machine learning in digital healthcare epidemiology. *Infect. Control Hosp. Epidemiol.*, **39** (12), 1457–1462. https://doi.org/10.1017/ice.2018.265

[3] Sarker I.H. (2021). – Machine learning: algorithms, real-world applications and research directions. *SN Comput. Sci.*, **2** (3), 160. https://doi.org/10.1007/s42979-021-00592-x

[4] Lo Vercio L., Amador K. […] & Forkert N.D. (2020). – Supervised machine learning tools: a tutorial for clinicians. *J. Neural Eng.*, **17** (6), 062001. https://doi.org/10.1088/1741-2552/abbff2

[5] Zhou Y.-H. & Gallins P. (2019). – A review and tutorial of machine learning methods for microbiome host trait prediction. *Front. Genet.*, **10**, 579. https://doi.org/10.3389/fgene.2019.00579

[6] Scott I.A. (2021). – Demystifying machine learning: a primer for physicians. *Intern. Med. J.*, **51** (9), 1388–1400. https://doi.org/10.1111/imj.15200

[7] Wang X., Bouzembrak Y., Oude Lansink A.G.J.M. & van der Fels-Klerx H.J. (2022). – Application of machine learning to the monitoring and prediction of food safety: a review. *Compr. Rev. Food Sci. Food Saf.*, **21** (1), 416–434. https://doi.org/10.1111/1541-4337.12868

[8] Wang M.W.H., Goodman J.M. & Allen T.E.H. (2021). – Machine learning in predictive toxicology: recent applications and future directions for classification models. *Chem. Res. Toxicol.*, **34** (2), 217–239. https://doi.org/10.1021/acs.chemrestox.0c00316

[9] Bollig N., DeBoer D. & Döpfer D. (2020). – A machine learning tutorial for veterinarians: examples using canine atopic dermatitis. Presentation to the Virtual Talbot Veterinary Informatics Symposium, 19 pp. Available at: https://avinformatics.org/resources/Documents/Proceedings/Nathan%20Bollig%20Article%20Talbot2020.pdf (accessed on 5 April 2023).

[10] Monaco A., Pantaleo E., Amoroso N., Lacalamita A., Giudice C.L., Fonzino A., Fosso B., Picardi E., Tangaro S., Pesole G. & Bellotti R. (2021). – A primer on machine learning techniques for genomic applications. *Comput. Struct. Biotechnol. J.*, **19**, 4345–4359. https://doi.org/10.1016/j.csbj.2021.07.021

[11] Tarca A.L., Carey V.J., Chen X.-W., Romero R. & Drăghici S. (2007). – Machine learning and its applications to biology. *PLoS Comput. Biol.*, **3** (6), e116. https://doi.org/10.1371/journal.pcbi.0030116

[12] Lever J., Krzywinski M. & Altman N. (2017). – Principal component analysis. *Nat. Methods*, **14** (7), 641–642. https://doi.org/10.1038/nmeth.4346

[13] Ausmees K. & Nettelblad C. (2022). – A deep learning framework for characterization of genotype data. *G3 (Bethesda)*, **12** (3), jkac020. https://doi.org/10.1093/g3journal/jkac020

[14] Ferro C., López M., Fuya P., Lugo L., Cordovez J.M. & González C. (2015). – Spatial distribution of sand fly vectors and eco-epidemiology of cutaneous leishmaniasis transmission in Colombia. *PLoS One*, **10** (10), e0139391. https://doi.org/10.1371/journal.pone.0139391

[15] Altman N. & Krzywinski M. (2017). – Clustering. *Nat. Methods*, **14** (6), 545–546. https://doi.org/10.1038/nmeth.4299

[16] Warns-Petit E., Morignat E., Artois M. & Calavas D. (2010). – Unsupervised clustering of wildlife necropsy data for syndromic surveillance. *BMC Vet. Res.*, **6**, 56. https://doi.org/10.1186/1746-6148-6-56

[17] Hernández B., Reilly R.B. & Kenny R.A. (2019). – Investigation of multimorbidity and prevalent disease combinations in older Irish adults using network analysis and association rules. *Sci. Rep.*, **9** (1), 14567. https://doi.org/10.1038/s41598-019-51135-7

[18] Nahar J., Imam T., Tickle K.S. & Chen Y.-P.P. (2013). – Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Syst. Appl.*, **40** (4), 1086–1093. https://doi.org/10.1016/j.eswa.2012.08.028

[19] Raddatz B.B., Spitzbarth I., Matheis K.A., Kalkuhl A., Deschl U., Baumgärtner W. & Ulrich R. (2017). – Microarray-based gene expression analysis for veterinary pathologists: a review. *Vet. Pathol.*, **54** (5), 734–755. https://doi.org/10.1177/0300985817709887

[20] Kraemer M.U.G., Hay S.I., Pigott D.M., Smith D.L., Wint G.R.W. & Golding N. (2016). – Progress and challenges in infectious disease cartography. *Trends Parasitol.*, **32** (1), 19–29. https://doi.org/10.1016/j.pt.2015.09.006

[21] Buzdugan S.N., Chang Y.M., Huntington B., Rushton J., Guitian J., Alarcon P. & Blake D.P. (2020). – Identification of production chain risk factors for slaughterhouse condemnation of broiler chickens'. *Prev. Vet. Med.*, **181**, 105036. https://doi.org/10.1016/j.prevetmed.2020.105036

[22] Crotta M., Prakashbabu B.C., Holt H., Swift B., Kaur P., Bedi J.S., Pedada V.C., Shaik T.B., Tumati S.R. & Guitian J. (2022). – Microbiological risk ranking of foodborne pathogens and food products in scarce-data settings. *medRxiv*, 2022.04.07.22273592. https://doi.org/10.1101/2022.04.07.22273592

[23] Schwalbe N. & Wahl B. (2020). – Artificial intelligence and the future of global health. *Lancet*, **395** (10236), 1579–1586. https://doi.org/10.1016/S0140-6736(20)30226-9

[24] Bonicelli L., Trachtman A.R., Rosamilia A., Liuzzo G., Hattab J., Alcaraz E.M., Del Negro E., Vincenzi S., Capobianco Dondona A., Calderara S. & Marruchella G. (2021). – Training convolutional neural networks to score pneumonia in slaughtered pigs. *Animals*, **11** (11), 3290. https://doi.org/10.3390/ani11113290

[25] Romero M.P., Chang Y.-M., Brunton L.A., Parry J., Prosser A., Upton P. & Drewe J.A. (2022). – Machine learning classification methods informing the management of inconclusive reactors at bovine tuberculosis surveillance tests in England. *Prev. Vet. Med.*, **199**, 105565. https://doi.org/10.1016/j.prevetmed.2021.105565

[26] Shahinfar S., Khansefid M., Haile-Mariam M. & Pryce J.E. (2021). – Machine learning approaches for the prediction of lameness in dairy cows. *Animal*, **15** (11), 100391. https://doi.org/10.1016/j.animal.2021.100391

[27] Yoo D.-S., Song Y.-H., Choi D.-W., Lim J.-S., Lee K. & Kang T. (2021). – Machine learning-driven dynamic risk prediction for highly pathogenic avian influenza at poultry farms in Republic of Korea: daily risk estimation for individual premises. *Transbound. Emerg. Dis.*, **69** (5), 2667–2681. https://doi.org/10.1111/tbed.14419

[28] Araújo Rosas M., Benjamin Bezerra A.F. & Duarte-Neto P.J. (2013). – Use of artificial neural networks in applying methodology for allocating health resources. *Rev. Saúde Pública*, **47** (1), 8 pp. https://doi.org/10.1590/s0034-89102013000100017

[29] Patyk K.A., McCool-Eye M.J., South D.D., Burdett C.L., Maroney S.A., Fox A., Kuiper G. & Magzamen S. (2020). – Modelling the domestic poultry population in the United States: a novel approach leveraging remote sensing and synthetic data methods. *Geospat. Health*, **15** (2), 913. https://doi.org/10.4081/gh.2020.913

[30] Buzdugan S.N., Alarcon P., Huntington B., Rushton J., Blake D.P. & Guitian J. (2021). – Enhancing the value of meat inspection records for broiler health and welfare surveillance: longitudinal detection of relational patterns. *BMC Vet. Res.*, **17** (1), 278. https://doi.org/10.1186/s12917-021-02970-2

[31] Brock J., Lange M., Tratalos J.A., More S.J., Graham D.A., Guelbenzu-Gonzalo M. & Thulke H.-H. (2021). – Combining expert knowledge and machine-learning to classify herd types in livestock systems. *Sci. Rep.*, **11** (1), 2989. https://doi.org/10.1038/s41598-021-82373-3

[32] Worsley-Tonks K.E.L., Escobar L.E., Biek R., Castaneda-Guzman M., Craft M.E., Streicker D.G., White L.A. & Fountain-Jones N.M. (2020). – Using host traits to predict reservoir host species of rabies virus. *PLoS Negl. Trop. Dis.*, **14** (12), e0008940. https://doi.org/10.1371/journal.pntd.0008940

[33] Dórea F.C., Sanchez J. & Revie C.W. (2011). – Veterinary syndromic surveillance: current initiatives and potential for development. *Prev. Vet. Med.*, **101** (1–2), 1–17. https://doi.org/10.1016/j.prevetmed.2011.05.004

[34] Anholt R.M., Berezowski J., Jamal I., Ribble C. & Stephen C. (2014). – Mining free-text medical records for companion animal enteric syndrome surveillance. *Prev. Vet. Med.*, **113** (4), 417–422. https://doi.org/10.1016/j.prevetmed.2014.01.017

[35] Arguello-Casteleiro M., Jones P.H., Robertson S., Irvine R.M., Twomey F. & Nenadic G. (2019). – Exploring the automatisation of animal health surveillance through natural language processing. *In* Artificial Intelligence XXXVI (M. Bramer & M. Petridis, eds), Vol. 11927. Springer, Cham, Switzerland, 213–226. https://doi.org/10.1007/978-3-030-34885-4_17

[36] Bollig N., Clarke L., Elsmo E. & Craven M. (2020). – Machine learning for syndromic surveillance using veterinary necropsy reports. *PLoS One*, **15** (2), e0228105. https://doi.org/10.1371/journal.pone.0228105

[37] Giordano J.O., Sitko E.M., Rial C., Pérez M.M. & Granados G.E. (2022). – Symposium review: use of multiple biological, management, and performance data for the design of targeted reproductive management strategies for dairy cows. *J. Dairy Sci.*, **105** (5), 4669–4678. https://doi.org/10.3168/jds.2021-21476

[38] Sturm V., Efrosinin D., Öhlschuster M., Gusterer E., Drillich M. & Iwersen M. (2020). – Combination of sensor data and health monitoring for early detection of subclinical ketosis in dairy cows. *Sensors (Basel)*, **20** (5), 1484. https://doi.org/10.3390/s20051484

[39] Wang H., Cui W., Guo Y., Du Y. & Zhou Y. (2021). – Machine learning prediction of foodborne disease pathogens: algorithm development and validation study. *JMIR Med. Inform.*, **9** (1), e24924. https://doi.org/10.2196/24924

[40] Sperschneider J. (2020). – Machine learning in plant–pathogen interactions: empowering biological predictions from field scale to genome scale. *New Phytol.*, **228** (1), 35–41. https://doi.org/10.1111/nph.15771

[41] Walsh D.P., Ma T.F., Ip H.S. & Zhu J. (2019). – Artificial intelligence and avian influenza: using machine learning to enhance active surveillance for avian influenza viruses. *Transbound. Emerg. Dis.*, **66** (6), 2537–2545. https://doi.org/10.1111/tbed.13318

[42] Valentin S., Arsevska E., Rabatel J., Falala S., Mercier A., Lancelot R. & Roche M. (2021). – PADI-web 3.0: a new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. *One Health*, **13**, 100357. https://doi.org/10.1016/j.onehlt.2021.100357

[43] Mäntylä Noble P.-J., Appleton C., Radford A.D. & Nenadic G. (2021). – Using topic modelling for unsupervised annotation of electronic health records to identify an outbreak of disease in UK dogs. *PLoS One*, **16** (12), e0260402. https://doi.org/10.1371/journal.pone.0260402

[44] Hill A.A., Crotta M., Wall B., Good L., O'Brien S.J. & Guitian J. (2017). – Towards an integrated food safety surveillance system: a simulation study to explore the potential of combining genomic and epidemiological metadata. *R. Soc. Open Sci.*, **4** (3), 160721. https://doi.org/10.1098/rsos.160721

[45] Ajayi T., Dara R. & Poljak Z. (2019). – Forecasting herd-level porcine epidemic diarrhea (PED) frequency trends in Ontario (Canada). *Prev. Vet. Med.*, **164**, 15–22. https://doi.org/10.1016/j.prevetmed.2019.01.005

[46] Machado G., Vilalta C., Recamonde-Mendoza M., Corzo C., Torremorell M., Perez A. & VanderWaal K. (2019). – Identifying outbreaks of porcine epidemic diarrhea virus through animal movements and spatial neighborhoods. *Sci. Rep.*, **9** (1), 457. https://doi.org/10.1038/s41598-018-36934-8

[47] Romero M.P., Chang Y.-M., Brunton L.A., Prosser A., Upton P., Rees E., Tearne O., Arnold M., Stevens K. & Drewe J.A. (2021). – A comparison of the value of two machine learning predictive models to support bovine tuberculosis disease control in England. *Prev. Vet. Med.*, **188**, 105264. https://doi.org/10.1016/j.prevetmed.2021.105264

[48] Romero M.P., Chang Y.-M., Brunton L.A., Parry J., Prosser A., Upton P., Rees E., Tearne O., Arnold M., Stevens K. & Drewe J.A. (2020). – Decision tree machine learning applied to bovine tuberculosis risk factors to aid disease control decision making. *Prev. Vet. Med.*, **175**, 104860. https://doi.org/10.1016/j.prevetmed.2019.104860

[49] Becker D.J., Albery G.F. […] & Carlson C.J. (2022). – Optimising predictive models to prioritise viral discovery in zoonotic reservoirs. *Lancet Microbe*, **3** (8), e625–e637. https://doi.org/10.1016/S2666-5247(21)00245-7

[50] Wardeh M., Blagrove M.S.C., Sharkey K.J. & Baylis M. (2021). – Divide-and-conquer: machine-learning integrates mammalian and viral traits with network features to predict virus-mammal associations. *Nat. Commun.*, **12** (1), 3954. https://doi.org/10.1038/s41467-021-24085-w

[51] Munck N., Njage P.M.K., Leekitcharoenphon P., Litrup E. & Hald T. (2020). – Application of whole-genome sequences and machine learning in source attribution of Salmonella Typhimurium. *Risk Anal.*, **40** (9), 1693–1705. https://doi.org/10.1111/risa.13510

[52] Deneke C., Rentzsch R. & Renard B.Y. (2017). – PaPrBaG: a machine learning approach for the detection of novel pathogens from NGS data. *Sci. Rep.*, **7** (1), 39194. https://doi.org/10.1038/srep39194

[53] Sunuwar J. & Azad R.K. (2021). – A machine learning framework to predict antibiotic resistance traits and yet unknown genes underlying resistance to specific antibiotics in bacterial strains. *Brief. Bioinform.*, **22** (6), bbab179. https://doi.org/10.1093/bib/bbab179

[54] Njage P.M.K., Leekitcharoenphon P. & Hald T. (2019). – Improving hazard characterization in microbial risk assessment using next generation sequencing data and machine learning: predicting clinical outcomes in shigatoxigenic *Escherichia coli*. *Int. J. Food Microbiol.*, **292**, 72–82. https://doi.org/10.1016/j.ijfoodmicro.2018.11.016

[55] Hollings T., Robinson A., van Andel M., Jewell C. & Burgman M. (2017). – Species distribution models: a comparison of statistical approaches for livestock and disease epidemics. *PLoS One*, **12** (8), e0183626. https://doi.org/10.1371/journal.pone.0183626

[56] Smith R.P., Gavin C., Gilson D., Simons R.R.L. & Williamson S. (2020). – Determining pig holding type from British movement data using analytical and machine learning approaches. *Prev. Vet. Med.*, **178**, 104984. https://doi.org/10.1016/j.prevetmed.2020.104984

[57] Krzywinski M. & Altman N. (2017). – Classification and regression trees. *Nat. Methods*, **14** (8), 757–758. https://doi.org/10.1038/nmeth.4370

[58] Altman N. & Krzywinski M. (2017). – Ensemble methods: bagging and random forests. *Nat. Methods*, **14** (10), 933–934. https://doi.org/10.1038/nmeth.4438

[59] Natekin A. & Knoll A. (2013). – Gradient boosting machines, a tutorial. *Front. Neurorobot.*, **7**, 21. https://doi.org/10.3389/fnbot.2013.00021

[60] Bzdok D., Krzywinski M. & Altman N. (2018). – Machine learning: supervised methods. *Nat. Methods*, **15** (1), 5–6. https://doi.org/10.1038/nmeth.4551

[61] Lever J., Krzywinski M. & Altman N. (2016). – Regularization. *Nat. Methods*, **13** (10), 803–804. https://doi.org/10.1038/nmeth.4014

[62] Mehlig B. (2021). – Introduction. *In* Machine learning with neural networks: an introduction for scientists and engineers. Cambridge University Press, Cambridge, United Kingdom, 1–12. https://doi.org/10.1017/9781108860604.001

[63] Asmussen C.B. & Møller C. (2019). – Smart literature review: a practical topic modelling approach to exploratory literature review. *J. Big Data*, **6** (1), 93. https://doi.org/10.1186/s40537-019-0255-7

[64] Franceschini S., Grelet C., Leblois J., Gengler N., Soyeurt H. & GplusE consortium (2022). – Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health? *J. Dairy Sci.*, **105** (8), 6760–6772. https://doi.org/10.3168/jds.2022-21975

Table I

Selected machine learning methods applied to address tasks in the context of animal and veterinary public health surveillance: description of the method and suggested reference for further details

Machine learning method	Description	Reference
Classification and regression tree	A decision tree approach that explains how a target variable's values can be predicted based on other values, where the target variable may be categorical (classification tree) or a quantitative (regression tree)	[57]
Random forest	An ensemble of decision trees is generated from different bootstrap samples before individual bootstrap predictions are combined into a consensus estimate (bagging approach)	[58]
Gradient boosting machine	A decision tree approach that can be used for classification and regression problems. The algorithm iteratively improves the predictive power by refining its performance in areas it is performing poorly. It differs from random forest in that it sequentially improves each individual tree (boosting) rather than building an ensemble of a whole set of trees	[59]
Support vector machine	Supervised machine learning method for classification, regression and outlier detection. It works out how best to create rules that separate items into different classes. It does this by finding the features that maximise the distance between the classes	[60]
Regularised regression	A regression approach which penalises models with more parameters in order to prevent overfitting	[61]
Artificial neural networks	Networks of interconnected nodes grouped in layers and including an input layer, an output layer and one or more hidden layers. The nodes (or neurons) receive signals from antecedent nodes and may forward signals to subsequent nodes. The existence of hidden layers implies that in contrast to other (shallow) machine learning algorithms, in artificial neural networks inputs and outputs are not directly connected	[62]

Convolutional neural networks	Artificial neural networks with layers (convolutional layers) where the input is transformed before it is passed to the next layer through the application of a filter. Originally developed to process pixel data that could otherwise result in a vast number of nodes in a conventional artificial neural network. They are mostly applied to analyse visual imagery	[24]
Recurrent neural networks	Artificial neural networks that allow the output of previous steps to be fed as input to the current step. While classical (feed-forward) artificial neural networks allow signal to travel only from input to output, recurrent neural networks include feedback loops allowing outputs from some nodes to affect input into the same nodes. This feature of recurrent neural networks allows them to display temporal behaviour and capture sequential data such as text or time series	[36]
Latent Dirichlet allocation	A probabilistic Bayesian model frequently used for unsupervised classification of documents into topics, which are discovered based on the co-occurrence of individual terms. Documents are allocated to specific topics based on how relevant they are to them	[63]
K-means cluster	An unsupervised algorithm that groups observations into *k* clusters by minimising the total sum of squared distances between observations and cluster centres	[15]
K-nearest neighbour	A supervised method for classification based on finding the k-nearest neighbours in a reference training set and assigning the class of a new observation based on a 'majority vote' among the k neighbours	[60]
Ward's agglomerative hierarchical clustering	An unsupervised method aimed at building a hierarchy of clusters. The algorithm starts considering each observation as a cluster and proceeds by combining pairs of clusters at each iteration. The pairs of clusters to be merged at each step are identified in order to minimise total within cluster variance	[64]

Table II

Selected tasks achieved by applying machine learning methods in the context of animal and veterinary public health surveillance

Task	Machine learning method applied	Data requirements	Considerations	Reference
Identify underlying structure in large volumes of abattoir inspection data	Association rule mining	Features (e.g. reasons for condemnation in abattoir) for which co-occurrence is of interest	The number of possible association rules can be vast, requiring the specification of constraints for rule selection and pruning of redundant rules	[21, 30]
Identify factors that are predictive of farms becoming infected with a particular pathogen	CART, random forest, regularised regression, support vector machines, gradient boosting machines	Infection history data from a number of farms along with data on the potential risk factors	Transparency as need to know each factor and its importance so farms can be targeted for surveillance/control	[25, 46, 48]
Prediction of future outbreak size category (e.g. zero, small, medium, large)	Random forest, CART, neural networks	Previous incidence data for the area of interest and data on the potential factors; influencing outbreak size	Transparency not critical for making predictions of outbreak size, but would be beneficial if you also want to know which farms are more likely to become infected	[45]
To predict the likelihood of a pathogen being detected from a sample	Gradient boosting machine	Viral RNA from avian cloacal and oropharyngeal swab samples	Although probability of prediction was slightly higher if the variables age, sex and bird-type were included a more parsimonious model was selected	[41]

To better understand the structure and composition of the livestock population	Self-organising feature map, k-means clustering	Demographic information of the cattle herd and transport statistics; registered pig movement data		[31, 56]
Farm-level assessment of animal health-status	Ward's agglomerative hierarchical clustering	Individual-animal level health and production data and biomarkers	Can be very computationally intensive	[64]

CART: classification and regression trees

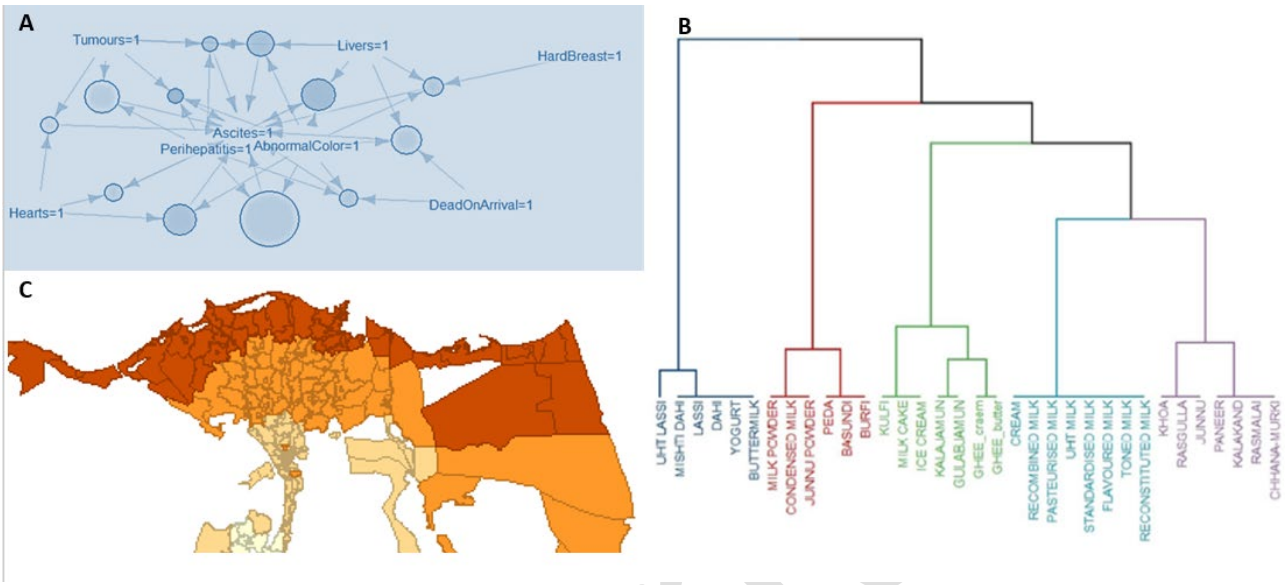

Figure 1

Data visualisation examples generated by unsupervised machine learning algorithms for association rules mining, clustering and dimension reduction

(A) Networks of associated morbidities in 18 rules generated by association rules mining analysis using poultry condemnation data [21]

(B) Dendrogram showing hierarchical clustering of dairy products based on product characteristics favouring or preventing microbial growth/survival [22]

(C) Choropleth map of districts of Northern Egypt displaying values of a single principal component explaining 52% of the variation among 18 bioclimatic variables