REVIEW ARTICLE

ADVANCING VETERINARY EPIDEMIOLOGY BY INTEGRATION OF MACHINE LEARNING: CURRENT STATUS AND FUTURE PERSPECTIVES

Abdullah Muftić[1]*, Nihad Fejzić[1]

[1]University of Sarajevo – Veterinary Faculty, Department of Pathobiology and Epidemiology, Sarajevo, Bosnia and Herzegovina

***Corresponding author**: Abdullah Muftić, DVM, Senior Assistant

Address: Zmaja od Bosne 90, 71 000 Sarajevo, Bosnia and Herzegovina
Phone: +38761439209
ORCID: 0000-0003-0404-0002
E-mail: abdullah.muftic@vfs.unsa.ba

Original Submission:
18 June 2024
Revised Submission:
05 July 2014
Accepted:
19 July 2024

How to cite this article: Muftić A, Fejzić N. 2024. Advancing veterinary epidemiology by integration of machine learning: Current status and future perspectives. Veterinaria, 73(2), 93-104.

ABSTRACT

The integration of machine learning (ML) in veterinary epidemiology offers transformative potential for data analysis and disease management, a significant shift from traditional statistical methods. This review explores the burgeoning role of ML, emphasizing its capacity to handle complex, high-dimensional data and uncover nonlinear relationships, which are pivotal in epidemiology. Key ML methodologies, including supervised, unsupervised, and reinforcement learning, provide robust frameworks for predictive modeling, pattern recognition, and decision-making processes. Applications in veterinary medicine are already evident in diagnostic imaging and animal behavior monitoring, showcasing ML's ability to enhance diagnostic accuracy and welfare monitoring.

Despite these advancements, the field faces challenges such as imbalanced datasets, data quality issues, and the need for interdisciplinary collaboration. Strategies like Synthetic Minority Over-sampling Technique and ensemble methods help address class imbalance, while robust preprocessing techniques mitigate data noise. Future advancements in natural language processing and reinforced learning promise further integration, optimizing disease surveillance and intervention strategies.

The review highlights the transformative potential of ML in veterinary epidemiology, advocating for continued research and collaboration to overcome existing hurdles. By leveraging ML's capabilities, veterinary professionals can improve disease prediction, develop targeted preventive programs, and enhance overall animal health and food security, marking a significant advancement in veterinary science.

Keywords: Animal health, artificial intelligence, disease surveillance, predictive modeling, veterinary medicine

doi.org/10.51607/22331360.2024.73.2.93

INTRODUCTION

Historically, data analysis and inferences made by applying analytical and statistical methods were epidemiological research's predominant roles and outcomes. However, the fast and rapid development of data science and artificial intelligence (AI) tools seen in the past couple of years, mainly in the domain of integration of powerful machine analysis of big data aggregation, opened a new window of opportunities for all biomedical disciplines including epidemiology. Although there has been increasing evidence of the use of machine learning (ML) as one of the growing tools under the AI umbrella in veterinary medicine, significant clinical and research potential and opportunities involving ML remain largely untapped, putting veterinary science behind other biomedical research fields.

This review does not attempt to offer an exhaustive elaboration of examples and progress in the integration of these new tools in the analysis of data collected through traditional surveillance approaches. Instead, it offers a glimpse into basic terms, principles, and current ML practices and their potential use within epidemiology and broader scope of veterinary medicine issues. Several comprehensive review articles delve deeper into AI and ML strategies in human medicine and data science. However, we recognised the need to explore the current status of integration of AI methods in the analysis of biological data in the animal domain, which might be very complex considering relations in the traditional epidemiological triangle of agent – host – environment. We were aware of the fact that certain levels of scepticism and ignorance exist yet in the scientific and professional community related to fast and progressive game-changing process we face today with AI and ML. This review is aimed at unpacking potential of ML methods in integration with veterinary medicine, and more specifically, with its data analysis discipline – epidemiology. We believe this can serve to research community in better understanding of the ongoing and future processes by exploring the transformative potential of ML in veterinary epidemiology. To prove this, the aim of this manuscript was to elaborate and highlight the advantages, challenges, and prospects of integrating advanced ML techniques into this field, offering a basic comparison with the most common traditional statistical methods.

Statistics and Machine Learning in the Domain of Epidemiological Researches

Frequentist inference is the predominant tool in epidemiological research. These methods, which might be complex and hardly understandable for non-statisticians, are grounded in hypothesis testing and probability calculations to support or refuse research thesis (Rothman et al., 2008). Common practices include basic statistical tests and multivariable regression models that are pivotal in identifying associations or treatment effects between variables. In epidemiology, traditional regression models serve dual functions: they are utilized either to predict a dependent variable from multiple independent variables (predictive models) or to measure the effect or association of specific independent variables on a dependent variable (explanatory models). These modeling strategies provide essential insights for both researchers and decision makers, and rely on several data assumptions such as linearity, absence of multicollinearity, and proportional risks/odds/hazards over time, familiar to epidemiologists (Vittinghoff et al., 2012). Data challenges stemming from inconsistencies, imbalances, or weak connections can impede model performance (Molenberghs and Kenward, 2007). To address these challenges, incorporating advanced data analysis has significant advantages, allowing the model to autonomously learn from the data's structure and features. This approach prioritizes predictive capabilities over explanatory or causal inference, showing potential for enhancing disease management strategies. ML is particularly suited for this role, offering distinct advantages in predictive accuracy (James et al., 2013).

ML is a subset of AI dedicated to developing algorithms that enable computers to glean

insights and make predictions or decisions from data. While ML and statistics are interconnected disciplines centered around data analysis, they diverge significantly in their methodologies, objectives, and focal points. ML aims to develop algorithms capable of identifying patterns and making informed predictions or decisions, emphasizing predictive precision and the ability to generalize. This contrasts with conventional programming paradigms, where tasks are defined by explicit instructions, and ML algorithms autonomously discern and refine patterns from data, adapting their performance iteratively to diverse scenarios (Hastie et al., 2009). Conversely, statistics is focused on the comprehension and interpretation of underlying data structures and relationships, primarily for inference, hypothesis testing, and parameter estimation (Casella and Berger, 2002). ML algorithms are designed to optimize performance across specific tasks such as classification, regression, clustering, or reinforcement learning, whereas statistical methodologies aim to draw population inferences from sample data and estimate probability distribution parameters. ML techniques emphasize algorithmic intricacy and scalability, facilitating the handling of diverse data types through feature engineering and preprocessing mechanisms, contrasting with statistics, which predominantly deals with structured datasets, drawing upon specific distributional assumptions, and emphasizing theoretical rigor and interpretive insights (Bishop, 2006).

Over the past decade, ML techniques have gathered considerable attention within biomedical research, although their integration into population health remains somewhat limited. ML methods provide a flexible alternative to traditional statistical approaches by handling complex, high-dimensional data and capturing nonlinear relationships prevalent in epidemiological studies. Unlike conventional regression models, ML models can adaptively learn from the data, proving invaluable when data structures are intricate or unknown. Despite their advantages, ML models also present several challenges. They often require large datasets to perform effectively, which may not always be available in epidemiological research, especially in the application of active data collection systems/surveillance with limited resources. Moreover, the interpretability of ML models can be problematic. Unlike traditional statistical methods that offer clear and interpretable coefficients, ML models can be perceived as "black boxes" that provide limited insights into the underlying mechanisms driving the predictions. This lack of transparency can hinder their acceptance in epidemiology, where understanding causality and mechanisms is crucial.

Nevertheless, the advantages of ML over traditional statistical methods in epidemiological research are substantial. ML excels in processing large volumes of data swiftly, a critical capability in a field where datasets can be extensive and heterogeneous. By automating feature extraction and model selection, ML reduces the manual effort and time typically required for complex data preprocessing and model building. This efficiency not only accelerates the pace of research but also conserves valuable resources, allowing epidemiologists to focus more on interpreting results and translating findings into actionable insights. Furthermore, ML's ability to handle nonlinear relationships and complex interactions among variables enhances its predictive accuracy compared to traditional regression models, offering promising avenues for improving disease management strategies and public health interventions. These advantages underscore ML's transformative potential in advancing epidemiological research beyond the constraints of traditional statistical methodologies.

Exploring the Triad of ML Methodologies

While exhibiting similarities, ML introduces terminology distinct from that commonly encountered in the realm of statistics. A knowledge of these unique terms is imperative for the effective utilization and interpretation of ML algorithms and models. As presented by Wiemken and Kelley (2019), the terms are explained in Table 1.

Table 1 Comparison of terminology between biostatistics/epidemiology and machine learning

Term in biostatistics and epidemiology	Term in Machine Learning	Explanation
Dependent variable	Label/class	Main factor of interest in a study
Independent variable	Feature	Factors that are thought to influence the dependent variable
Contingency table	Confusion matrix	Displays the counts of true positive, true negative, false positive, and false negative outcomes of a classification algorithm
Number of variables	Dimensions	Refers to the number of measurable properties or characteristics of each data point in a dataset (e.g. age, weight, breed, and vaccination status of an animal)
Sensitivity	Recall	Measures the proportion of true positives (animals correctly identified as having the disease) among all animals that actually have the disease
Positive predictive value	Precision	Measures the proportion of true positives among all positive test results
Outcome group with the highest frequency	Majority class	Refers to the most prevalent category within the dataset (e.g. the most common diagnostic test result observed in a population of animals)
Outcome group with the lowest frequency	Minority class	Refers to the least prevalent category within the dataset (e.g. rare condition or diseases that occur infrequently observed in a population of animals)
Proportion of cases in each category of the outcome variable (when outcome is categorical)	Class balance	Distribution of different categories or classes within the outcome variable.

ML encompasses three primary methodologies: supervised learning, unsupervised learning, and reinforcement learning. Each of these approaches has distinct methods for handling data and is tailored to achieve different objectives, yet they can be integrated to address complex datasets comprehensively. The compiled brief explanations and prominent examples of essential ML methods obtained from relevant scientific literature are provided in Table 2.

Table 2 Overview of supervised, unsupervised, and reinforcement learning techniques with corresponding algorithms and references.

Methodology	Explanation	Example algorithms	References
Supervised Learning	Involves training a model on a labeled dataset, where the input data is paired with the correct output. The goal is for the model to learn a mapping from inputs to outputs, allowing it to make predictions on new, unseen data.	Linear Regression, Support Vector Machines, Neural Networks, Random Forests	Goodfellow et al. (2016), Kotsiantis et al. (2007), Bishop (2006)
Unsupervised Learning	Involves training a model on data that has no labels or predefined outcomes. The goal is to discover underlying patterns or structures within the data.	K-means Clustering, Hierarchical Clustering, Principal Component Analysis, t-Distributed Stochastic Neighbor Embedding	Murphy (2012), Jain (2010), Hastie et al. (2009)
Reinforcement Learning	Involves training a model through interactions with an environment, where the model receives feedback in the form of rewards or penalties. The goal is to learn a policy that maximizes cumulative rewards over time.	Q-Learning, Deep Q-Networks, Policy Gradient Methods, Actor-Critic Methods	Sutton and Barto (2018), Mnih et al. (2015), Kaelbling et al. (1996)

Supervised learning is characterized by its use of labeled datasets, where each input is associated with a specific output. This method aims to learn the mapping from inputs to outputs, enabling the model to make accurate predictions on new, unseen data. Supervised learning is particularly effective for tasks that involve classification, where the objective is to categorize inputs into predefined classes, and regression, which focuses on predicting continuous outcomes. This approach is underpinned by a variety of algorithms, such as linear and logistic regression, support vector machines, neural networks, and random forests, each offering unique advantages in different scenarios (Bishop, 2006; Kotsiantis et al., 2007; Goodfellow et al., 2016). One of the earliest applications of supervised learning in the medical field was for the diagnosis of diseases. In the 1960s, researchers began using linear regression models to predict the presence of certain diseases based on patient data (Bishop, 2006). The development of neural networks in the 1980s further revolutionized medical diagnostics, enabling more accurate and complex pattern recognition (Albahra et al., 2023).

In epidemiology, supervised learning has been used to predict disease outbreaks and understand the spread of infectious diseases. One of the pioneering studies in this field used logistic regression to model the risk factors associated with the spread of malaria in populations (Kuang et al., 2009).

In contrast, unsupervised learning deals with data that lack labeled outputs, aiming to uncover hidden patterns, structures, or relationships within the data. This approach is instrumental for exploratory data analysis and is essential when the goal is to identify inherent groupings or reduce data dimensionality. Common applications include clustering, where similar data points are grouped together; dimensionality reduction, which simplifies data while retaining essential information; and anomaly detection, which identifies outliers within the data. Key algorithms in unsupervised learning include k-means clustering, hierarchical clustering, principal component analysis (PCA), and t-distributed stochastic neighbor embedding (t-SNE) (Hastie et al., 2009; Jain, 2010; Murphy, 2012). Unsupervised learning was first applied in the medical field to cluster patient data into meaningful groups. In the 1970s, clustering algorithms were used to group patients based on symptoms and laboratory results, aiding in the identification of disease subtypes (Jain, 2010). These early applications helped in understanding the heterogeneity of diseases and improving personalized medicine approaches. In epidemiology, unsupervised learning was first used to identify patterns in disease incidence data. For example, clustering algorithms were applied to group geographic regions based on the similarity of disease outbreak patterns, which helped in identifying hotspots and potential sources of outbreaks (De Silva, 2007). This approach has been essential in monitoring and controlling infectious diseases.

Reinforcement learning (RL) stands apart by focusing on learning through interaction with an environment. In RL, an agent learns to make decisions by performing actions and receiving feedback in the form of rewards or penalties. The primary goal is to develop a policy that maximizes cumulative rewards over time. This approach is highly applicable to scenarios that require sequential decision-making and adaptability, such as robotics, game playing, and autonomous systems. Notable algorithms in reinforcement learning include Q-learning, which focuses on learning the value of action-reward pairs; deep Q-networks, which utilize neural networks to approximate Q-values; policy gradient methods, which optimize the policy directly; and actor-critic methods, which combine value and policy learning to enhance performance (Kaelbling et al., 1996; Mnih et al., 2015; Sutton and Barto, 2018). RL has more recently been applied in the medical field, particularly in personalized treatment planning. In the 2009, researchers used RL algorithms to develop adaptive treatment strategies for blood glucose control in diabetic patients, optimizing the timing and dosage of interventions based on patient responses (Yasini et al., 2009). In epidemiology, RL has been explored to optimize public health interventions. One of the most

important applications was in the development of strategies for vaccination campaigns, where the goal was to determine the most effective allocation of limited vaccine supplies to minimize disease spread (Wei et al., 2021; Lu et al., 2023; Rey et al., 2023), which effectiveness was evident during COVID-19 pandemic (Beigi et al., 2021; Hao et al., 2022). Unfortunately, in the field of veterinary epidemiology no studies explored and utilized potential of RL on disease control strategies.

These three approaches to ML-supervised, unsupervised, and reinforcement learning — provide robust frameworks for analyzing and interpreting complex data. By leveraging the strengths of each method, researchers and practitioners can develop more effective and accurate models tailored to a wide range of applications, from predictive analytics to automated decision-making systems.

Use of ML in Veterinary Medicine

Significant applications of ML in veterinary medicine are already evident in diagnostic imaging, where ML algorithms, particularly deep learning models, have demonstrated considerable promises. These models analyze images from ultrasounds, MRIs, and X-rays to detect abnormalities such as tumors or fractures (Currie and Rohren, 2022; Tahghighi et al., 2023). In veterinary medicine, convolutional neural networks have been employed to diagnose conditions in pets and livestock with high accuracy, occasionally surpassing that of human experts (McEvoy, 2015). This application not only accelerates the diagnostic process but also enhances its accuracy, providing substantial support to veterinary professionals. Additionally, ML is utilized to monitor animal behavior and welfare. Sensors and cameras gather data on animal movements and behaviors, which ML algorithms analyze to detect signs of illness or stress. This application is particularly crucial in large-scale farming operations where manual monitoring is impractical. By providing continuous, automated monitoring, ML tools help ensure the well-being of large animal populations, highlighting unusual behaviors that might indicate health issues (Bidder et al., 2014).

ML emerged as a game-changing force in veterinary epidemiology, offering new tools to better understand zoonotic diseases spillover and enhance surveillance, detection response and recovery. The current applications of ML span a broad spectrum, from predictive modeling to pattern recognition, fundamentally altering how epidemiological data are analyzed and interpreted. Looking forward, ML is poised to further integrate into this field, addressing complex epidemiological challenges with increasingly sophisticated algorithms. At present, one of the primary applications of ML in veterinary epidemiology is in the development of predictive models. These models utilize historical data to forecast disease outbreaks, allowing for timely preventive measures. Supervised learning techniques, such as logistic regression and random forests, have been particularly effective in identifying the likelihood of disease occurrences based on various risk factors. For instance, models have been developed to enhance surveillance (Hepworth et al., 2012; Fountain-Jones et al., 2019; Bollig et al., 2020; Hyde et al., 2020; Zhang et al., 2021), aid in decision-making (Romero et al., 2020), and identify risk factors for animal diseases (Silva et al., 2019; Ghafoor and Sitkowska, 2021). These capabilities are crucial for allocating resources effectively and mitigating potential outbreaks before they become widespread.

The primary obstacle to the increased utilization of ML in veterinary epidemiology is likely the frequent occurrence of imbalanced class distributions in the data, where one class is significantly underrepresented compared to others. As it is often the case, especially in countries with diversity of production systems and lack of population data, the sample size, although considered representative of the population, is not deemed large enough to produce robust predictive models for diseases. These obstructions are extremely discouraging for researchers tackling rare diseases and/or small populations in building adequate predictive models for diseases. However, in the initial phases of applied ML, Weiss (2004) highlighted the challenge of imbalanced datasets, which can be classified into two categories:

relative rarity (where imbalance occurs within a large sample size) and absolute rarity (where the sample size is small and imbalanced). In many practical examples, a problem that commonly occurs concerning imbalanced datasets is that the cost of misclassifying a minority class example is significantly higher than that of misclassifying a majority class example. Since the development of control measures in epidemiology heavily relies on the use of statistical data calculations to construct disease models, it is crucial to develop classifiers that minimize overall misclassification costs, as stated by Weiss et al. (2007). This approach typically improves performance on the minority class compared to treating all misclassification costs equally. It also prevents the classifier from always predicting the majority class in highly imbalanced scenarios.

Several strategies are effective for handling skewed class distributions with unequal misclassification costs (He and Garcia, 2009). The most straightforward approach involves using a learning algorithm that inherently accounts for these costs when constructing the classifier (Weiss et al., 2007). Another approach for addressing skewed data with varying misclassification costs is to use sampling techniques to modify the class distribution of the training data. Two primary sampling methods can be employed for this purpose: oversampling and undersampling (Shelke et al., 2017; Mohammed et al., 2020). Oversampling involves replicating existent or creating new synthetic minority-class examples (Zheng et al., 2015), whereas undersampling entails removing examples from the majority class (Gosain and Sardana, 2017), thus eliminating potential bias due to misclassification (He and Garcia, 2009; Mohammed et al., 2020).

Among the oversampling techniques, Synthetic Minority Oversampling Technique (SMOTE) stands out as a particularly influential method for addressing imbalanced datasets in veterinary epidemiology. Introduced by Chawla et al. (2002), SMOTE generates synthetic samples from the minority class instead of creating exact copies. This method involves selecting examples that are close in the feature space, drawing a line between the minority samples in the space and generating new samples along that line. This approach helps create a more generalizable decision boundary for the minority class, effectively enhancing the performance of ML models (Chawla et al., 2002). In addition to SMOTE, advanced dimensionality reduction techniques are pivotal when dealing with high-dimensional data, which is typical in biological data relevant to veterinary epidemiology. Dimension reduction techniques like PCA and t-SNE are utilized to simplify the complexity of data while retaining its most significant features. PCA, for example, reduces the dimensionality of the data by transforming it into a new set of variables, which are linear combinations of the original variables and capture the maximum variance within the data (Jolliffe, 2002). t-SNE, on the other hand, is particularly well-suited for the visualization of high-dimensional datasets and can help identify clusters of similar data points, which is valuable for understanding the underlying structures of disease outbreaks (van der Maaten and Hinton, 2008). Another corrective method in the arsenal for tackling imbalanced datasets is the use of ensemble methods, such as Random Forests and Boosted Trees, which combine multiple weak classifiers to form a robust classifier. These methods can be particularly effective because they naturally handle imbalance by constructing multiple decision trees and aggregating their predictions, thereby providing a balanced perspective on the classes (Breiman, 2001; Schapire, 2003).

To implement these strategies effectively, it is essential for researchers in veterinary epidemiology to apply these techniques, but also rigorously validate their models. This includes cross-validation techniques and stratified sampling to ensure that the models are robust and perform well across different subsets of data (Kohavi, 1995). Moreover, ongoing research into novel ML techniques that can address the specific challenges of imbalanced data in veterinary contexts is crucial. Additionally, the construction of successful ML models can be significantly hindered by various data quality issues, commonly referred to as “noise”.

Noise in ML is defined as random or irrelevant data within a dataset that does not contribute to the pattern or structure the model is designed to learn (Gupta and Gupta, 2019). This can obscure the true relationships between features and the target variable, thereby complicating the model's ability to generalize effectively to new, unseen data (Saseendran et al., 2019). One prevalent source of noise is measurement errors, which occur due to inaccuracies or inconsistencies in data collection methods or instruments used in field studies or clinical environments. These errors can include everything from error-prone measurement tools to biological variability in samples, which can skew data and mislead analytical models (Greenland et al., 2016). Label noise also poses a significant challenge, particularly when diagnostic criteria are subjective or when multiple evaluators are involved without sufficient calibration. Incorrect or inconsistent labeling of training data can lead to misclassifications and reduce the overall effectiveness of predictive modeling (Garcia et al., 2015).

Furthermore, veterinary epidemiological datasets often include irrelevant features-data points that do not have a meaningful relationship with the target variable. This can happen due to the inclusion of extensive and diverse data sources, which may dilute the predictive power of the model (Van der Waal et al., 2017). Environmental factors introduce another layer of complexity: external conditions such as temperature or humidity at the time of sample collection can introduce variability that is not relevant to the study's objectives, but can affect the data collected (Weatherhead et al., 1998). Lastly, the inherent randomness in the data generation process, which is not captured by the model, contributes to noise. This includes biological variability among subjects or uncontrolled environmental influences during data collection, adding further uncertainty to the data analysis process (Louppe, 2014). The presence of such noise in the dataset can lead to several complications such as overfitting, where the model erroneously learns to capture the noise as if it were a true signal, thereby harming the model's performance on new data. It can also reduce the overall accuracy and increase the complexity of the model, resulting in higher computational costs and inefficient learning processes (Domingos, 2012).

To mitigate these issues, it is crucial for researchers in veterinary epidemiology to employ robust data preprocessing techniques to clean and normalize data, utilize feature selection algorithms to reduce dimensionality and discard irrelevant features, and apply advanced ML techniques that are inherently more resistant to noise, such as ensemble methods or regularized regression models (Kumar and Minz, 2014). These strategies help ensure that the predictive models developed are both accurate and robust, capable of providing reliable insights for disease management and control.

Future of ML

Looking to the future, the integration of ML in veterinary epidemiology is expected to expand further with the advancement of technologies such as natural language processing (NLP) and reinforced learning (RL). NLP can be utilized to shift through vast amounts of unstructured data from veterinary records, social media, and literature to identify disease trends and outbreaks. This capability could be particularly transformative in managing zoonotic diseases, where early detection and rapid response are critical (Clark et al., 2020).

RL, on the other hand, offers potential for optimizing decision-making processes in epidemiology. For example, RL could be used to develop dynamic vaccination strategies, adjusting in real-time based on feedback from ongoing disease surveillance data. This approach could maximize the effectiveness of interventions by tailoring them to the specifics of an outbreak scenario, thereby minimizing the spread and impact of diseases (Libin et al., 2021).

Despite these advances, the application of ML in veterinary epidemiology is not without challenges. Issues such as data quality, privacy concerns, and the need for robust computational infrastructure pose significant hurdles. Additionally, there is a

need for interdisciplinary collaboration between veterinarians, epidemiologists, and data scientists to ensure that ML models are effectively integrated into veterinary practices and that they address practical, real-world problems (Van der Waal et al., 2017).

In summary, based on our review and current understanding, we propose that the incorporation of ML into veterinary epidemiology heralds a potential paradigm shift. This integration could significantly improve disease surveillance, detection, and management throughout the value chain. By leveraging ML's predictive accuracy and ability to handle complex data, veterinary professionals can improve hazard recognition and detection, develop better "tailor made" preventive programs and improve health outcomes for animal populations and food safety and security trends. However, overcoming challenges such as data quality, privacy concerns, and interdisciplinary collaboration is crucial. Continued research and development in ML applications will undoubtedly pave the way for more robust and effective epidemiological practices, ultimately advancing veterinary science and animal health on a global scale.

ACKNOWLEDGEMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

The authors declared that there is no conflict of interest.

CONTRIBUTIONS

Concept and Design – AM, NF; Supervision – NF; Literature review, analysis and interpretation of data - AM, NF; Writing and Critical review - AM, NF

REFERENCES

Albahra S, Gorbett T, Robertson S, D'Aleo G, Kumar SVS, Ockunzzi S, et al. 2023. Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. Semin Diagn Pathol, 40(2), 71-87.

Currie G, Rohren E. 2022. Intelligent imaging: Applications of machine learning and deep learning in radiology. Vet Radiol Ultrasound, 63, 880-8.

Tahghighi P, Appleby RB, Norena N, Ukwatta E, Komeili A. 2023. Machine learning can appropriately classify the collimation of ventrodorsal and dorsoventral thoracic radiographic images of dogs and cats. Am J Vet Res, 84(7).

Gupta S, Gupta A. 2019. Dealing with noise problem in machine learning data-sets: A systematic review. Procedia Comput Sci, 161, 466-74.

Saseendran AT, Setia L, Chhabria V, Chakraborty D, Barman Roy A. 2019. Impact of noise in dataset on machine learning algorithms. Mach Learn Res, 1, 1-8.

Beigi A, Yousefpour A, Yasami A, Gómez-Aguilar JF, Bekiros S, Jahanshahi H, et al. 2021. Application of reinforcement learning for effective vaccination strategies of coronavirus disease 2019 (COVID-19).Eur Phys J Plus, 136(5), 1-22.

Bidder OR, Campbell HA, Gómez-Laich A, Urgé P, Walker J, Cai Y, et al. 2014. Love thy neighbour: automatic animal behavioural classification of acceleration data using the k-nearest neighbour algorithm. PloS One, 9(2), e88609.

Bishop C. 2006. Pattern recognition and machine learning. New York, USA: Springer.

Bollig N, Clarke L, Elsmo E, Craven M. 2020.Machine learning for syndromic surveillance using veterinary necropsy reports. PloS One, 15(2), e0228105.

Breiman L. 2001. Random Forests. Mach Learn, 45(1), 5-32.

Casella G, Berger RL. 2002. Statistical Inference. Duxbury Resource Center.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res, 16, 321-57.

Clark NJ, Tozer S, Wood C, Firestone SM, Stevenson M, Caraguel C, et al. 2020. Unravelling animal exposure profiles of human Q fever cases in Queensland, Australia, using natural language processing. Transbound Emerg Dis, 67(5), 2133-45.

De Silva D, Alahakoon D, Dharmage S. 2007. Cluster analysis using the GSOM: Patterns in epidemiology, in: 2007 Third Int Conf Inf Automat Sustain. IEEE, 63-9.

Domingos P. 2012. A Few Useful Things to Know about Machine Learning. Commun ACM, 55(10), 78-87.

Fountain□Jones NM, Machado G, Carver S, Packer C, Recamonde□Mendoza M, Craft ME. 2019. How to make more from exposure data? An integrated machine learning pipeline to predict pathogen exposure. J Anim Ecol, 88(10), 1447-61.

Garcia LP, de Carvalho AC, Lorena AC. 2015. Effect of label noise in the complexity of classification problems. Neurocomputing, 160, 108-19.

Ghafoor N, Sitkowska B. 2021. MasPA: A machine learning application to predict risk of mastitis in cattle from AMS sensor data. Agri Engineering, 3(3), 575-583.

Goodfellow I, Bengio Y, Courville A. 2016. Deep Learning. MIT Press.

Gosain A, Sardana S. 2017. Handling class imbalance problem using oversampling techniques: A review, in: 2017 Int Conf Adv Comput Commun Informatics (ICACCI). IEEE, 79-85.

Greenland S, Daniel R, Pearce N. 2016. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. Int J Epidemiol, 45(2), 565-75.

Hao Q, Huang W, Xu F, Tang K, Li Y. 2022. Reinforcement learning enhances the experts: Large-scale covid-19 vaccine allocation with multi-factor contact network, in: Proc 28th ACM SIGKDD Conf Knowl Discov Data Min. pp. 4684-94.

Hastie T, Tibshirani R, Friedman J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York, USA: Springer.

He H, Garcia EA. 2009.Learning from imbalanced data. IEEE Trans Knowl Data Eng, 21(9), 1263-84.

Hepworth PJ, Nefedov AV, Muchnik IB, Morgan KL. 2012. Broiler chickens can benefit from machine learning: support vector machine analysis of observational epidemiological data. J R Soc Interface, 9(73), 1934-42.

Hyde RM, Down PM, Bradley AJ, Breen JE, Hudson C, Leach KA, et al. 2020. Automated prediction of mastitis infection patterns in dairy herds using machine learning. Sci Rep, 10(1), 4289.

Jain AK. 2010. Data clustering: 50 years beyond K-means. Pattern RecognLett, 31(8), 651-66.

James G, Witten D, Hastie T, Tibshirani R. 2013. An Introduction to Statistical Learning. New York, USA: Springer.

Jolliffe IT. 2002. Principal Component Analysis for Special Types of Data, in: Principal Component Analysis. 2nd ed. New York, USA: Springer, 338-72.

Kaelbling LP, Littman ML, Moore AW. 1996. Reinforcement Learning: A Survey. J Artif Intell Res, 4, 237-85.

Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, in: IJCAI. Vol. 14, No. 2, 1137-45.

Kuang R, Gu J, Cai H, Wang Y. 2009. Improved prediction of malaria degradomes by supervised learning with SVM and profile kernel. Genetica, 136, 189-209.

Kumar S, Minz S. 2014. Feature Selection: A Literature Review. SmartCR, 4(3), 211-29.

Libin PJ, Moonens A, Verstraeten T, Perez-Sanjines F, Hens N, Lemey P, et al. 2021. Deep reinforcement learning for large-scale epidemic control, in: Mach Learn KnowlDiscov Databases. Appl Data Sci Demo Track: Eur Conf, ECML PKDD 2020, Ghent, Belgium, Sep 14–18, 2020, Proc, Part V. Springer Int Publish, 155-70.

Louppe G. 2014. Understanding random forests: From theory to practice. arXiv preprint arXiv:1407.7502. https://arxiv.org/abs/1407.7502 (accessed 27 May 2024).

Lu Y, Wang Y, Liu Y, Chen J, Shi L, Park J. 2023. Reinforcement learning relieves the vaccination dilemma. Chaos: Interdiscip J Nonlinear Sci, 33(7).

McEvoy FJ. 2015. Grand challenge veterinary imaging: technology, science, and communication. Front Vet Sci, 2, 38.

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. 2015. Human-Level Control through Deep Reinforcement Learning. Nature, 518(7540), 529-33.

Mohammed R, Rawashdeh J, Abdullah M. 2020. Machine learning with oversampling and undersampling techniques: overview study and experimental results, in: 2020 11th Int Conf Inf Commun Syst (ICICS). IEEE, 243-8.

Molenberghs G, Kenward MG. 2007. Missing Data in Clinical Studies. Chichester, UK: Wiley.

Murphy KP. 2012. Machine Learning: A Probabilistic Perspective. MIT Press.

Rey D, Hammad AW, Saberi M. 2023. Vaccine allocation policy optimization and budget sharing mechanism using reinforcement learning. Omega, 115, 102783.

Romero MP, Chang YM, Brunton LA, Parry J, Prosser A, Upton P, et al. 2020. Decision tree machine learning applied to bovine tuberculosis risk factors to aid disease control decision making. Prev Vet Med, 175, 104860.

Rothman KJ, Greenland S, Lash TL. 2008. Modern Epidemiology. 3rd ed. Philadelphia, USA: Lippincott Williams & Wilkins.

Schapire RE. 2003. The boosting approach to machine learning: An overview, in: MSRI Workshop on Nonlinear Estimation and Classification.

Shelke MS, Deshmukh PR, Shandilya VK. 2017. A review on imbalanced data handling using under sampling and oversampling technique. Int J Recent Trends Eng Res, 3(4), 444-9.

Silva GS, Machado G, Baker KL, Holtkamp DJ, Linhares DC. 2019. Machine-learning algorithms to identify key biosecurity practices and factors associated with breeding herds reporting PRRS outbreak. Prev Vet Med, 171, 104749.

Sutton RS, Barto AG. 2018. Reinforcement Learning: An Introduction. MIT Press.

Van der Maaten L, Hinton G. 2008. Visualizing Data using t-SNE. J Mach Learn Res, 9(Nov), 2579-605.

Van der Waal K, Morrison RB, Neuhauser C, Vilalta C, Perez AM. 2017. Translating big data into smart data for veterinary epidemiology. Front Vet Sci, 17(4), 110.

Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. 2012. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. 2nd ed. New York, USA: Springer.

Wei X, Pu C, He Z, Liò P. 2021. Deep reinforcement learning-based vaccine distribution strategies, in: 2021 2nd Int Conf Electron Commun Inf Technol (CECIT). IEEE, 427-36.

Weiss GM. 2004. Mining with rarity: a unifying framework. SIGKDD Explor, 6(1), 7-19.

Weatherhead EC, Reinsel GC, Tiao GC, Meng XL, Choi D, Cheang WK, et al. 1998. Factors affecting the detection of trends: Statistical considerations and applications to environmental data. J Geophys Res: Atmos, 103(D14), 17149-61.

Zhang S, Su Q, Chen Q. 2021. Application of machine learning in animal disease analysis and prediction. Curr Bioinformatics, 16(7), 972-82.

Zheng Z, Cai Y, Li Y. 2015. Oversampling method for imbalanced classification. Comput Inform, 34(5), 1017-37.

UNAPREĐENJE VETERINARSKE EPIDEMIOLOGIJE INTEGRACIJOM MAŠINSKOG UČENJA: TRENUTNO STANJE I PERSPEKTIVE

SAŽETAK

Integracija mašinskog učenja (MU) u veterinarsku epidemiologiju ima transformativni potencijal u polju analize podataka i menadžmenta, što predstavlja značajan zaokret u odnosu na tradicionalne statističke metode. U ovom Pregledu istražujemo rastuću ulogu MU sa naglaskom na njegov kapacitet u upravljanju složenim, višedimenzionalnim podacima i otkrivanju nelinearnih odnosa koji su od presudnog značaja u epidemiologiji. Ključne metodologije MU, uključujući *supervised, unsupervised* i *reinforcement learning* osiguravaju robustne okvire za kreiranje prediktivnih modela, obrasce prepoznavanja i procese donošenja odluka. Već se primjenjuju u veterinarskoj medicini u oblasti dijagnostičkog snimanja i monitoringu životinjskog ponašanja, pokazujući sposobnost MU da poveća dijagnostičku preciznost i ojača zdravstveni monitoring.

Uprkos ovakvom napredovanju, praktični izazovi uključuju neizbalansirane skupove podataka, probleme kvalitete podataka te potrebu za interdisciplinarnom suradnjom. Strategije poput *Synthetic Minority Oversampling Technique* i ansambl metode su od pomoći kod neuravnotežene distribucije klasa, dok robustne tehnike prethodne obrade ublažavaju podatkovne šumove. Budući razvoj u obradi prirodnog jezika i pojačanog učenja obećava daljnju integraciju, optimiziranje praćenja bolesti i interventne strategije.

Ovaj Pregled naglašava transformativni potencijal MU u veterinarskoj epidemiologiji promovirajući kontinuirano istraživanje i suradnju sa ciljem prevazilaženja postojećih problema. Iskorištavanjem mogućnosti MU, veterinari mogu unaprijediti predviđanje bolesti, razviti ciljane preventivne programe i u cjelosti unaprijediti zdravlje životinja i ispravnost hrane, dajući značajan doprinos razvoju veterinarske znanosti.

Ključne riječi: Kreiranje prediktivnih modela, praćenje bolesti, veterinarska medicina, vještačka inteligencija, zdravlje životinja