





# Data Mining for Animal Health to Improve Human Quality of Life: Insights from a University Veterinary Hospital

Oscar Tamburis<sup>1</sup><sup>a</sup>, Elio Masciari<sup>2</sup><sup>b</sup>, Christian Esposito<sup>3</sup><sup>c</sup> and Gerardo Fatone<sup>1</sup><sup>d</sup>

<sup>1</sup>*Dept. of Veterinary Medicine and Animal Productions, Federico II University, Via Delpino 1, Naples, Italy*

<sup>2</sup>*Dept. of Computer Science and Electrical Engineering, Federico II University, Naples, Italy*

<sup>3</sup>*Dept. of Computer Science, University of Salerno, Fisciano (SA), Italy*


**Keywords:** Veterinary Medicine, Electronic Medical Record, Decision Tree Algorithm, One Health.


**Abstract:** The increasing importance of Veterinary Informatics is driving the implementation of integrated veterinary information management systems (VIMS) for the capture, storage, analysis and retrieval of animal data. In this paper, a decision tree algorithm was implemented, starting from the database of the University Veterinary Hospital at Federico II University of Naples, aiming at building a predictive model for an effective recognition of neoplastic diseases and zoonoses for cats and dogs focusing to Campania Region, in order to figure out, according to the One (Digital) Health perspective specifics, the connection between humans, animals, and surrounding environment.


## 1 INTRODUCTION


Animals, be they categorized as pets, livestock, or wildlife, stand as essential element in the evolution of human race for countless reasons. In particular, animal healthcare-related aspects play a prominent role because of their strict connections with human health. The monitoring of both wildlife and syntropic species' health state can provide in fact valuable information about (i) the quality of the environment they live in, and that they share with humans, in terms of pollution level, as well as food safety and traceability management; (ii) the occurring of zoonotic phenomena (for instance, leptospirosis and the recent COVID-19 pandemic). Furthermore, many non-infectious diseases (e.g. diabetes, cancer, and renal failure) are similar in both animals and humans (Smith-Akin et al., 2007). As a consequence, the need for an effective tracking of veterinary information to facilitate integration of animal medical data to support Public Health, has become essential. As a matter of fact, under the epidemiological perspective the advantages of using animals as sentinels or comparative models of human diseases are well

known, as animals – or better, animal sentinels – may be sensitive indicators of environmental hazards and provide an early warning system for public health interventions (Vilhena et al., 2020). With specific reference to Campania Region, this kind of studies are of particular concern due to the widely known so-called “Terra dei Fuochi/Land of Fires” phenomenon (see e.g. Zaccaroni et al., 2014; Cavallo et al., 2018). An as important aspect relates then to the control of the zoonoses, i.e. those diseases that can be transmitted from the animals to the human beings via faeces, urine, saliva, or blood. It is the case of e.g. intestinal parasites and ticks (that use the animal as a vector), or rabies (transmitted via the saliva). Such risks have to be carefully taken into account when it comes to the cohabitation between humans and (conventional as well as non-conventional) pets (Salyer et al., 2017; Mhlanga, 2020). To this end, it becomes useful to resort to data mining computational methods for extracting knowledge also in the case of animal large databases deployed in integrated veterinary information management systems (VIMS) (Plavšić et al., 2009). Among the most diffused data mining algorithms (Masciari, 2012; Ficco et al., 2015;

<sup>a</sup> <https://orcid.org/0000-0002-0130-7915>

<sup>b</sup> <https://orcid.org/0000-0002-1778-5321>

<sup>c</sup> <https://orcid.org/0000-0002-0085-0748>

<sup>d</sup> <https://orcid.org/0000-0003-2578-7420>

Ianni et al., 2020), decision tree provides a tree-based classification for developing a predictive model according to independent variables (Bernardi et al., 2017; Haq et al., 2020).

In this paper the main results will be shown from the analysis of the data extracted from PONGO software ©, i.e. the first EMR solution implemented in the University Veterinary Teaching Hospital (it.: OVUD, acronym for Ospedale Veterinario Universitario Didattico) of the “Federico II” University of Naples, Italy. The main goal was to establish, by means of decision tree algorithm, a predictive model for an effective recognition of neoplastic diseases and zoonoses using clinical data, according to clinical, para-clinical, and demographic attributes. The investigation on the quality of clinical data of OVUD’s patients is intended for helping, at least on a region-wide scenario, to find out the presence of specific connections between people’s health, animal health, and their surrounding environment, thus conveying the specific Public Health dimension into the greater One (Digital) Health scenario (Gamache et al., 2018; Magnuson & Dixon, 2020; Benis et al., 2021).

## 2 MATERIALS AND METHODS

### 2.1 Subjects

The data extracted from PONGO sw in form of MS Access DB relate to the general physical examination (GPE), that is the first visit performed from the veterinarian when the animal arrives to the hospital. The database contains about 10360 rows (one row per animal access) which span over a period going from 2010 to mid-2020. The visits were mainly performed on pets, i.e. dogs (n = 8925; 86%) and cats (n = 1181; 11%). Horses occurred to be treated in the hospital as well (n = 160; 2%). Only for a small part (n = 92; 1%) the animals examined belonged to other species (ducks, donkeys, bovines, buffaloes, goats, lagomorphs, rodents, tortoises, and birds). Besides animal species and date of the visit, the main fields of the DB also related to age and sex of the animal, main health issue (HI) acknowledged during the GPE, type of feeding (e.g. commercial vs. homemade), and vaccination status information. Also considered in the study were the kind of environments the animal used to live in (e.g. in an apartment, or outdoors), and the Italian province it came from. As for the latter point, the research was limited to the provinces of Campania Region, due to the marginal number of rows related to patients coming from other Italian regions. Table 1

describes the accesses to OVUD, on the basis of the geographic provenance, for dogs, cats, and horses. A number of OLAP operations (Pešić et al., 2009; Lu & Keech, 2015) were performed to investigate the quality of clinical data of OVUD’s patients for the considered time period. Given the situation, it was decided to focus the investigation only on dogs and cats.

Table 1: Distribution of dogs, cats, and horses that accessed the OVUD, according to the Italian provinces.

Species	Province	#	%
Dog	Avellino	145	2%
	Benevento	71	1%
	Caserta	753	8%
	Napoli	6967	78%
	Salerno	444	5%
	Other Italian provinces	545	6%
Cat	Avellino	22	2%
	Benevento	9	1%
	Caserta	68	6%
	Napoli	975	83%
	Salerno	41	3%
	Other Italian provinces	66	6%
Horse	Avellino	3	2%
	Benevento	7	4%
	Caserta	14	9%
	Napoli	64	40%
	Salerno	49	31%

### 2.2 Accesses per Animal Sex

Four types of sex specifications have to be considered for animals: male (M), castrated male (MC), female (F), spayed female (FS). Figure 2 reports the accesses to the OVUD of dogs and cats, respectively, for the time period considered. The number of rows/visits for which it was not possible to retrieve the sex of the animal, were also reported. Only in one case, the animal (dog) was reported as not visited after the access in the hospital. The lower number of accesses registered in 2016 in both cases, was due to a partial stop of the OVUD activities, as a structural collapse interested at the end of 2015 part of the University Department that hosts the hospital itself. The number of male dogs’ accesses is about twice as much the female accesses in almost all the years considered, with quite lower numbers for the neutered dogs. A different situation concerns cats, where the differences M/MC and F/FS tend to be proportionally shorter, sometimes in favour of the neutered exemplars.

### 2.3 Health Issues per Year

It was possible to identify about 140 different diagnoses from the GPE for the period considered.

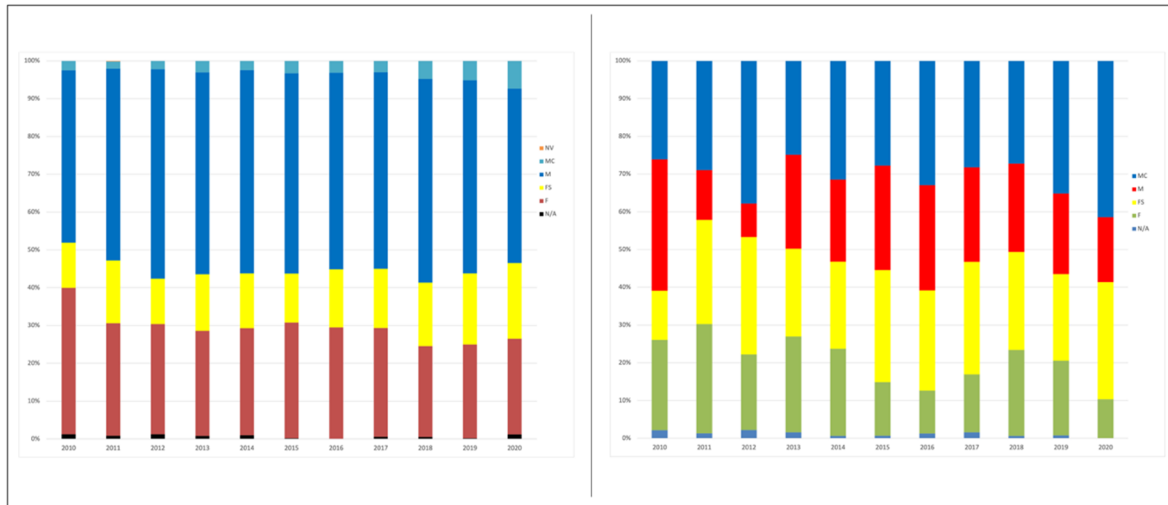


Figure 1: Accesses to OVUD of dogs (left) and cats (right).

For mere space reasons, it was decided for both dogs and cats to investigate, for each year, only the three most relevant health issues (HIs), as reported in Tables 2 and 3. In case of HIs featuring the same occurrences, they were all considered. The only exception is for cats' HIs in 2012, where the occurrences for HI #3 were equal to 1 for a very large set of issues, so it was decided not to report them in the table.

Table 2: The three most diagnosed health issues for dogs.

Year	HI #1	HI #2	HI #3
2010	Limping	Injury of abdomen	Firm lymph node (on exam.)
2011	Limping	On exam. - inspection of vomit	Pain in eye; Skin lesion (on exam.); Cough
2012	Alopecia	Skin lesion (on exam.); Limping	On exam. - inspection of vomit
2013	Limping	Alopecia	On exam. - inspection of vomit
2014	Limping	Neoplastic disease	Alopecia
2015	Limping	Neoplastic disease	Alopecia
2016	Limping	Injury of abdomen; Alopecia	Firm lymph node (on exam.)
2017	Limping	Firm lymph node (on exam.); Alopecia	Neoplastic disease
2018	Limping	Injury of abdomen	Neoplastic disease
2019	Injury of abdomen	Cough; Neoplastic disease	Limping
2020	Injury of abdomen	Alopecia	On exam. - inspection of vomit

Table 3: The three most diagnosed health issues for cats.

Year	HI #1	HI #2	HI #3
2010	On exam. - inspection of vomit; Pain in eye	Injury of abdomen	Urinary tract pain
2011	On exam. - inspection of vomit	Alopecia	Urinary tract pain
2012	On exam. - inspection of vomit; Pain in eye	Firm lymph node (on exam.); Alopecia	.
2013	On exam. - inspection of vomit	Pain in eye; Injury of abdomen	Alopecia
2014	Pain in eye	Injury of abdomen	Urinary tract pain
2015	Alopecia; On exam. - inspection of vomit	Neoplastic disease	Skin lesion (on exam.); Pain in eye
2016	On exam. - inspection of vomit	Injury of abdomen	Alopecia
2017	Injury of abdomen	Skin lesion (on exam.)	On exam. - inspection of vomit; Urinary tract pain; Neoplastic disease
2018	On exam. - inspection of vomit	Alopecia; Pain in eye; Injury of abdomen	Skin lesion (on exam.)
2019	Injury of abdomen	Cough	Pain in eye; Alopecia
2020	Injury of abdomen	Firm lymph node (on exam.)	Skin lesion (on exam.); Closed fracture of hip; Sore mouth

It can be noticed for dogs a diffuse presence of limping-related issues (N = 458), along with Neoplastic diseases (N = 263), and alopecia (N = 239). Injury of abdomen (N = 46), inspection of vomit (N = 46), and pain in eye (N = 40) appear instead among the most diffused issues reported for the cats that accessed the OVUD. The occurrences of such HIs during the years are reported in Figure 2. The total number of occurrences are depicted in Figure 3. In both cases, it is worth noticing the presence of neoplastic diseases (dogs: N = 263; cats: N = 9) and firm lymph node-related (dogs: N = 95; cats: N = 3) diagnoses. Moreover, considering animals' age of birth (spanning from 1984 to 2020), it was possible to compare for each trimester the diagnoses of firm lymph nodes and neoplastic diseases. This revealed that the 44% cases of dogs of the same age, and the 5% cases of cats of the same age presented a number of occurrences of firm lymph node-related diagnoses greater or at least equal to neoplastic diseases diagnoses, thus inducing – at least for dogs – the reasonable hypothesis of an existing connection between the two pathologies. Furthermore, Figure 5 reports the occurrences of those diagnoses which can be somehow related to the transmission of zoonoses,

from tetanus (N = 1 for dogs) to vomit (dogs: N = 254; cats: N = 53). The number of occurrences of such diagnoses is the 7% of the total occurrences registered in OVUD for dogs and cats for the period considered.

## 2.4 Dataset

A preliminary step of dataset cleansing was necessary, especially for what concerns the health diagnoses, as no form of clinical standardized terminology had been deployed. Moreover, for about 30% rows (N = 3729), such type of data was actually missing, and only in a limited number of cases it was possible to get to it anyway by means of the analysis of the remainder fields of the database. Eventually, the total number of participants considered in the model were 10108. Given the mentioned importance of identifying the presence of neoplastic diseases-related and/or zoonoses-related diagnoses, the need emerged to figure out a way to predict the presence of symptoms for both the issues considered – for both dogs and cats, who also happen to live very close to humans. In particular, according to what depicted in Figure 4, for what concerns zoonoses it was decided to consider for the analysis the diagnosis of “inspection of vomit”.



Figure 2: Distribution of the three most diagnosed health issues per year, for dogs (up) and cats (down).

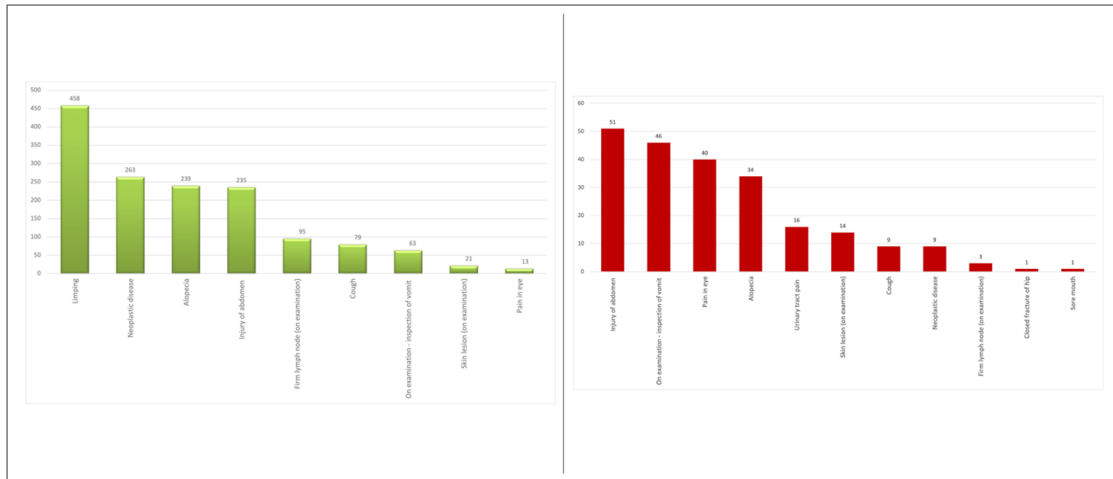


Figure 3: Total occurrences for the three most diagnosed health issues, for dogs (up) and cats (down).

### 2.5 DT ID3 Feature Selection Algorithm

The implementation of a Decision Tree Algorithm (DT) appeared as the most suitable way to investigate the membership of the subjects to different categories (diagnosed with neoplastic disease, or not; diagnosed with vomit, or not), taking into account the values of specific attributes (predictor variables), which in our case were identified for both cases as: animal sex, diet, vaccination, feeding routines, and living environment (plus the eventual presence of diagnosis of firm lymph nodes, for neoplastic diseases). In order to achieve these goals, a filter-based strategy using DT ID3 (Iterative Dichotomiser 3) was proposed (Gharehchopogh et al., 2012; Tayefi et al., 2017).

As it is common in data mining methods to divide the dataset into two parts, also in our case the original sample was split into a training set (to train the model), and a test set (to evaluate the performance of DT ID3). In particular, the original Training dataset for the DT (oTrDS) featured all the accesses of dogs and cats to the OVUD between 2010 and 2018 (N = 8643; 86%), while the original Testing dataset (oTeDS) comprised the remaining accesses between 2019 and 2020 (N = 1465; 14%). The reason why it was not respected the common rule according to which oTrDS  $\approx$  70% sampling data, and oTeDS  $\approx$  remaining 30%, mainly depends on two factors: (i) the reduced accesses to OVUD in 2016 due to the mentioned structure collapse, and; (ii) available data from year 2020 only cover the first six months. Since the aim of the study was to make prediction for two kind of health issues, each per two animal species, four specific Training datasets (sTrDS) and four specific Testing datasets (sTeDS) were extracted

from oTrDS and oTeDS, respectively. For each case, a confusion matrix was used to evaluate the performance of the DT for classification of participants. Accuracy, sensitivity, and specificity were then measured for comparison. For sake of simplification, decision tree and confusion matrix have been represented in the following for one case only (presence of symptoms for neoplastic disease in dogs). A comparison was instead conducted for the performances of all four algorithms.

## 3 RESULTS

A decision tree was built starting from the sTrDS related to the recognition of neoplastic disease for dogs (N = 8927). The sTeDS (N = 1305) was used to evaluate the model. The input variables were animal sex, diet, vaccination, feeding routines, living environment, and eventual presence of diagnosis of firm lymph nodes. As seen, since for dogs the possibility of a correlation was recognized between the diagnoses of neoplastic disease and firm lymph nodes, the number of subjects positive for both health issues (ND+ and L+) was reported in the algorithm. ID3 uses two metrics to measure the importance of the input variables, or features, such as entropy (the measure of the amount of uncertainty) and information gain (the difference between the entropy of the DS, and the one related to the single feature). So, be DS a given dataset, and X the set of variables in DS. For each  $x \in X$ , the less the entropy, the more the information gain. For each iteration, the algorithm selects the feature with the smallest entropy/largest information gain value. The final decision tree with size 15, 8 leaves and 5 layers is shown in Figure 5.

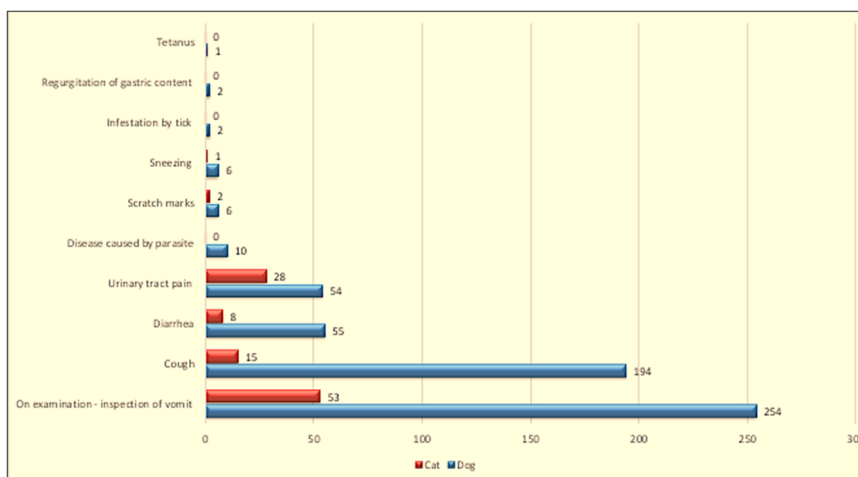


Figure 4: Total occurrences of zoonosis-related diagnoses, for both dogs and cats.

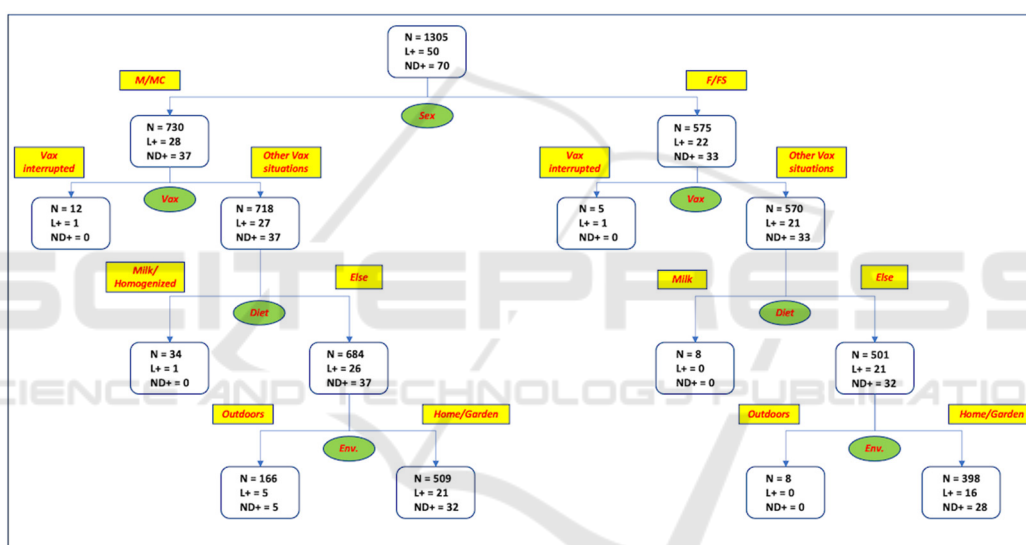


Figure 5: Decision Tree to evaluate the presence of symptoms for neoplastic disease in dogs.

The evaluation of the tree was undertaken using confusion matrix on a testing dataset, as shown in Table 4. The algorithm had an Accuracy of 99%: of the 70 animals diagnosed as ND+ in the sTeDS, 60 were correctly classified using the DT. In a subordinate position, of the 50 animals diagnosed as L+, 37 were correctly identified. The specificity and sensitivity of the tree were equal to 99,2 and 1, respectively. The performance of DT was also reported in Table 5.

Table 4: Confusion Matrix of sTeDS related to the recognition of neoplastic disease for dogs.

		Predicted outcome	
		ND+	ND-
Expected outcome	ND+	60 (TP)	10 (FP)
	ND-	0 (FN)	1235 (TN)

Table 5: Performance of the DT ID3 model for the case investigated.

Variable	Decision Tree Model
Sensitivity (95% CI)	1 (93,9 – 1)
Specificity (95% CI)	99,2 (98,5 – 99,6)
Accuracy (95% CI)	99 (98,3 – 99,4)

An overall comparison was instead conducted between the performances of the algorithm for the four cases investigated, as reported in Table 6. Although the numbers of cats-related diagnoses extracted from the PONGO DB were significantly lesser than the dogs-related ones, the overall results obtained confirmed anyway the validity of the data mining algorithm implemented, which turned as highly capable of modelling the process of healthcare

provision (Tamburis, 2019), as well as of setting forth reliable measurements of system performance and outcomes (Luzi et al., 2017).

Table 6: Comparison of the performances of the DT ID3 algorithm for all the cases investigated.

		Accuracy	Sensitivity	Specificity
Dogs	Neoplastic Disease	99%	100%	99,2%
	Zoonosis (Vomit)	100%	100%	100%
Cats	Neoplastic Disease	100%	100%	100%
	Zoonosis (Vomit)	100%	100%	100%

## 4 DISCUSSION AND CONCLUSIONS

In this paper a decision tree algorithm was implemented, starting from the database of the University Veterinary Teaching Hospital of the Federico II University of Naples, to work out a predictive model for an effective recognition of neoplastic diseases and zoonoses using clinical data, according to clinical, para-clinical, and demographic attributes. The main scope was to investigate whether and at what extent relations can stand between human and animal health, and their surrounding environments. The whole set of disciplines broadly dealing with the such kind of “connecting chain” goes under the name of One Health (OH), introduced for the first time as part of the twelve “Manhattan Principles” calling for an international, interdisciplinary approach to prevent diseases (Mackenzie & Jeggo, 2019) and specifically animal-human transmissible and communicable ones. Seen under this comprehensive point of view, the bursting of dynamics connected to the emerging and re-emerging of infectious diseases from national to supranational contexts, as well as the need to identify at global level risk factors and causes of health problems that arise at the human-animal-environment crossing, made even more remarkable the role of

veterinarians towards the protection of human health. This points out therefore the growing of veterinary informatics, as also encompassing the need for new paradigms, approaches and technologies to reinforce the capacity of traditional surveillance systems for prevention and control of zoonoses, in terms of i.e. inter-sectoral coordination, link between human and animal health data and consequent management of flows of reliable data and information, or proper use of infrastructures, systems and human resources to detect outbreaks (Choi et al., 2016).

## REFERENCES

- Smith-Akin, K. A., Bearden, C. F., Pittenger, S. T., & Bernstam, E. V., 2007. Toward a veterinary informatics research agenda: an analysis of the PubMed-indexed literature. *International journal of medical informatics*, 76(4), 306-312.
- Vilhena, H., Figueira, A. C., Schmitt, F., Canadas, A., Chaves, R., Gama, A., & Dias-Pereira, P. 2020. Canine and Feline Spontaneous Mammary Tumours as Models of Human Breast Cancer. In *Pets as Sentinels, Forecasters and Promoters of Human Health* (pp. 173-207). Springer, Cham.
- Zaccaroni, A., Corteggio, A., Altamura, G., Silvi, M., Di Vaia, R., Formigaro, C., & Borzacchiello, G., 2014. Elements levels in dogs from “triangle of death” and different areas of Campania region (Italy). *Chemosphere*, 108, 62-69.
- Cavallo, S., Serpe, F. P., Rea, D., Pellicanò, R., D'Amore, M., Martinis, C. D., ... & Baldi, L., 2018. The Land of Fires in Campania: the effects of exposure to dioxins on the progression of human breast cancer in an innovative animal model. In *XVIII Congresso Nazionale SI Di. LV, Perugia (PG), Italia, 7-9 Novembre 2018* (pp. 41-42). Società Italiana di Diagnostica di Laboratorio Veterinaria (SIDiLV).
- Salyer, S. J., Silver, R., Simone, K., & Behravesh, C. B., 2017. Prioritizing zoonoses for global health capacity building—themes from One Health zoonotic disease workshops in 7 countries, 2014–2016. *Emerging infectious diseases*, 23(Suppl 1), S55.
- Mhlanga, A., 2020. Assessing the Impact of Optimal Health Education Programs on the Control of Zoonotic Diseases. *Computational and Mathematical Methods in Medicine*, 2020.
- Plavšić, B., Nedić, D., Mićović, Z., Tešić, M., Stanojević, S., Ašanin, R., ... & Milanović, S., 2009. Veterinary information management system (VIMS) in the process of notification and management of animal diseases. *Acta veterinaria*, 59(1), 99-108.
- Masciari, E., 2012. SMART: stream monitoring enterprise activities by RFID tags. *Information Sciences*, 195, 25-44.
- Ficco, M., Palmieri, F., & Castiglione, A., 2015. Modeling security requirements for cloud-based system

- development, *Concurrency and Computation: Practice and Experience*, 27(8), Jun. 2015, pp. 2107-2124
- Ianni, M., Masciari, E., Mazzeo, G. M., Mezzanzanica, M., Zaniolo, C., 2020. Fast and effective Big Data exploration by clustering. *Future Generation Computer Systems*, 102, 84-94.
- Bernardi, M. L., Cimitile, M., Martinelli, F., & Mercaldo, F., 2017 (June). A time series classification approach to game bot detection. In *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics* (pp. 1-11).
- Haq, A. U., Li, J. P., Khan, J., Memon, M. H., Nazir, S., Ahmad, S., ... & Ali, A., 2020. Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data. *Sensors*, 20(9), 2649.
- Gamache, R., Kharrazi, H., & Weiner, J. P., 2018. Public and population health informatics: the bridging of big data to benefit communities. *Yearbook of medical informatics*, 27(1), 199.
- Magnuson, J. A., & Dixon, B. E. (Eds.). 2020. *Public health informatics and information systems*. Springer Nature.
- Benis, A., Tamburis, O., Chronaki, C., & Moen, A., 2021. One Digital Health: a unified framework for future health ecosystems. *Journal of Medical Internet Research*.
- Pešić, S., Stanković, T., & Janković, D. 2009. Benefits of using OLAP versus RDBMS for data analyses in health care information systems. *Faculty of Electrical Engineering University of Banja Luka*, 56.
- Lu, J., & Keech, M., 2015 (September). Emerging technologies for health data analytics research: a conceptual architecture. In *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)* (pp. 225-229). IEEE.
- Gharehchopogh, F. S., Mohammadi, P., & Hakimi, P., 2012. Application of decision tree algorithm for data mining in healthcare operations: A case study. *International Journal of Computer Applications*, 52(6).
- Tayefi, M., Tajfard, M., Saffar, S., Hanachi, P., Amirabadizadeh, A. R., Esmacily, H., ... & Ghayour-Mobarhan, M., 2017. hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. *Computer methods and programs in biomedicine*, 141, 105-109.
- Tamburis, O., 2019 (November). Bridging the gap between process mining and des modeling in the healthcare domain. In *2019 E-Health and Bioengineering Conference (EHB)* (pp. 1-4). IEEE.
- Luzi, D., Pecoraro, F., & Tamburis, O., 2017. Appraising Healthcare Delivery Provision: A Framework to Model Business Processes. *Studies in health technology and informatics*, 235, 511-515.
- Mackenzie, J. S., & Jeggo, M. 2019. The One Health approach—Why is it so important?
- Choi, J., Cho, Y., Shim, E., & Woo, H., 2016. Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC public health*, 16(1), 1-10.