Data Standardization and Machine Learning Models for Histopathology

Abdullah M. Awaysheh

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Biomedical and Veterinary Sciences

Kurt L. Zimmerman, Chair
Jeffrey Wilcke
François Elvinger
Loren Rees
Weiguo Fan

Feb 17, 2017
Blacksburg, VA

Keywords: Data Standardization, Machine Learning, Histopathology, Inflammatory Bowel Disease, Alimentary Lymphoma

Data Standardization and Machine Learning Models for Histopathology

Abdullah M. Awaysheh

Abstract (academic)

Machine learning can provide insight and support for a variety of decisions. In some areas of medicine, decision-support models are capable of assisting healthcare practitioners in making accurate diagnoses. In this work we explored the application of these techniques to distinguish between two diseases in veterinary medicine; inflammatory bowel disease (IBD) and alimentary lymphoma (ALA). Both disorders are common gastrointestinal (GI) diseases in humans and animals that share very similar clinical and pathological outcomes. Because of these similarities, distinguishing between these two diseases can sometimes be challenging. In order to identify patterns that may help with this differentiation, we retrospectively mined medical records from dogs and cats with histopathologically diagnosed GI diseases. Since the pathology report is the key conveyer of this information in the medical records, our first study focused on its information structure. Other groups have had a similar interest. In 2008, to help insure consistent reporting, the World Small Animal Veterinary Association (WSAVA) GI International Standardization Group proposed standards for recording histopathological findings (HF) from GI biopsy samples. In our work, we extend WSAVA efforts and propose an information model (composed of information structure and terminology mapped to the Systematized Nomenclature of Medicine - Clinical Terms) to be used when recording histopathological diagnoses (HDX, one or more HF from one or more tissues). Next, our aim was to identify free-text HF not currently expressed in the WSAVA format that may provide evidence for distinguishing between IBD and ALA in cats. As part of this work, we hypothesized that WSAVA-based structured reports would have higher classification accuracy of GI disorders in comparison to use of unstructured free-text format. We trained machine learning models in 60 structured, and independently, 60 unstructured reports. Results show that unstructured information-based models using two machine learning algorithms achieved higher accuracy in predicting the diagnosis when compared to the structured information-based models, and some novel free-text features were identified for possible inclusion in the WSAVA-reports. In our third study, we tested the use of machine learning algorithms to differentiate between IBD and ALA using complete blood count and serum chemistry data. Three

models (using naïve Bayes, neural networks, and C4.5 decision trees) were trained and tested on laboratory results for 40 Normal, 40 IBD, and 40 ALA cats. Diagnostic models achieved classification sensitivity ranging between 63% and 71% with naïve Bayes and neural networks being superior. These models can provide another non-invasive diagnostic tool to assist with differentiating between IBD and ALA, and between diseased and non-diseased cats. We believe that relying on our information model for histopathological reporting can lead to a more complete, consistent, and computable knowledgebase in which machine learning algorithms can more efficiently identify these and other disease patterns.

Keywords: Data Standardization, Machine Learning, Histopathology, Inflammatory Bowel Disease, Alimentary Lymphoma

Data Standardization and Machine Learning Models for Histopathology

Abdullah M. Awaysheh

Abstract (public)

Computational models play an important role in supporting the decision making process. In some areas of medicine, decision-support models assist healthcare practitioners to make accurate diagnoses. In this work, we explored the application of computational techniques to distinguish between two diseases; inflammatory bowel disease (IBD) and alimentary lymphoma (ALA). These are common gastrointestinal (GI) diseases in humans and animals that share very similar laboratory findings. Because of these similarities, distinguishing between these two diseases can sometimes be challenging. In order to identify patterns that may help with this differentiation, we mined medical records from dogs and cats diagnosed with GI diseases. Since the pathology report is a key source of information for the diagnosis of these two diseases, in our first study we focused on its information structure. Others with similar interest have also examined reports of this type. In 2008, a group proposed standards for recording histopathological findings (HF) from GI biopsy samples. In our work, we extend the group's efforts and propose an information model (composed of information structure and terminology) to be used when recording histopathological diagnoses (HDX, one or more HF from one or more tissues). Next, our aim was to identify free-text HF not currently expressed in the standardization group's format that may provide evidence for distinguishing between IBD and ALA in cats. We trained computational models in 60 structured, and independently, 60 unstructured reports. Results show that unstructured information-based models using two computational models achieved higher accuracy in predicting the diagnosis when compared to the structured information-based models. As a result, novel free text features, which improved the performance of the structured reports, were identified. In our third study, we tested the use of computational models to differentiate between IBD and ALA using routine laboratory results. Three models were trained and tested on laboratory results from 40 Normal, 40 IBD, and 40 ALA cats. Diagnostic models achieved classification sensitivity ranging between 63% and 71%. These models can provide another non-invasive diagnostic tool to assist with differentiating between IBD and ALA, and between diseased and non-diseased cats. We believe that relying on our information model for histopathological

reporting can lead to a more complete, consistent, and computable knowledgebase for the identification of these two diseases.

Keywords: Data Standardization, Machine Learning, Histopathology, Inflammatory Bowel Disease, Alimentary Lymphoma

Dedication

To:

The Creator – In his name, the Most Merciful, the Most Compassionate

His Prophets – Who were not sent except as a Mercy to the worlds

My Mother; Siham – Who carried me, stayed up for me, and suffered so I live

My Father; Mamdouh – Who made a man out of me, taught me to be strong and wise

My Sisters; Rawan, Suzan, Nadeen, and Fatimah – My parents' gifts to the heart

My Uncles; Ahmad and Ghaleb – Who stood for me before passing away

My Wife; Dana Al Kurdi – The comfort of my eyes and my best friend

My Adviser; Kurt L. Zimmerman – The instructor, who held my hand and guided me through this work step by step

&

The innocent people killed in the crisis of Aleppo, Syria (2016)

Acknowledgements

This work would not be successful without the contribution of a variety of people. They contributed in many ways; educating me, supporting, and encouraging. I feel very thankful for every one of them and I would like to acknowledge them for helping me to achieve my goals. I would like to start by acknowledging my PhD committee members; Dr. Kurt L. Zimmerman, Dr. Francosi Elvinger, Dr. Loren Rees, Dr. Patrick Fan, and special thanks to Dr. Jeff Wilcke. Without all of you, your guidance, and your instructions this work would not be valuable.

I would like to thank my friends and colleges around the world generally and at Virginia Tech specifically, in no particular order, Dr. Hamzeh Al qublan, Dr. Mostafa Ali, Dr. Mohamed Mohamedin, Dr. Hassan M. Mahsoub, Dr. Iman Tavassoly, Dr. Jeff Alexander, Dr. Hana Alkhalidy, Enaam Kharabsheh, Huthaifa Ashqar, Mohammad Aljammal, and Ahmad Tbaileh for their support. I would like also to thank my friends in Jordan, Samer Aljbarah, Anas Athieh, Omar Mazaraa, Osama Al-jaloudi, Mohammad Awwad, Salam Sunjoq, Saif Saleh, Sahem Alkharabsheh, Mohammed Ghunaimat, Ehab Ghunaimat, Omar Jabrieh, and Ibrahim Faoury for their prayers and encouragements. Thanks to my uncles Mahmoud Jabrieh, Faisal and Hussain Awaysheh, and thanks to the Awaysheh's family at large.

I would like also to recognize my nieces and nephews, Layan, Adam, Aseel, Jad, Tameem, and Asia for bringing the joy to our family, and I wish them a life full of success.

Table of Contents

List of Figures

List of Tables

Chapter 1 - Literature Review

1.1 Diagnosis and Decision-Making

The medical diagnosis is the procedure in which clinicians identify the nature of the disease or the condition considering the patient's symptoms and signs. [1] Diagnosis is a process more than a single event, as defined by the Random House Kernerman Webster's College Dictionary; it is "*the process of determining by examination the nature and circumstances of a diseased condition*." [2]

Diagnostic reasoning as a process consists of some cognitive activities (called "cognitive heuristics" as far back as 1974 [3]) including problem solving (summarized by pattern recognition of different conditions and understanding relationships between different pathological components), decision-making, and judging under uncertainty. Human reasoning is the process of information processing characterized in a human system. In the diagnostic decision-making process, information processing is characterized by applying the studied way of making an accurate diagnosis and the human reasoning that is gained by the psychological experiments. [4] When examining a new case, diagnosticians generate and test hypotheses. The selected hypotheses would be the ones that best align with the actual diagnostic state as supported by available evidence. In a real world scenario, testing hypotheses requires collection of the patient's history, physical examination findings, and direct patient questioning. [5] Diagnosticians further test hypotheses using data from appropriate laboratory testing procedures.

There is more known about how clinicians should reason than how reasoning actually occurs. In one study, researchers constructed patients' stories and listened to clinicians while making a diagnosis loudly to optimize the process. They found that successful clinicians didn't generate more hypotheses; instead, successful clinicians interpreted data from the same list of hypotheses more accurately. [6] This finding supports the conclusion that clinicians' interpretation of results sometimes can be subjective. The subjectivity can be due to not following the diagnostic guidelines or to the lack of existing standards that maintain evidence-based practice. [7] The scenario is similar for pathologists; studies have shown subjectivity in pathologists' interpretation of findings, different pathologists can interpret similar findings differently. [8, 9] Similarly, clinicians form differential list of diagnoses before taking a closer look at the patient, sometimes prejudicing their conclusions. [10] For example, during a flu epidemic condition, clinicians assume

that a patient with a fever automatically has the flu and they do not consider other possibilities such as bacterial pneumonia or pyelonephritis. Therefore, it is important that the diagnostician reform their list of candidate diagnoses after testing their hypotheses and ensure that actual findings are taken into account. Reason for errors in diagnoses by experts can be summarized as: 1) they do not give attention to some details, 2) lack of knowledge about a particular scenario, or 3) they make invalid shortcuts as a result of making invalid assumptions or quick decisions. [5]

To help address these human limitions in medical decision making, computer applications were created to assist healthcare providers. Developers of these systems believe that these applications were not designed to replace health experts but to support their work. Specifically, these decision-support systems were envisioned to support an ad hoc decision-making process in the concept of "human-assisted computer diagnosis". [11]

Extensive work has been done to develop computer-based decision-support systems to assist clinicians in making an accurate diagnosis with the least invasive approach at the point of care. [12-15] Also, other diagnostic computer-based systems were developed to help in other matters, such as laboratory measurement applications, digital microscopy, or enhancing image quality. The overall goal behind all of these systems is to help achieve a more accurate decision with the minimum cost. The cost in this context can be summarized by the effect of a decision on patient health (such as the invasiveness of the diagnostic testing tool on the body and the hospitalization duration of the patient) and any financial cost associated with such a decision. Clinicians can make mistakes which can lead to greater costs. [16, 17] Making mistakes motivated the development of clinical guidelines that computers have shown to be able to establish. [18]

Clinical classification rules can generate guidelines for practitioners to follow when making a diagnosis. Experts in the field can develop these guidelines based on previously known findings (called rules from knowledge-based learning). In one study, [19] researchers have developed rules to help clinicians in assessing the risk of cardiac complications from non-cardiac surgery. In this approach, researchers put diagnostic weights based on statistical differences (previously tested hypotheses) of discriminant analyses for each clinical finding, and the total score is used to place the patient in a group with a defined probability of cardiac complications. Other applications, such as INTERNIST, [20] CADUCEUS, [21] and Quick Medical Reference [22] were developed using rules shaped by experts in the field. However, deep knowledge and previously tested hypotheses about findings or diseases do not always exist; this is either because

no effort has been made to discover the knowledge, or because of not having a clear pattern that can explain a disease or a distinction between diseases. This challenge raised the concept of "non-knowledge-based learning," which represents learning from previous instances instead of previous knowledge shaped by experts. The simplest example of this approach is matching findings of a new patient with findings from a previous one to make assumptions based upon those previous findings.

Textbooks and articles represent a good source of information for clinicians (for knowledge-based learning). However, most of clinicians' acquired knowledge comes from examining numerous patients and recognizing different patterns (through non-knowledge-based learning). [10] Clinicians apply what they learn from previous patients for the diagnosis of new. Mimicking this human cognitive learning process, the concept of machine learning grew within the computer science and artificial intelligence community; computers learn from previous examples and use that information to make future predictions. [23] Literature has examples of these computational models that emulate the neuronal structure of the human brain as well as training for short-term and long-term memories. [24] These computational, machine learning methodologies have been shown to be able to identify patterns to support one decision over another. [25, 26]

1.2 Applications of Computer-Based Diagnostic Decision-Support

Medical decision-making requires clinicians to act on patient care with less than all-possible knowledge regarding the patients' health status. Every day, healthcare practitioners must operate in this area of uncertainty while caring for their patients. To help manage this uncertainty, computer tools have been developed to assist healthcare practitioners in making decisions. [219]

Information systems have been widely used in diagnostic medicine for a variety of reasons. Some systems were created to improve information retrieval, some to analyze patients' records, and others developed as intelligent tools (using machine learning) to assist in decision-making. [28] The use of computational systems to assist in the diagnosis of diseases goes back to the 1950s. [220, 221] In a 1979 review by Shortliffe, Buchanan, and Feigenbaum [222] of computer-based clinical decision-support systems, authors used examples of medical computing paradigms to

assess strengths and limitations of clinical algorithms, clinical databanks, mathematical models in physiology, pattern recognition, Bayesian statistics, decision analysis, and artificial intelligence.

In the 1970s, researchers at the University of Pittsburg developed INTERNIST [223] as one of the first clinical decision-support systems. It was a rule-based expert system designed for the diagnosis of complex problems within internal medicine. In 1980s, INTERNIST medical knowledge base was recognized and INTERNIST became a base for other systems including CADUCEUS and Quick Medical Reference. [21, 22] In the late 1970s, MYCIN was developed as another rule-based expert system that was designed to diagnose and suggest treatments for different blood infections. MYCIN's knowledge base was modeled as a set of If-Then rules and certainty factors associated with every diagnosis were given. [222]

In the 1980s, RECONSIDER [224] was developed as a program for generating differential diagnoses given a list of patient attribute values. RECONSIDER's knowledge base was composed of a corpus of 3,262 disease definitions in the form of structured natural language text. DXplain is another system that was developed in the late 80s for the purpose of supporting the decision-making process and to diagnose patients with common diseases, such as anemia or heart failure. The system accepts a list of clinical manifestations and then proposes diagnostic hypotheses. DXplain knowledge was based on signs and symptoms data with the associated differential diagnosis. [225]

The previously illustrated systems were the core works that encouraged the development of many other systems. Since the 1990s, many studies were conducted to develop expert systems that can assist medical practitioners in making accurate diagnoses. One study [226] systematically assessed the use of different computer classification systems for the interpretation of electrocardiograms. Computer-based diagnoses were compared with clinical diagnoses independent from electrocardiogram information and compared with those of the cardiologists. Results showed that some but not all computer systems performed almost as well as cardiologists in identifying seven major cardiac disorders. In breast cancer, [227-231] the literature shows that machine learning (in specific, models of neural networks algorithms) can diagnose patients more accurately than experts (such as radiologists). Literature shows that such systems can achieve high accuracy in predicting malignancy tumors and distinguish them from benign ones. Authors of another study of breast cancer [232] showed that the neural network algorithm can be constructed with a small number of internal connections and still achieve a high accuracy. The same study

demonstrated a way of extracting rules from the generated neural network to be used in classification. Another study [233] also reviewed the literature for different tools and their application in screening for breast cancer. Authors found that most of the screening technologies utilized theoretical frameworks. In another study of breast cancer diagnosis systems, [84] authors reviewed different prediction models with a reduction in the number of features selected for the purpose of reducing the complexity of the developed models. In classifying cases into either malignant or benign breast cancer, although accuracy rates slightly decreased after reducing the number of features from 30 into one-dimension (using independent component analysis), the sensitivity rates increased when the one-dimension attribute was used with radial basis function neural network and support vector machine algorithms. Another study conducted in 2007 [234] reviewed the application of machine learning and computational systems in diagnosis and prediction of urological cancer behavior. The study found applications of such systems in areas of prostate, bladder, and kidney cancers. Authors concluded that machine learning has the flexibility and learning capability necessary to assist physicians in making decisions. Moreover, machine learning applications can be superior to standard statistical methods and allow for more flexibility. Authors suggested that understanding the basis of machine learning and its potential will allow these intelligent techniques to be developed further and implemented to assist physicians in the diagnosis and management of cancer cases. Many more recent studies have shown applications of decision-support systems in the diagnosis of different types of cancers, such as thyroid, [235] gastric, [236] cervical, [237] pancreatic, [238] brain tumors, [66, 239, 240] and lymphoma. [241]

For diseases other than cancer, a study in the 1990s assessed the use of machine learning algorithms to diagnose patients with different sports injuries. [242] The study showed a classification accuracy of up to 70% with the naïve Bayes being superior using fuzzy discretization as a discretization methodology of numerical attributes. In another study conducted in 1998, authors developed a system to assist physicians in the use of anti-infective agents, thus improving the quality of care. [243] The system was designed to give recommendations on anti-infective regimens and courses of therapy for particular patients. A prospective study of the scheme showed that its usage led to significant reductions in orders for drugs, excess drug usages, antibiotic-susceptibility mismatches, and costs.

In 2005, a review [25] examined studies of the effects of computer-based clinical decision-support systems in practitioner's performance (97 studies) and patient's outcome (52 studies). The

reviewed studies reported systems to be used as diagnostic tools, reminders, disease management systems, and treatment guidelines systems. The analysis showed that decision-support systems improved the practitioners' performance in 64% of the studies and improved the patients' outcome in 13% of studies. Diagnostic support studies included in the review covered areas such as mental disorders, [244-246] cardiac disorders, [247, 248] and abdominal pain. [249]

Using decision trees, naïve Bayes, and neural networks algorithms, a 2008 study developed an intelligent system to predict the likelihood of heart diseases. [64] The study showed that the most efficient model to predict patients with heart diseases appeared to utilize the naïve Bayes algorithm followed by neural networks and decision trees. Authors showed that the models established in their study were able to answer complex queries and give detailed information. Authors also suggested the application of text mining techniques to extract knowledge from unstructured data available in healthcare databases.

In 2012, one study tested the use of decision-support systems to diagnose many diseases, such as Jaundice in newborns. [250] Study utilized machine learning algorithms such as decision trees, neural networks, naïve Bayes, and others. Another study conducted in 2014 developed decision-support systems to distinguish between acute respiratory distress syndrome and cardiogenic pulmonary edema. Study used routinely available clinical data and an established clinical prediction score methodology. [251] In another 2014 study, [76] decision-support systems were developed to diagnose mild cognitive impairment, especially Alzheimer's disease. Using 10-fold cross-validation with the C4.5 algorithm, the model achieved a performance (80.2% sensitivity) that is higher than the ones achieved by other models developed using algorithms of support vector machine, Bayesian network, and neural networks. Therefore, authors concluded that a decision-support system using C4.5 algorithm would be useful to assist physicians efficiently in real clinical diagnosis for this disease. A more recent study (2016) tested the application of machine learning to predict coronary artery disease using data from non-invasive techniques. [252] Authors applied supervised machine learning algorithms and showed that the multi-layer perceptron neural networks algorithm achieved the best performance with 88.4% prediction accuracy.

In the 1980s, and for classifying text documents, experts had to define a set of rules encoding expert's knowledge on how to classify a document or extract knowledge from its content. In the 1990s many studies started to rely on machine learning to learn from free-text and

automatically apply what it has been learned to make classifications. A 2012 study [253] examined the use of clinical decision-support systems for cervical cancer screening by learning from a corpus of 49,293 Papanicolaou reports. The systems were developed to access patient records and generate patient-specific recommendations. In the study, the developed systems gave recommendations for 7 of 74 patients and identified two patients for gynecology referral that were missed by the physician. Authors highlighted the ability to learn from free-text and suggested the use of a more standardized reporting format for a more efficient information retrieval. Many other applications focused on learning from free-text for the purpose of supporting the diagnostic process. [38]

Decision-support systems using machine learning were not only trained using text records. Previous studies showed applications of machine learning on images from biopsies (to grade gastric atrophy according to Sydney system by neural networks algorithm), [254] colposcopy (to classify patients infected with papillomavirus by neural networks algorithm), [255] ultrasound (to classify breast tumor into benign or malignant by support vector machine algorithm), [256] histopathology (to diagnose basal cell carcinoma by unsupervised learning), [257] mammography (to classify breast tumors into benign or malignant by multi-classifiers [230] and to classify breast tissues into normal or abnormal by an improved neural networks algorithm [231]). More than text and images, researchers extended the application of decision-support systems to classify sound signals. In one study, [14] researchers developed a system to classify heart sound signals taken by stethoscope from patients with normal, pulmonary and mitral stenosis heart valve diseases. Using neural networks algorithm and a feature selection methodology, authors showed that such a system could reach up to 97% classification accuracy and up to 95% area under the ROC curve.

In veterinary medicine, many studies highlighted the applications of informatics to solve veterinary problems. [45, 48, 136] Studies suggested requirements that may advance veterinary informatics and reviewed options to overcome challenges surrounding some areas in veterinary medicine. One study [258] tested the application of expert systems such as neural networks, case-based expert systems, rule-based ES, and fuzzy logic on fish diseases. The study found that the use of such systems proved usefulness. Another study [259] developed a decision-support system to help veterinarians in many aspects, such as animal physical examination. A 2013 study [260] evaluated the ability of machine learning algorithms to extract syndromic information from laboratory tests submitted to a veterinary diagnostic laboratory. In the study, the naïve Bayes,

decision tree, and rule-based methodologies have shown to achieve relatively high performance. Another study used machine learning to assist in the diagnosis of canine visceral leishmaniasis. [261] Although many other studies showed that machine learning proved one way to solve veterinary problems, [262-264] our review of literature revealed that the application of machine learning in veterinary practice is limited.

1.3 Data Mining and Machine Learning for Decision-Support

1.3.1 Introduction

Data mining is an area of computer science which denotes the process of using computational techniques to convert raw data into useful information for knowledge discovery. Data mining involves the application of a variety of methods from computer science, statistics, and artificial intelligence. The main goal is extracting or "mining" knowledge from a data set (raw material). Most researchers currently agree that learning is the basis of intelligence. Machine learning is an area of artificial intelligence that uses algorithms to learn from existing data and uses what it has been learned to make classifications or future predictions. Based on the nature of the learning strategy, computer scientists classify machine learning into three main categories: supervised, unsupervised, and reinforcement learning. [27]

In supervised machine learning, the algorithmic model takes a set of instances as an input (called training instances), in which every instance belongs to a particularly known class (label) and presents values of a set of features. The model then outputs a classifier that can predict the classes of new instances given their features' values; the resulted model is associated with a particular sensitivity (called accuracy in information science) evaluating its performance. The unsupervised type of learning focuses on learning from instances without providing the actual classes; clustering algorithms are examples of this type. The reason behind not giving the classes can be due to the lack of knowledge about them, or it can be for the purpose of desiring a new way of classification structure. In the third type of learning; reinforcement, the learning algorithm is not informed about the actions that must be taken, but instead, it tests the reward of different actions to reach the most rewarding one, in which the reward can be either immediate, or after series of actions. [23]

In medicine, a variety of machine learning models have been designed to assist health care practitioners in the decision-making process. Clinical decision-support systems have been

developed to assist with drug dosing, health maintenance, diagnosis, and with other non-hospital based purposes, such as patient support or educational systems. [28, 29] As an example, in one study of the application of different machine learning algorithms, researchers were able to discriminate metastatic brain tumors from gliomas with 96.9% classification accuracy using values of features extracted from magnetic resonance images. [30] In the same study, machine learning techniques were used to discriminate between low-grade and high-grade neoplasms and achieved 94.5% classification accuracy.

1.3.2 Data Preparation

Machine learning extracts knowledge from computable information. However, the process of learning is dependent on the efficient representation of this information. It has been shown that the quality of data has a high impact on the machine learning process. [31, 32] Therefore, data preparation for a future processing usually consumes the bulk of the effort in the knowledge discovery process.

As discussed previously, developers of a supervised machine learning scheme represent the input of information as a set of instances (examples), and these instances are to be used as the basis for training and the reference when testing the developed prediction models. In most of the preprocessing situations, instances are expressed in an independent and non-redundant format. In every relation (table), rows are instances and columns are features (attributes) which represent variables to be recorded for every instance. The feature values can be numeric, nominal, images, sounds, or others. However, data are transformed into numerical values so they can be analyzed computationally.

Data Transformation and Free-Text Pre-processing

Data transformation is the method of converting the data values from the source format into an input format to be processed. Successful data mining and machine learning involves more than selecting a learning algorithm. Most algorithms used in learning have various parameters and different value settings. Some of these settings control the ability to transform the input into a computable, sometimes more suitable format. The right group of settings has been a focus of research for the past decade. Studies have shown that prediction results can be improved if developers selected the suitable value settings. For example, discretization is one of the most common data transformation methodologies to transform continuous values into discrete

counterparts. Discretization has received a great deal of attention in the data mining community. [33-35] Moreover, there are classifiers that are called discrete classifiers, which take discrete values to achieve quickly and efficient performance; a decision tree is an example of a discrete classifier that is very commonly used. [36]

Free-text is a common type of information encountered for use with machine learning algorithms. Various transformation methodologies have been developed to prepare free-text documents for such use. A "bag of words" is an example of a very common transformation method in which each document is represented by a set of words (called features) that are extracted from its text. [37, 38] The frequency of occurrences of all word within the bag and across other bags can be used as quantitative measurements to represent the content of each document. A previous study [39] tested the effect of different techniques of transforming free-text into a vector of numerical descriptors. The study reviewed techniques such as Term Frequency (frequency of a term in a document), Term Frequency with Inverse Document Frequency (in which the terms that appear in all documents are overlooked), and Term Frequency with Inverse Class Frequency (in which terms are weighted according to their relationship with the document categories or classes). Results of the study showed that the Inverse Document Frequency and Inverse Class Frequency weighting factors have shown to either improve the performance when learning from free-text or keep it not changed.

Different methodologies have been studied to optimize extracting the right set of words that can distinguish the class of each document. Among these methodologies is text tokenization, also called text segmentation, which is simply dividing the text into meaningful units presented as words, sentences, or topics. [40] Text segmentation focuses on extracting alphabetical content from the text corpus and ignores any non-alphabetical contents. Words stemming is another technique that has shown to have a good impact in the free-text analysis. [39, 41, 42] In stemming, words are reduced to their stems or roots so words with similar roots may be gathered together. Stemming usually results in the removal of derivational affixes, an assumption that similar roots represent words that are synonyms. In a 2015 study, [32] researchers found that stemming reduced the set of features from 9793 to 936.

Besides the simple and easy way of searching free-text through word indexing, another approach to searching and retrieval is through the use of taxonomies (concept indexing). This methodology has the advantage of searching for related concepts (such as synonyms) without

having to declare every one of them explicitly. Taxonomies will allow users to not only browse and search free-text documents but also comprehensively retrieve information by traversing a path of categories; the recognition of cognitive terms without having to recall keywords. In "bag of words" representation, taxonomy words can be added to the set of words extracted from the text and can be used to substitute some of them as shown previously. [43] Moreover, taxonomy categories can introduce new information that is not conveyed in the free-text within the corpus. For example, when experts use the concept class "inflammation" to substitute different terms that convey the same meaning (such as "infiltration" or "inflammatory"), they can classify different inflammation relevant documents under one category. One study [44] represented 91% of specimens and critical pathological findings from 275 pathology reports using Unified Medical Language System (UMLS) codes. Study used a hierarchy for the intention of achieving more efficient document retrieval and analysis. In veterinary medicine, previous studies used a variety of text mining techniques for evidence-based practice, [45] veterinary syndromic surveillance, [46] disease detection, [47] and extracting enteric syndrome cases. [48]

Another factor that has shown to improve learning from free-text is excluding the list of words which are not dependent on the class or topic. A list of words called *stop words* is excluded from being considered in the "bag of words" to improve the tokenization accuracy and speed. [49] For example, literature has shown that the word "*the*" appears in almost all documents, accounts for a large percentage of the words, and has no relevance to any particular category when learning. Therefore, it is of a great advantage to exclude the word "the" before preparing the data input. A recent study, [32] showed that *stop words* counted for 9% of the extracted text features and, therefore, hamper the effort of learning by machines and introduce more complexity.

Feature Selection

Many data preparation methodologies have shown to improve success when applying machine learning techniques. The success is not limited to improve the performance accuracy on the training data set; the best choice for a real world scenario is not necessarily the one that improves the accuracy on the training dataset. Previous studies showed that optimizing the performance on the training data may cause a problem of overfitting, in which the trained model is too specific to the data that was used for training and it is less representative of the real world's scenario. To maximize the success of the machine learning models, developers select only a subset of the dataset

as an input in a process called feature selection. Feature selection is another commonly used pre-processing methodology.

In learning from free-text, the number of words mined from a text with a moderate length can easily reach 10,000 words. [38, 50] There is tendency to gather as much data as possible, "more is better." However, in machine learning this may or may not be true. There is evidence that machine learning models do better when subsets of features are selected for learning. [50-52] Irrelevant or redundant data can affect the performance of the computational models negatively and add more noise. The effect can be an increase in running time, introducing more complexity (so the results are hard to interpret), and overfitting the training data set. [53]

A study conducted in 2010 compared the effect of threshold-based feature selection techniques on three different datasets. Authors evaluated the effectiveness of different feature selection methodologies using 8 different metrics; area under the curve (AUC), precision-recall plot (PRC), default F-measure (corresponds to a decision threshold value of 0.5), best F-measure (the largest value of F-measure when varying the decision threshold value between 0 and 1), default geometric mean, best geometric mean, default arithmetic mean, and best arithmetic mean. [54] The study found that the performance of the software quality models they developed either improve or remain unchanged despite the removal of 96% of the features; in fact, they found that in 95% of the cases the results were improved. In a 2013 study for detecting the effect of feature selection methodologies on the predicting accuracy of machine learning algorithms, [55] authors tested the application of five algorithms; *Trees.J45*, *Bayes.BayesNet*, *Functions.Logestic*, *Meta.Bagging*, and *Rules.ZeroR*. Authors found that decreasing the number of features to one-fourth (25%) would either improve the classification performance or keep it the same. They also concluded that the performance is dependent on the selected subset of features, and the overall detection behavior is independent of the number of selected features.

Filter and wrapper are two different methodologies of selecting the relevant features prior to classification. These two techniques have shown to improve the performance of the prediction models significantly. [52] In the filter approach, any irrelevant features that are not correlated to one of the class labels are filtered out based on some general characteristics of the training data, such as statistical dependencies. This approach is considered faster than wrapper because it acts independently from the induction algorithm (the algorithm that evaluates each subset). However, this approach tended to select a high number of features that would require a threshold for a subset

selection. [56] In wrapper, an optimal subset of features is selected and tailored to a particular induction algorithm. Unlike the filter approach, wrapper uses an induction algorithm as part of the evaluation process of different feature subsets. The wrapper algorithm searches for features that best suit the machine learning algorithm used for prediction, and this makes it more computationally expensive than the filter model. [57]

In searching the attribute space, several methods have been developed. The majority of these methods search for the set of attributes that is most likely to predict the class. Greedily searching of the space is one of the typical methods. In *Greedy*, space is searched either forward or backward, by adding or removing a single attribute at each step. The forward direction starts with no attributes and then adds one at a time. The backward elimination starts with all attributes and deletes one at a time. In Greedy, features adding or removing stops when the performance drops. Another method for searching is Best First, [58] in which the search does not stop when the performance of the new set drops. Instead, in *Best First* the searching method keeps looking for new subsets while keeping the old ones in the memory, then it sorts subsets by their performance measurements. The *Best First* method is considered to be more computationally expensive as a result of the memory and time requirements. However, this methodology assesses the entire space (unless otherwise specified) of attributes to guarantee the selection of the best subset and it has shown to be very effective. [59] One study conducted in 2011 investigated the use of machine learning techniques in classifying brain neoplasms based on magnetic resonance imaging. [30] In the study, authors selected a subset of attributes in learning using two filtering methods and a wrapper approach in a combination of three different searching algorithms (*Best First*, *Greedy Stepwise* and *scatter*). Authors found that the feature selection method of using wrapper in combination with the *Best First* search algorithm achieved the highest average classification accuracy when learned by the K-nearest neighbor algorithm. Combinations of different pre-processing and transformation techniques have shown to improve the performance of the developed models significantly. In a 2011 study of the diagnosis of patients with Alzheimer's disease, [60] classification models were built to distinguish between healthy and diseased patients. Results showed that the models' performances improved by applying techniques like bootstrap resampling, spatial normalization, smoothing, intensity normalization, multivariate image analysis based on principal component analysis and Fisher discriminant analysis.

1.3.3 Classification and Prediction Schemes

This section reviews the three classification (prediction) algorithms: naïve Bayes (NB), C4.5 decision trees (DT), and artificial neural networks (ANN) and their usage as supervised machine learning models.

Naïve Bayes

The naïve Bayes classifier is considered to be simple and efficient. It is derived from Bayes' theorem which can be used to predict the class of new events using probabilities learned from training. But unlike Bayesian classifier, the naïve Bayes allows for a computationally inexpensive learning while keeping a reasonable classification performance. The computational complexity in the Bayesian algorithm lies in calculating dependencies not only between the training attributes and the class that is to be predicted, but also within the training attributes. The naïve Bayes model assumes conditional independencies between the random attributes that are used for training. [61] Based on the Bayes theorem the selected class (or predicted one) should be the one that maximizes $P(X_i \mid E) = P(X_i)\, P(E \mid X_i) / P(E)$, where X_i represents the ith class, E represents the test example, $P(A \mid B)$ denotes the conditional probability of A given B, and the probabilities are estimated from the training sample. If n represents the number of attributes that are independent given the class, then $P(E \mid X_i)$ can be decomposed into the product $P(v_1 \mid X_i) \ldots P(v_n \mid X_i)$, where v_k is the value of the kth attribute in the example E. Therefore, based on the naïve Bayes the chosen class should be the one that maximizes:

$$P(X_i \mid E) = \frac{P(X_i)}{P(E)} \prod_{k=1}^{n} P(v_k \mid X_i) \qquad (2.1)$$

Theoretically, the naïve Bayes model should achieve the best performance when trained on attributes that are independent of each other, and decline in performance as attributes become more dependent. However, previous studies examining systems trained on attributes that were not independent have shown this decline is not always the case. For example, in one study of classifying schizophrenia patients using electroencephalogram data, [62] the naive Bayes classifier performed better than other classifiers that take dependencies into account, such as Adaboost, random forest, and support vector machine. A recent study [63] of prostate cancer staging prediction using clinical data showed that the naïve Bayes classifier is competitive and it achieved a performance equivalent to the one achieved by more complex classifiers, such as neuro-fuzzy, Fuzzy C-Means, support vector machine, and artificial neural network. Other studies that tested

the usage of a naïve Bayes algorithm in machine learning showed similar findings in areas like Heart Disease diagnosis, [64] neonatal jaundice diagnosis, [65] and brain tumor classification. [66]

C4.5 Decision Tree

Decision tree algorithms represent another classification methodology that Hunt and his co-authors described in the 1960s. [67] Decision tree represents each instance using a collection of attributes, and each instance belongs to one of the so-called exclusive classes. The decision tree algorithm uses a training set of instances labeled with classes to develop a mapping from attribute values to classes. In the developed map, each attribute represents a dimension and each instance becomes a point on a description space. The decision tree algorithm then splits the description space into regions in which each one is associated with a particular class. The created map can then be used to predict the class of a new instance given its set of attribute values. A classification tree creates a hierarchical data structure composed of nodes; the first tree node on the tree is called the root node, and child nodes are referred to as internal nodes. Each of these internal nodes carries a particular test used to classify instances, e.g. "Is the patient male or female?" For each possible outcome of a test, a child node is present. In cases of discrete attributes, an attribute A has h possible outcomes $A = d_1, \ldots, A = d_h$, where $d_1, \ldots, d_h$ are known A attribute values. In a case of a continuous attribute, there are two possible outcomes; $A \leq t$ and $A > t$, where t is a value of a threshold that is to be determined at the node. The nodes at the end of the tree are termed leaf nodes and they are used to identify the class which will be the target, e.g. "Patient with cancer". Decision tree classification techniques are embodied in packages, such as CART, [68] ID3, [69] and C4.5 that was developed as an extension to the earlier ID3 algorithm. [70] The C4.5 algorithm generates a decision tree by following the following methodology: Attribute a^* is selected as a root node from the set of attributes A for having the maximum information gain (a synonym for Kullback–Leibler divergence that originated in 1951 [71]). Then, the sample L is divided into subsets L_v in which each subset contains a different value v of attribute a^* values V_{a^*} and each value is represented by an edge on the tree that leads to a new node. If in any subset L_v only one class d exists, the new node becomes a leaf node and carries the subset L_v class d. Otherwise, the attribute with the next highest information gain is selected and linked to the new node by an edge. It is split recursively further on each subset of attribute values.

The C4.5 decision tree classifier is distinguished for its user-friendly structure. It provides a tree that is easy to apply at the point of practice. Unlike the naïve Bayes classifier, decision trees

do not assume independencies between attributes, making them applicable in many scenarios. Previous studies have shown that the C4.5 classifier can perform well in many areas, such as traffic management, [59] marketing, [72] health insurance industry, [73] gene identification, [74, 75] and medical diagnosis. [64, 76, 77]

Artificial Neural Network

In the 1940s, a study reported computational models that represent biological neural networks mathematically. [82] Since then, modeling neural networks mathematically became the area of interest for many groups, some of which desired to understand the biological processes in the brain and nervous system, while others focused on the application of neural network models in artificial intelligence. In artificial intelligence, artificial neural network algorithms were developed to mimic the structure and the function of the biological neural networks. Scientists have shown that neural networks can mathematically model neuron biological structure, memory function, and knowledge storage and retrieval. [78, 79]

In artificial neural network algorithms, knowledge is acquired through a learning process called back propagation and stored within the interconnection strength between neurons (called nodes). Developers can build such algorithms out of single-layer neurons (called single-layer perceptron) or neurons arranged in multiple layers (multi-layer perceptron). While single-layer perceptron algorithms classify instances into categories using direct relationships between inputs and outputs, they cannot be used to solve every problem. Some sets of instances cannot be divided into distinct categories by a simple linear relationship. Multi-layer perceptron algorithms were developed to deal with more complex scenarios. In Multi-layer perceptron, more than one layer of neurons is used and non-linear relationships can be produced. For illustration, if $x_1, \ldots, x_n$ are input layer attributes and y is the class attribute (output layer) of dataset D, and $w_1, \ldots, w_n$ are the weights of the relationships between the input and the output nodes (adjusted based on information from previous instances) as shown in **Figure 1.1**, then the single-layer perceptron algorithm classifies new instances given their input attribute values, where θ is a threshold value designated to make a classification to a particular output node as the following:

$$\text{output} = \begin{cases} 1, & if \sum_{i=1}^{n} w_i x_i > \theta \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

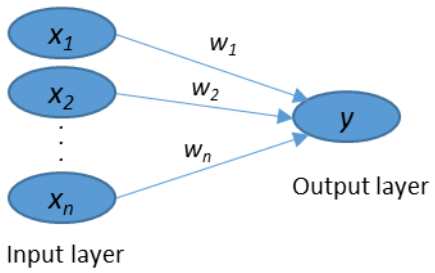

Figure 1.1 Diagram showing a single-layer perceptron network. x_{1-z} are input attributes, y is the output attribute (class), w_{1-n} are the weights of the relationships between the input and the output nodes.

One the other hand, a multi-layer perceptron approach introduces a one or more layers (called hidden layers) that introduce a role of complexity in dealing with the non-linear learning patterns of many examples. For illustration, a hidden layer that is composed of nodes h_1, …, h_n can be introduced to the single-layer perceptron approach (**Figure 1.2**).

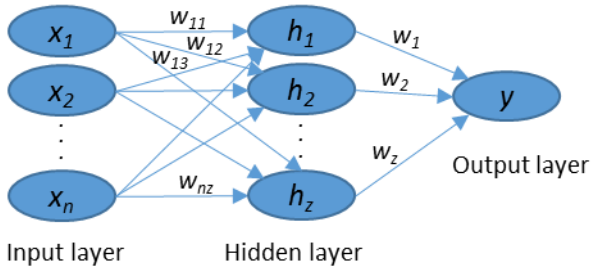

Figure 1.2 Diagram shows a multi-layer perceptron network. x_{1-n} are input attributes, h_{1-z} are internal nodes, y is the output attribute (class), w_{11-nz} are the weights of the relationships between the input and the internal nodes, w_{1-z} are the weights of the relationships between the internal and the output nodes.

Unlike single-layer perceptron and the direct relationships between the input and the output, multilayer-perceptron has the ability of training the new hidden nodes by adjusting their weights. Adjusting the weights of the multi-perceptron nodes is done most commonly through a process called backpropagation. [80] Backpropagation is a method that was introduced in 1986 to repeatedly adjust the weights of the connections in the network for the goal of minimizing the difference between the actual output vector of the network and the desired one. [81]

The fact that "multi-layer perceptron uses a hidden layer of nodes" leads to the generation of classification results that are hard to understand and justify. However, the added complexity of this hidden layer has increased the ability of this approach to solve scenarios that decision tree or naïve Bayes algorithms might not be able to adequately address. Multi-layer neural network

algorithm has shown to be very useful in research and applications for its ability to approximate solutions and learn from complex datasets. [64, 65, 82-84]

Datasets for Training and Testing

In non-knowledge based systems, where the learning is based on previous instances, a dataset is used to train the classifiers. After being trained, the classifier will be tested for its ability to predict the class given its attributes' values. Usually, experts evaluate the classifier's performance on a dataset that was not used in training; this is to test the classifier's performance on new examples. There are two common ways of splitting any dataset into training and testing sets. The first technique is to divide based on a particular percentage, such as randomly dividing a dataset into 60% used for training and 40% used for testing. Some studies used 50% for training and 50% for testing, [85, 86] others chose to split into 70% training and 30% testing. [87] Authors of a particular study challenged the classifier after using less than 10% of instances for training. [88] These researchers found that training a naïve Bayes algorithm on 10% of instances and testing with 90% resulted in performance with 95.20% accuracy and an F-Measure value of over 97%. Although researchers did not agree upon an exact division, most train using 60-80% of the instances and test on the reminder.

Cross-validation is another common approach to split a dataset into training and testing. Unlike simple random split, cross-validation facilitates the usage of all instances in training but by performing multiple rounds of divisions. One round of cross-validation involves dividing a dataset into complementary subsets, train the model on one, and test on the other. In an alternate approach, the subset used for training is held out for testing and the testing set will be used for training. Commonly multiple rounds of cross-validation are performed and results are averaged to reduce variability. [89, 90] There are several cross-validation approaches that can be performed based on the number of folds (rounds) selected. K-fold cross-validation is the most common approach while K is the number of folds to be created with K=10 being the most common selection. Leave-p-out cross-validation is another approach, in which p is equal to the number of instances used in testing and the rest used in training. Leave-one-out cross-validation is a common type of selection which represents k-fold cross-validation taken to its extreme, with k equals to the number of instances in the dataset. Typically, after training an algorithm on a dataset, its performance in classifying new instances is reported using values of sensitivity, specificity, accuracy, and area under the receiver operator characteristics (ROC) curve.

1.3.4 Data Standards

In 2010, the Department of Health and Human Services assigned rules and regulations through its electronic health record incentive program regarding medical recording. These rules required all US health care providers and institutions participating in Medicare and Medicaid programs to implement an electronic health record system by 2015. [91] With increased adoption of electronic health records and the incentive to promote the "meaningful use" of health information technology by the Department of Health and Human Services, [92, 93] the volume of the health information recorded in the organizations' repertories will grow exponentially. Therefore, healthcare in the United States will become an information-intensive industry. Moreover, a 2015 report by the Council for Affordable Quality Healthcare (a non-profit alliance of health plans and trade associations that intends to streamline the business of healthcare) showed that healthcare providers and commercial health plans could save up approximately $8.5 billion annually by transiting from manual to electronic processes. [94] This transition from paper to electronic record systems provided an opportunity for these health records to be used in secondary activities, such as research, public health, and statistics. Reporting of clinical or pathological findings in systematized manner has been shown to improve the efficiency of these secondary uses. [95] However, it is difficult to achieve improvement in secondary activities without corresponding standards for information storage and retrieval. It is also challenging to compare practice patterns and outcomes without a common language. The integration and maturation of clinical terminologies to standardize information recording will result in more accurate and detailed clinical information which facilitate more efficient and effective health care delivery and information recall. [96, 97]

In 1994, the Board of Directors of the American Medical Informatics Association (AMIA) [98] recommended specific approaches to standardization in the areas of the patient, provider, and site of care identifiers. Standards covered computerized health care message exchange, medical record content and structure, and medical codes and terminologies. In the 1994 study, the AMIA Board of Directors suggested coding systems for several subject domains, such as the World Health Organization (WHO) drug record codes for drugs, the International Classification of Diseases (ICD) and the Systematized Nomenclature of Medicine - Clinical Terms (*SNOMED-CT*) for diagnoses, Health Level 7 (HL7) for messaging, and others.

In healthcare messages exchange, HL7 International is a standard developing organization that provides framework and standards for the exchange, integration, sharing, and retrieval of electronic health information. The HL7 Reference Information Model (RIM) is an example of an object model created to provide conceptual standards which identify the life cycle that a message or groups of related messages will carry. [99] HL7 was recommended by the Board of Directors at AMIA to serve within-institution transmission of orders, clinical observations, and clinical data; admission, transfer, and discharge records; and charge and billing information. Moreover, in 2004, the Consolidated Health Informatics (CHI) in the U.S. suggested the use of HL7 vocabulary standards for demographic information, units of measure, immunizations, and clinical encounters. CHI also suggested HL7's Clinical Document Architecture standard for text-based reports. [100, 101] Other standards were recommended by the CHI as well, such as the Laboratory Logical Observation Identifier Name Codes (LOINC) to standardize the electronic exchange of laboratory test orders and drug label section headers.

The ICD (which is maintained by WHO) has been one of the most common classification systems for epidemiology, health management, and clinical purposes. ICD is used by physicians, nurses, and other health providers to classify diseases and other health problems. ICD codes are currently being used by over 50 countries to improve consistency in recording patients' symptoms and diagnoses for the purpose of reimbursement and clinical research. [102]

In 1965 the College of American Pathologists (CAP) published the Systematized Nomenclature of Pathology (SNOP) to describe morphology and anatomy. In 1975, led by the efforts of Dr. Roger Cote, CAP expanded SNOP to create the Systematized Nomenclature of Medicine (SNOMED). In 2000, CAP created a new logic-based version of SNOMED called SNOMED-RT. Over that the same time period, Dr. James Read developed the Read codes which evolved into Clinical Terms Version 3 (CTV-3) under the British National Health Service. In 2002, the CTV-3 and SNOMED-RT were combined to create SNOMED-CT a joint development project of the British National Health Service (NHS) and CAP. SNOMED-CT is now the intellectual property of the International Health Terminology Standards Development Organization (IHTSDO) which determines global standards for medical terms for clinical findings, procedures, body structures, organisms, qualifier values, and more. [103, 104] With more than 344,000 unique medical concepts and 900,00 synonyms or alternate descriptions, SNOMED-CT is considered to be the most comprehensive clinical vocabulary that can be used to report and

represent medical information. The components of SNOMED-CT define standard terms for health processes, indicate relationships between processes, and describe these processes through the use of synonyms. The terms for health processes are organized in multiple hierarchies with different degrees of granularity to provide flexibility in data recording and representation. [105, 106] SNOMED-CT's current approach adopts a foundation based in descriptive logic (DL) that has many advantages, such as establishing formal semantics for SNOMED-CT's assertions and suggesting a formal syntax. It also provides a basis for understanding expressiveness and computational complexity. [107] SNOMED-CT concepts meanings are derived from their arrangement in the hierarchy and from axioms that connect them across hierarchies. The connections across different hierarchies provide definitions, which classify concepts into sufficiently defined or non-sufficiently defined.

SNOMED-CT usage has grown with an increase in access facilitated by the National Library of Medicine's U.S. public license free access agreement through the Unified Medical Language System (UMLS). [105] Moreover, in 2004, the CHI recommended the use of SNOMED-CT as the standard for diagnoses and problem lists, anatomy, and procedures. [101] SNOMED-CT provides also an extension mechanism, in which organizations can create the concepts' classes and descriptions that fulfill their needs and still fit into the SNOMED-CT framework. The logic-based framework of SNOMED-CT also increases its abilities by allowing for post-coordination (i.e., the creation of new concepts' classes by combining existing ones). SNOMED-CT User [104] and Starter [108] Guides explained the concepts' classes and relationships that developers can combine to post-coordinate clinical findings at the point of care. Post-coordination can be straightforward, and it can be complex, this potential complexity falls on the shoulders of coders and implementation experts who would have to code each single concept by more than one SNOMED-CT concept code.

A 40 years review of SNOMED-CT [109] showed no difference in SNOMED-CT usage among various medical specialties. The study showed that SNOMED-CT is broadening beyond pathology to be used to cover nursing, cancer, cardiology, primary care, gastroenterology, HIV, orthopedics, nephrology, and anesthesia. A more recent review [110] has shown that SNOMED-CT is reportedly used in more than 50 countries and its implementation has been steadily increasing. However, only a few publications showed SNOMED-CT utilization in operational settings. In data capture and implementation, SNOMED-CT has been very commonly used to

represent terms from documents and forms (by concepts mapping), as an interface terminology (where users browse through SNOMED-CT terms and descriptions), and as an automatic indexing system for clinical text (through the application of natural language processing). Previous studies [110-114] facilitated mapping SNOMED-CT to over 40 standard terminologies with ICD (versions 9 and 10) being the most common one. SNOMED-CT medical domains of application reached over 39, with problem list/diagnoses, nursing, drugs, and pathology being the most common ones. [110] These studies showed that mapping SNOMED-CT to other terminologies identified gaps in concepts and synonyms that need to be incorporated within SNOMED-CT to improve its completeness.

As in human medicine, veterinarians see the need for standards when recording animal patients' information. [115-119] A good example of standardizing recording of veterinary information is the use of SNOMED-CT as a standard nomenclature for pathology reporting. [120] The Veterinary Terminology Services Laboratory (VTSL) at Virginia Tech provides a browser to query SNOMED-CT content. VTSL maintains an extension of SNOMED-CT that integrates content that is considered to be "veterinary only" (is not applicable in human medicine). [121] Because veterinary medical terminology is shared with human medicine, the VTSL browser presents an integrated view of the international release (core) of SNOMED-CT combined with the veteirnary extension.

SNOMED-CT Applications

Searching medical records using traditional information retrieval methodologies (such as keywords filtering or text-mining) is a challenge. There is need for a more robust data retrieval methodology, such as semantic searches, that would be enabled by using terminologies such as SNOMED-CT. A previous study [122] has shown that SNOMED-CT's concept-based approach can effectively deal with the two types of queries: keyword mismatch and specialization/generalization (granularity). In another study, [123] authors used SNOMED-CT searching terms to recall incidents from a database relating to neuromuscular blockade in anesthesia. In their study, authors concluded that a keyword search is only as good as the terms selected, but the use of SNOMED-CT reveals more incidents. Another retrospective study [124] examined the feasibility of using electronic laboratory and admission-discharge-transfer data to retrieve cases with *Clostridium difficile* infection. Date was recalled from 44 hospitals. The study showed that 40 hospitals sent test results recorded using LOINC codes and five hospitals sent

SNOMED-CT codes. Study showed cases retrieval using 25 LOINC and 20 SNOMED-CT codes. In their conclusion, authors emphasized the need for a widespread adoption of standard vocabularies to facilitate public health use of electronic data. Other studies used SNOMED-CT to retrieve terms and classify cases from information systems for chronic diseases, [125] infectious symptoms, [126] and cancer morphologies, [42, 127, 128] such as breast and prostate malignancies. [129] Although SNOMED-CT has shown to be very comprehensive and useful for many domains, its usage as knowledge base ontology has been a challenge. [130] SNOMED-CT has shown the ability to facilitate semantic interoperability (i.e., the ability of two parties to exchange data with unambiguous and shared meaning [131]) aspects, such as classifying terms by meaning, placing terms with similar meanings into one hierarchy, and enriching them with synonyms or descriptions. Logic-based concept definitions were then introduced and consolidated in SNOMED-CT. [132] As a result, the current SNOMED-CT constitutes a blend of diverging and even contradicting architectural principles. Other challenges related to using SNOMED-CT as knowledge base ontology have been reported. [130] In histopathology, a recent study [133] investigated the use of SNOMED-CT to represent histopathological diagnoses. Authors of that study investigated the ability to represent diagnostic tissue morphologies and tissue architectures described within a pathologist's microscopic examination report for 24 breast biopsy cases. Results found that 75% of the 95 diagnostic statements could be represented by one valid SNOMED-CT pre-coordinated and 73 SNOMED-CT post-coordinated expressions. Authors concluded that development of the SNOMED-CT model and content to cover microscopic examination of histopathological tissues is required in order for SNOMED-CT to be effective in surgical pathology knowledge capture.

In veterinary medicine, reference and terminology standards are core components of the informatics infrastructure. In 2002, The American Veterinary Medical Association (AVMA) endorsed the use of SNOMED, HL7, and LOINC as the official informatics standards in veterinary medicine. [134] This subset of SNOMED-CT contains more than 5,600 clinical signs and diagnostic terms commonly used in companion animal practice in North America. The subset was drawn from both the SNOMED-CT core and Veterinary Extension. With the availability of different standard terminologies for veterinary medicine, several studies tested the implementation, evaluation, and application of these standards. One study [136] highlighted some of the needs and challenges surrounding different areas of veterinary informatics. Authors

suggested the application of LOINC, HL7, and SNOMED-CT for laboratory tests and messaging would promote the use of evidence-based practice in veterinary medicine. SNOMED-CT contains veterinary concepts in addition to a great number of anatomical structures that are for non-humans. A previous study [120] examined SNOMED-CT coverage of veterinary clinical pathology concepts taken from textbooks and industry pathology reports. Authors of the study found that SNOMED-CT representation of veterinary clinical pathology content was limited (ranging from 45 – 48 % of good matching). However, the missing and problem concepts were relevant to a relatively small area of terminology. In their findings, authors concluded that revising and enriching some of the SNOMED-CT contents will optimize it for veterinary clinical pathology applications. Another study [137] examined organizing microbiome data from mammalian hosts using biomedical informatics methodologies and the Foundational Model of Anatomy (FMA [138, 139]) and SNOMED-CT. The study concluded that researchers can use the current biomedical infrastructure in organizing microbiome data from animals, therefore, bring a new source of knowledge to facilitate testing comparative biomedical hypotheses pertaining to human health. Another study [140] tested the applications of utilizing standardized technologies, such as HL7 and SNOMED-CT in veterinary hospital information system. The study conducted a survey that showed an improvement in the management of hospital data and the retrieval of useful clinical information when following standards.

Finally, a consistent coding system should be adopted in order to track animal health data on a national or international scale and facilitate public health use of human and animal records. [106, 141] Such systems should be interoperable and should receive a widespread adoption to be beneficial. Recommendations for perfect standard terminologies to improve interoperability and receive a widespread adoption were previously studied. [142] The study covered aspects regarding terminology content, concept orientation, polyhierarchy structure, granularity and others.

Structured Reporting

Electronic health data comprises various data types from structured information, such as patient blood test results or diagnoses coded, to unstructured (free-text) data such as admission notes, description of patient physical examination, and histopathological descriptions of tissue specimens. This variation in data types and the use of a free format presents a challenge for data integration and reuse. When it comes to processing unstructured information, such as free-text

documents (which have shown to account for 80% of the data in business and health record systems [143]), the task becomes more time and effort consuming. Previous studies showed techniques to learn from free-text and used them widely to identify special patterns and features that can be of interest. One study [128] showed a way of converting scanned images of pathology reports into a free-text report so researchers can analyze information and classify reports computationally into cancer or non-cancer categories. However, applying text mining techniques ignores some relationships between information stored in the text. A more efficient computational approach (such as natural language processing) requires human guidance for information to be extracted and interpreted properly from free-text.

While unstructured (free-text) reports can be more fluid and explicit of findings, [144] they are not easily converted into a structured, computable data format. Recording standards will help ensure consistent reporting of patients' information (such as physicians' observations, diagnoses, and treatment) to improve data integration and uses. Standards have to be followed when recording patients' information in order to automate computer-based technological tools which can assist in a variety of hospital processes and secondary uses, such as reminders, procedures, and decision-making activities. Such computer-based systems minimize errors in medical practice, control costs, and billing matters. [145]

Clinicians tended to value flexibility and efficiency in reporting information, while informaticists (users of information in secondary activities) value the information structure. Structured reports can require more time than unstructured ones at the point of care, however, structured reporting saves time and efforts when it comes to the secondary activities. More importantly, supporting the automation of the secondary activities would have a good impact at the point of care (the primary activity). One study [146] showed that information could be found four times faster if recorded using structured format than using an unstructured one. In information coverage, a previous assessment across the two formats showed that 10% of the studied features could not be found in unstructured clinical records. [147] Another study [148] showed that structured, well-organized data can influence care decisions by offering a complete picture of the patient state to the decision maker. The well-organized display of information provided to physicians has shown in another study [149] to reduce physicians ordered tests by 15%, therefore reducing the cost and invasiveness of diagnostic tests on patients. Although structured reports have shown to have some advantages over the unstructured ones, structured reports may lack the

expressiveness that the narrative format has. This lack of expressiveness can make it challenging to gain physician's acceptance of the structured format. However, a partial capture of information in a structured manner has been more readily accepted. For example, a 1988 study [147] showed that the best the institution could do was to capture some value or items from physicians in a structured format with the rest being coded as free-text. However, the right (computationally sufficient and medically representative of a particular scenario) amount of information to be structured presents a challenge. Literature has shown that the right balance between report expressiveness and report structure is an optimization problem. [97, 150, 151] In order to achieve the optimum balance, a study conducted in 2008 [152] suggested the use of a *structured narrative* design, in which unstructured text and coded data are fused into a single model. A *structured narrative* report is marked up with standardized codes used within a narrative structure to enable an efficient retrieval of information while keeping the report on its original expressive shape. Natural Language Processing (NLP) and relying on post-hoc text processing can be used then to identify fine structure. However, NLP can't assure the full understanding achieved using the structured format and its optimum accuracy. In other words, analyzing and understanding yes/no based attributes is straightforward for machines, and it is more challenging for machines to draw a similar conclusion if the information was reported using the natural language. Previous studies have discussed the effects of structured versus unstructured (free-text) data on the retrieval and analysis of records. [153, 154] Other studies examined the use of structured reporting formats in reporting variety of laboratory information, such as clinical notes, [155] history and physical examination, [152] magnetic resonance imaging, [156] fracture risk assessment, [157] endoscopy, [150] radiology, [158] and cytology. [159]

In the field of pathology, the pathology report is the formal communication link between pathologists and clinicians. An unstructured format has been commonly used for reporting pathological findings. Standardizing the terminology used within the pathology reports has been evaluated in many studies. In one study [160] researchers surveyed 96 veterinary clinical pathologists. The study showed that 68 unique terms were used to express probability or likelihood of cytological diagnoses, among them was: "possible," "suggestive of," "consistent with," "most likely," "probable," and more. With including a numerical interpretation of the terms, the study drew a good probability diagram by translating each term into numerical expression.

In other studies, different formats used in reporting pathology information have shown to have an effect on the clarity of the message transferred. A 1991 study assessed the completeness of 1,565 upper endoscopy structured reports in comparison to 360 free-text reports. [161] The study found that structured reports had less missing data rate (18%) comparing to free-text reports (45%) and contained 60% more relevant information. Another study of radiology reports concluded that free-text reports contained 13% to 75% of forensically relevant findings comparing to 100% coverage by structured reports. [162]

Since 1986, the College of American Pathologists (CAP) has been working to establish guidelines for the assessment and recording of pathological examinations. [163] In 1992, one study by CAP [164] examined 15,940 pathology reports for colorectal carcinoma from 532 laboratories. The study found that standardized or checklist structure of pathology reports was the only factor significantly associated with increased likelihood of providing complete or adequate pathology information. In the same year, a study was published declaring the adequacy needed within structured surgical pathology reports. [165] Following the release of the 1992 CAP standards, and using them as references, a study conducted in 1993 proposed standardized surgical pathology forms for reporting major tumor types. The study included recommendations for reporting histologic grade, patterns of growth, local invasion, vascular invasion, and others for many body sites. [166] A study published in 1994 [167] described the use of standardized, synoptic surgical pathology reports to improve the accuracy with which critical information can be obtained consistently and easily regardless of the institution of origin. The authors of the study reviewed variables found to be relevant to the clinical management of patients, proposed a standardized reporting format, and showed a simple query for extracting information about tumors from a database for patients with right partial mastectomies.

Another study in 1996 by CAP [168] reviewed over 8,300 lung carcinoma cases from 464 institutions. The study assessed the adequacy of lung carcinoma surgical pathology reports in covering gross and microscopic findings. Authors examined the presence or absence of 23 descriptors. Results showed that a standardized or a checklist format was used in 21% of the cases and is associated with increased likelihood of recording nine of the 23 descriptors. Authors proposed a standardized reporting format for resected bronchogenic carcinomas and recommended the use of a standardized report or a checklist reports. In 1999, for the intention of assisting

pathologists in making diagnoses, CAP published a checklist reporting format to be used when recording specimens' information for patients with carcinoma of renal tubular origin. [169]

A 1999 study published guidelines for a more precise morphological evaluation and reporting of coeliac disease diagnoses. [170] Authors developed histopathological guidelines for the evaluation of biopsy specimens for animals with coeliac disease as well as a structured report composed of a checklist of findings to be recorded. Researchers also developed other reporting standards for cancer pathology; among them is a report that was released by the Center for Disease Control and Prevention (CDC) on reporting of pathology protocols for colon and rectum cancers. [171] Although the structured data format seemed to be the appropriate reporting structure, a 2000 study [172] highlighted a communication gap between pathologists and surgeons as a result of the unfamiliarity with the structured representation; they emphasized the need for a more complete, clear, and standardized format.

In 2008, CAP established the Diagnostic Intelligence and Health Information Technology Committee with a mission "to advance the implementation of the CAP Cancer Checklists using health information technology." In 2009, the number of available checklists was expanded to 55. Extensible markup language versions of the checklists became available for the intention of improving compatibility of checklists to a variety of platforms and to integrate checklist records into information systems. [163, 173] On the same year, another study [174] showed that implementing some of the CAP standards resulted in improvements in rates of synoptic reporting and completeness of cancer pathology reporting. The study showed a significant increase in completeness rates when structured reports were used to report cancer of prostate, colon\rectum, lung, and breast.

A 2010 CAP study [175] evaluated the frequency with which surgical pathology cancer reports contain all the scientifically validated elements by the American College of Surgery Commission on Cancer. Authors evaluated a total of 2125 cancer reports generated by 86 different institutions. The study found that institutions in which practitioners routinely used checklists reports reported all required elements at a higher rate (88 %) than those did not use checklists (34 %). The same study also showed that the institutions that had a system in place to track errors, which can only be achieved by computable reports, reported all required elements at a higher rate (88 %) when compared with those that did not have such a system (68 %). A more recent study, conducted in 2014 [176] found expressions of uncertainty in 35 % of 1,500 human surgical

pathology reports. In the same study, authors surveyed 76 clinicians seeking a percentage of certainty to be given to expressions of uncertainty reported within diagnoses. Results of the previous two studies showed a wide variation in the certainty percentages assigned by the clinicians; this suggests subjectivity introduced by the non-standardized pathology report and a significant source of miscommunication.

Although a "diagnosis" is the major product of pathology reports, a previous study found that the pathologist's product is not simply the correct diagnosis; there is pertinent and credible information that is useful in addressing patient care needs in other areas of the report, and structured reports can help ensure its capture. [177] A 2016 study [178] of microscopic assessments of rectal tumour specimens examined the effect of structured reporting formats on the data recorded. Authors found that structured reporting dictation template improves data collection and may reduce the subsequent administrative burden when evaluating rectal specimens.

In veterinary pathology reports, the unstructured format has been the traditional format used in recording histopathological findings. In 2005, the World Small Animal Veterinary Association (WSAVA) GI International Standardization Group undertook the responsibility of standardizing the histologic evaluation of the gastrointestinal tract of dogs and cats. [116] In 2008, the group proposed standards for the assessment of microscopic findings from GI biopsy samples. [115] The standards were found to minimize variation among pathology reports of microscopic findings in the GI tract. The group established criteria for quantifying and ranking key endoscopic, histologic observations from the gastric body, antrum, duodenum, and colon. These criteria provided specific numerical intervals to categorize the severity (mild, moderate, and marked) of different observations. In 2010, the same group published endoscopic, biopsy, and histopathologic guidelines for the evaluation, and partial reporting of GI inflammation in companion animals. [117] However, the WSAVA has not proposed standardization for recording the histopathological diagnosis, and there has been no formal assessment of the effect of the WSAVA structured format on information capture.

Therefore, in healthcare, to ensure reporting required elements, increase data interoperability across different systems, and standardize healthcare practice, clinicians and pathologists are recommending using structured reporting formats. [93, 166, 179-184]

1.4 Inflammatory Bowel Disease and Alimentary Lymphoma in Dogs and Cats

1.4.1 Introduction

Alimentary lymphoma (ALA, a type of neoplasia) is a common GI disease in companion animals. [185] Lymphoma is the most common feline malignancy, and the GI tract has been the most common location for this disease. [186] ALA is associated with the infiltration of lymphoid cells and can affect the upper or lower GI tract, liver, or pancreas. In ALA, the lymphoid cells infiltrate the lamina propria of the mucosa and possibly efface the epithelial lining, submucosa, tunica muscularis, and serosa. [185, 187, 188] As with many other types of neoplasia, the primary cause of lymphoma is often unclear. However, previous studies highlighted some disease risk factors. One study showed that 16 of 24 cats tested with ALA were positive for *Helicobacter heilmannii* in their biopsy samples. [189] Exposure to cigarette smoke has been suggested as another risk factor, in which cats in smoking households have been shown to have a 2.4-fold increase risk of lymphoma. [190, 191] Intestinal lymphoma in cats has been associated with retrovirus infection. Feline leukemia virus (FeLV) has shown to increase the risk of lymphoma by 60-folds in cats. [186] Feline immunodeficiency virus (FIV) is another virus that is known to be a risk for lymphoma, specifically the B-cell type. [192] The most common symptoms of the disease in dogs and cats are anorexia, weight loss, vomiting and/or diarrhea. [185, 187, 188]

Clinically, ALA can be classified into three major categories; Intermediate/high-grade, low-grade, and large granular. [193] In most cats, the GI lymphoma is histologically characterized by small-to-medium sized T-lymphocytes and occurs at the mucosal lining level of the intestine. By immunophenotype representation, most of the cells (90%) that comprise the low-grade and large-granular lymphocytic lymphomas are of a T cell-type, while the intermediate/high-grade can consist of either B or T cells. [194, 195]

Previous studies [196-201] have shown some common laboratories findings in dogs and cats with intestinal lymphoma. The complete blood count (CBC) often showed a monocytosis, neutrophilia, and a nonregenerative anemia that is due to chronic disease. Serum biochemistry data showed common abnormalities such as increased activity of hepatocyte leakage enzymes (ALT etc.). Hypoalbuminaemia is the most common blood chemistry abnormality in dogs and cats with intestinal lymphoma (especially the intermediate-to-high grade type). [191, 199, 202] This finding occurs as a result of the loss of albumin through the intestinal wall.

Accurate diagnosis of different ALA types and non-neoplastic disorders is essential. It is important to identify patients with lymphoma so that proper treatment can be provided. The histopathological evaluation of biopsy specimens under the microscope has been the reference standard for this purpose.

Inflammatory bowel disease (IBD, lymphoplasmacytic enteritis/gastritis) is another common GI disorder in companion animals. [185, 200] Clinically, IBD is very similar to ALA with the most common IBD symptoms being anorexia, weight loss, vomiting and/or diarrhea. [185, 204] A study conducted in 2014 [205] showed that serum 25(OH) vitamin D concentrations were significantly lower in cats with IBD and small cell ALA compared to healthy and hospitalized cats. The same study highlighted a positive correlation between serum albumin and 25(OH) D concentrations in the two diseases.

Similar to the ALA, the histopathologic evaluation is the reference standard for the diagnosis of IBD. Histologically, IBD in most cats is characterized by the accumulation of lymphocytes and plasma cells, and less often neutrophils, eosinophils, or macrophages, in the lamina propria of the stomach and small intestine. [204, 206] Unlike ALA, in IBD the lymphocyte infiltration does not extend significantly beyond the lamina propria. However, the histologic distinction between these two disorders is not always clear as shown in previous studies. [207, 208] These studies have also suggested there is disagreement among pathologists' interpretation of endoscopic findings and the diagnosis of ALA or IBD. [207, 209]

1.4.2 Distinguishing between ALA and IBD

Inflammation and neoplasia are common diseases that can affect the GI tract. The severity and duration of inflammation may be a causative agent in a variety of neoplasms including lymphoma. This relationship is not new; it has been shown in the literature that the longer inflammation lasts, the higher the risk of cancer. [210] In fact, IBD has been shown to mimic ALA in cats. [202] The continued recruitment and proliferation of lymphoid cells in an environment that is rich with inflammatory cells, growth factors and associated DNA damage provides a microenvironment that promotes the risk of cancer. [210]

Due to the interconnection of inflammatory and neoplastic disorders, the diagnosis of these two processes and the distinction between chronic inflammation and cancer tends to be challenging. [211] The challenge is a result of similarities in clinical and pathological

representation, especially when the lymphoma is of a small-cell type. [9, 212] Therefore, other studies have explored the use of supplemental diagnostics for differentiation.

One study [207] suggested the evaluation of full-thickness biopsies instead of relying upon endoscopic samples. However, the procedure of obtaining a full-thickness biopsy (surgically) is considered to be more invasive in which the animal is exposed to a higher morbidity risk due to surgical complications and is associated with a higher financial cost for the owner. [207] Cats with intestinal lymphoma tended to have a monoclonal population of B or T lymphocytes. Polymerase chain reaction (PCR) is a technique that can be used to assist in identifying lymphoma. PCR can assist in the detection of a clonal expansion of B or T lymphocytes that favors a diagnosis of lymphoma, while a mixed population of lymphoid cells supports a diagnosis of inflammatory bowel disease. One study suggested immunohistochemistry analysis be used to look for B-cell and T-cell markers to help distinguish intestinal lymphoma from inflammation. [194] Moreover, molecular clonality techniques are used to identify unique expression components that can assist in the diagnosis of lymphoma. [213] The expression of T-cell receptor-γ repertoire is one of the variables that was developed to detect T cells clonality status in feline intestinal lymphoma. [214] One study of clonality assessment of mucosal T-cell ALA revealed clonal rearrangement of T-cell receptor-γ in 78% (66/84) of cats. An assessment of transmural T-cell lymphoma showed clonal rearrangement of T-cell receptor-γ in 79% (15/19) in cats with ALA. In the assessment of B-cell ALA, clonality assessment revealed clonal rearrangement of feline immunoglobulin heavy chain repertoire in 50% of cats with lymphoma. Clonality testing has also been used to help classify different subtypes of lymphomas. [187] Another study investigated the usage of ultrasonography to distinguish between ALA and IBD. [215] The study found that cats with ALA were more likely to have a thickening of the muscularis propria (detected by ultrasonography) when compared with IBD or normal cats. A 2008 review highlighted the challenges of identifying ALA cases from cases with non-malignancies. [212] Other studies evaluated use of cell expression of CD3, CD79a, BLA.36, class II molecules of the major histocompatability complex, [12, 194, 216] immunosignature of serum antibodies (identified lymphoma dogs with 88% accuracy), [217] fecal α(1)-PI concentrations (as a marker of GI protein loss), [218] and thymidine kinases enzymes as potential tumor markers. [199] In one study, [13] immunolabeling for the critical lymphocyte survival factor Bcl-2 was performed on small intestinal biopsy sections. The study determined that the expression of Bcl-in was significantly higher in ALA than it was in cats with IBD. A more

recent study [217] developed an immunosignature assay and applied it to spontaneous canine lymphoma as proof of concept. The study showed an ability to predict lymphoma in 42 dogs with 97% accuracy. Although many studies have demonstrated a variety of techniques and guidelines to assist in the diagnosis of ALA and IBD, [115-117] research continues for more reliable, accurate, and less invasive diagnostic procedures.

1.5 References

1. Martin, E.A., *Concise medical dictionary*. 8th ed. Oxford paperback reference. 2010, Oxford ; New York: Oxford University Press. 832 p.

2. Random House Kernerman Webster's College Dictionary. [cited 2016 Aug 3]; Available from: http://www.kdictionariesonline.com/DictionaryPage.aspx?ApplicationCode=18&DictionaryEntry=diagnosis&SearchMode=Entry&TranLangs=18

3. Tversky, A. and D. Kahneman, *Judgment under Uncertainty: Heuristics and Biases.* Science, 1974. **185**(4157): p. 1124-31.

4. Simon, H.A., and Allen Newell, *Human problem solving: The state of the theory in 1970.* American Psychologist, 1971. **26(2)**: p. 145.

5. Mengel, M.B., W.L. Holleman, and S.A. Fields, *Fundamentals of clinical practice*. 2nd ed. Vol. Chapter 10 2002, New York, N.Y.: Kluwer Academic/Plenum Publishers. xxx, 837 p.

6. Arthur S. Elstein, L.S.S., Sarah A. Sprafka, *Medical problem solving a ten-year retrospective.* Evaluation & the Health Professions. **13(1)**: p. 5-36.

7. Garb, H.N., *Studying the clinician : judgment research and psychological assessment*. 1st ed. 1998, Washington, DC: American Psychological Association. x, 333 p.

8. Wilcock, B., *Endoscopic biopsy interpretation in canine or feline enterocolitis.* Semin Vet Med Surg (Small Anim), 1992. **7**(2): p. 162-71.

9. Willard, M.D., et al., *Interobserver variation among histopathologic evaluations of intestinal tissues from dogs and cats.* J Am Vet Med Assoc, 2002. **220**(8): p. 1177-82.

10. Sox, H.C., M.C. Higgins, and D.K. Owens, *Medical decision making*. 2nd ed. 2013, Chichester, West Sussex, UK: John Wiley & Sons. xvi, 347 p.

11. Miller, R.A., *Why the standard view is standard: people, not machines, understand patients' problems.* J Med Philos, 1990. **15**(6): p. 581-91.

12. Briscoe, K.A., et al., *Histopathological and immunohistochemical evaluation of 53 cases of feline lymphoplasmacytic enteritis and low-grade alimentary lymphoma.* J Comp Pathol, 2011. **145**(2-3): p. 187-98.

13. Swanson, C.M., et al., *Expression of the Bcl-2 apoptotic marker in cats diagnosed with inflammatory bowel disease and gastrointestinal lymphoma.* J Feline Med Surg, 2012. **14**(10): p. 741-5.

14. Uguz, H., *A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases.* J Med Syst, 2012. **36**(1): p. 61-72.

15. Abbass, H.A., *An evolutionary artificial neural networks approach for breast cancer diagnosis.* Artif Intell Med, 2002. **25**(3): p. 265-81.

16. Levinson, W. and P.M. Dunn, *A piece of my mind. Coping with fallibility.* JAMA, 1989. **261**(15): p. 2252.

17. Lee, T.S., G. Lansbury, and G. Sullivan, *Health care interpreters: A physiotherapy perspective.* Aust J Physiother, 2005. **51**(3): p. 161-5.

18. Bates, D.W., et al., *Reducing the frequency of errors in medicine using information technology.* J Am Med Inform Assoc, 2001. **8**(4): p. 299-308.

19. Palda, V.A. and A.S. Detsky, *Perioperative assessment and management of risk from coronary artery disease.* Ann Intern Med, 1997. **127**(4): p. 313-28.

20. Pople, H.E., Myers, J., & Miller, R. , *DIALOG: A Model Of Diagnostic Logic For Internal Medicine.* IJCAI 1975. **4**: p. 848-855.

21. Miller, R., *INTERNIST-l/CADUCEUS: Problems Facing Expert Consultant.* Meth. Inform. Med, 1984. **23**(1): p. 9-14.

22. Miller, R.A., et al., *The INTERNIST-1/quick medical REFERENCE project—Status report.* Western Journal of Medicine, 1986. **145**(6): p. 816.

23. Witten, I.H., et al., *Data mining practical machine learning tools and techniques*. 2011, Elsevier/Morgan Kaufmann: Amsterdam. p. xxxiii, 629 p.

24. Kourtzi, Z. and A.E. Welchman, *Adaptive shape coding for perceptual decisions in the human brain.* J Vis, 2015. **15**(7): p. 2.

25. Garg, A.X., et al., *Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review*. JAMA, 2005. **293**(10): p. 1223-38.

26. McEvoy, F.J. and J.M. Amigo, *Using machine learning to classify image features from canine pelvic radiographs: evaluation of partial least squares discriminant analysis and artificial neural network models.* Veterinary Radiology & Ultrasound, 2013. **54**(2): p. 122-126.

27. Bishop, C. *Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn.* New York: Springer, 2007.

28. Berner, Eta S. *Clinical decision support systems*. New York: Springer Science+ Business Media, LLC, 2007.

29. Kononenko, I., *Machine learning for medical diagnosis: history, state of the art and perspective.* Artificial Intelligence in medicine, 2001. **23**(1): p. 89-109.

30. Zacharaki, E.I., V.G. Kanas, and C. Davatzikos, *Investigating machine learning techniques for MRI-based classification of brain neoplasms.* International journal of computer assisted radiology and surgery, 2011. **6**(6): p. 821-828.

31. Crone, S.F., S. Lessmann, and R. Stahlbock, *The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing.* European Journal of Operational Research, 2006. **173**(3): p. 781-800.

32. Sharma, D. and S. Jain, *Evaluation of Stemming and Stop Word Techniques on Text Classification Problem.* 2015.

33. Dougherty, J., Ron Kohavi, and Mehran Sahami. *Supervised and unsupervised discretization of continuous features*. in *Machine learning: proceedings of the twelfth international conference*. 1995.

34. Zelič, I., et al., *Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries.* Journal of Medical Systems, 1997. **21**(6): p. 429-444.

35. Butterworth, R., et al., *A greedy algorithm for supervised discretization.* J Biomed Inform, 2004. **37**(4): p. 285-92.

36. Canisius, S., Antal van den Bosch, and Walter Daelemans. *Discrete versus probabilistic sequence classifiers for domain-specific entity chunking*. in *Proceedings of the Eighteenth Belgian-Dutch Conference on Artificial Intelligence (BNAIC-2006)*. 2006. Namur, Belgium.

37. Needham, C.D., *Organizing knowledge in libraries; an introduction to information retrieval*. 2d rev. ed. A Grafton book. 1971, London: Andre Deutsch. 448 p.

38. Sebastiani, F., *Machine learning in automated text categorization.* ACM computing surveys (CSUR), 2002. **34**(1): p. 1-47.

39. Jouhet, V., et al., *Automated classification of free-text pathology reports for registration of incident cases of cancer.* Methods of information in medicine, 2012. **51**(3): p. 242.

40. Choi, F.Y. *Advances in domain independent linear text segmentation*. in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. 2000.

41. Savoy, J., *A stemming procedure and stopword list for general French corpora.* Journal of the Association for Information Science and Technology, 1999. **50**(10): p. 944.

42. Spasic, I., et al., *Text mining of cancer-related information: review of current status and future directions.* Int J Med Inform, 2014. **83**(9): p. 605-23.

43. Wu, H., M.D. Gordon, and W. Fan, *Collective taxonomizing: A collaborative approach to organizing document repositories.* Decision Support Systems, 2010. **50**(1): p. 292-303.

44. Schadow, G. and C.J. McDonald, *Extracting structured information from free text pathology reports.* AMIA Annu Symp Proc, 2003: p. 584-8.

45. Anholt, R.M., et al., *The application of medical informatics to the veterinary management programs at companion animal practices in Alberta, Canada: a case study.* Prev Vet Med, 2014. **113**(2): p. 165-74.

46. Furrer, L., Küker, S., Berezowski, J., Posthaus, H., Vial, F., Rinaldi, F. . *Constructing a syndromic terminology resource for veterinary text mining*. in *Proc. Conf. Terminol. Artif. Intell* 2015.

47. Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., Dufour, B., *Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web.* Computers and Electronics in Agriculture, 2016. **123**.

48. Anholt, R.M., et al., *Mining free-text medical records for companion animal enteric syndrome surveillance.* Prev Vet Med, 2014. **113**(4): p. 417-22.

49. Reynar, J.C., *Topic segmentation: Algorithms and applications.* IRCS Technical Reports Series, 1998. **66**.

50. Blum, A.L., Pat Langley, *Selection of relevant features and examples in machine learning.* Artificial intelligence, 1997. **97(1)**: p. 245-271.

51. Piramuthu, S., *Evaluating feature selection methods for learning in data mining applications.* European journal of operational research, 2004. **156**(2): p. 483-494.

52. Hall, M.A., and Lloyd A. Smith *Feature subset selection: a correlation based filter approach*. 1997.

53. Langley, P. *Selection of relevant features in machine learning*. in *Proceedings of the AAAI Fall symposium on relevance*. 1994.

54. Wang, H., Taghi M. Khoshgoftaar, and Jason Van Hulse. *A comparative study of threshold-based feature selection techniques*. in *Granular Computing (GrC), 2010 IEEE International Conference on. IEEE*. 2010.

55. Noureldien, N.A., R.A. Hussain, and A. Khalid. *The Effect of Feature Selection on Detection Accuracy of Machine Learning Algorithms*. in *International Journal of Engineering Research and Technology*. 2013. ESRSA Publications.

56. Liu, H., and Lei Yu, *Toward integrating feature selection algorithms for classification and clustering.* IEEE Transactions on knowledge and data engineering, 2005. **17(4)**: p. 491-502.

57. Kohavi, R., and George H. John, *Wrappers for feature subset selection.* Artificial intelligence, 1997. **97(1)**: p. 273-324.

58. Korf, R.E., *Linear-space best-first search.* Artificial Intelligence, 1993. **62**(1): p. 41-78.

59. Williams, N., S. Zander, and G. Armitage, *A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification.* ACM SIGCOMM Computer Communication Review, 2006. **36**(5): p. 5-16.

60. Merhof, D., et al., *Optimized data preprocessing for multivariate analysis applied to 99mTc-ECD SPECT data sets of Alzheimer's patients and asymptomatic controls.* J Cereb Blood Flow Metab, 2011. **31**(1): p. 371-83.

61. Pedro Domingosm, M.P. *Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier*. in *Proc. 13th International Conference on Machine Learning*. 1996.

62. Laton, J., et al., *Single-subject classification of schizophrenia patients based on a combination of oddball and mismatch evoked potential paradigms.* J Neurol Sci, 2014. **347**(1-2): p. 262-7.

63. Cosma, G., et al., *Prediction of Pathological Stage in Patients with Prostate Cancer: A Neuro-Fuzzy Model.* PLoS One, 2016. **11**(6): p. e0155856.

64. Palaniappan, S. and R. Awang. *Intelligent heart disease prediction system using data mining techniques*. in *2008 IEEE/ACS International Conference on Computer Systems and Applications*. 2008. IEEE.

65. Ferreira, D., A. Oliveira, and A. Freitas, *Applying data mining techniques to improve diagnosis in neonatal jaundice.* BMC Med Inform Decis Mak, 2012. **12**: p. 143.

66. Tsolaki, E., et al., *Fast spectroscopic multiple analysis (FASMA) for brain tumor classification: a clinical decision support system utilizing multi-parametric 3T MR data.* Int J Comput Assist Radiol Surg, 2015. **10**(7): p. 1149-66.

67. Hunt, E.B., J. Marin, and P.J. Stone, *Experiments in induction*. 1966, New York,: Academic Press. xi, 247 p.

68. Breiman, L., *Classification and regression trees*. Wadsworth statistics/probability series. 1984, Belmont, Calif.: Wadsworth International Group. x, 358 p.

69. Quinlan, J.R., *Induction of decision trees*. 1986.

70. Salzberg, S.L., *C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993.* Machine Learning, 1994. **16**(3): p. 235-240.

71. Kullback, S. and R.A. Leibler, *On information and sufficiency.* The annals of mathematical statistics, 1951. **22**(1): p. 79-86.

72. Karim, M. and R.M. Rahman, *Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing.* 2013.

73. Araújo, F.H., A.M. Santana, and P.d.A.S. Neto, *Using machine learning to support healthcare professionals in making preauthorisation decisions.* International Journal of Medical Informatics, 2016. **94**: p. 1-7.

74. Liao, Z., et al., *In Silico Prediction of Gamma-Aminobutyric Acid Type-A Receptors Using Novel Machine-Learning-Based SVM and GBDT Approaches.* BioMed Research International, 2016. **2016**.

75. Jowkar, G.-H. and E.G. Mansoori, *Perceptron ensemble of graph-based positive-unlabeled learning for disease gene identification.* Computational Biology and Chemistry, 2016. **64**: p. 263-270.

76. Zhang, X., et al., *Ontology driven decision support for the diagnosis of mild cognitive impairment.* Comput Methods Programs Biomed, 2014. **113**(3): p. 781-91.

77. Ichikawa, D., et al., *How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach.* Journal of Biomedical Informatics, 2016. **64**: p. 20-24.

78. McCulloch, W.S. and W. Pitts, *A logical calculus of the ideas immanent in nervous activity.* The bulletin of mathematical biophysics, 1943. **5**(4): p. 115-133.

79. Rosenblatt, F., *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychol Rev, 1958. **65**(6): p. 386-408.

80. Haykin, S.S., et al., *Neural networks and learning machines*. Vol. 3. 2009: Pearson Upper Saddle River, NJ, USA:.

81. Williams DR, H.G., *Learning representations by back-propagating errors.* Nature 1986. **323**: p. 533-536.

82. Wasserman, P.D. and T. Schwartz, *Neural networks. II. What are they and why is everybody so interested in them now?* IEEE Expert, 1988. **3**(1): p. 10-15.

83. Venkatasubramanian, V., and King Chan, *A neural network methodology for process fault diagnosis.* AIChE Journal, 1989. **35(12)**: p. 1993-2002.

84. Mert, A., et al., *Breast Cancer Detection with Reduced Feature Set.* Comput Math Methods Med, 2015. **2015**: p. 265138.

85. Gramatica, P., P. Pilutti, and E. Papa, *Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling.* J Chem Inf Comput Sci, 2004. **44**(5): p. 1794-802.

86. Pankratz, V.S., et al., *Model for individualized prediction of breast cancer risk after a benign breast biopsy.* Journal of Clinical Oncology, 2015: p. JCO. 2014.55. 4865.

87. Kwon, O.H., W. Rhee, and Y. Yoon, *Application of classification algorithms for analysis of road safety risk factor dependencies.* Accid Anal Prev, 2015. **75**: p. 1-15.

88. Xhemali, D., C.J. Hinde, and R.G. Stone, *Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages*, 2009.

89. Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *Ijcai*, 1995.

90. Dietterich, T.G., *Machine-learning research.* AI magazine, 1997. **18**(4): p. 97.

91. DHHS, *Medicare and Medicaid Programs; Electronic Health Record Incentive Program; Final Rule*, 2010.

92. CDSPS, *Key Capabilities of an Electronic Health Record System*. Institute of Medicine of the National Academies, Washington, DC, 2003.

93. van der Meijden, M.J., et al., *Development and implementation of an EPR: how to encourage the user.* Int J Med Inform, 2001. **64**(2-3): p. 173-85.

94. CAQH, *2016 CAQH INDEX: A Report of Healthcare Industry Adoption of Electronic Business Transactions and Cost Savings*. [cited 2017 Jan 20]; Available from: http://www.caqh.org/sites/default/files/explorations/index/report/2016-caqh-index-report.pdf

95. Elkin, P.L., Brett E. Trusko, Ross Koppel, Ted Speroff, Daniel Mohrer, Saoussen Sakji, Inna Gurewitz, Mark Tuttle, and Steven H. Brown, *Secondary use of clinical data.* Stud Health Technol Inform, 2010. **155**: p. 14-29.

96. Chute, C.G., S.P. Cohn, and J.R. Campbell, *A framework for comprehensive health terminology systems in the United States: development guidelines, criteria for selection, and public policy implications. ANSI Healthcare Informatics Standards Board Vocabulary Working Group and the Computer-Based Patient Records Institute Working Group on Codes and Structures.* J Am Med Inform Assoc, 1998. **5**(6): p. 503-10.

97. Jensen, P.B., L.J. Jensen, and S. Brunak, *Mining electronic health records: towards better research applications and clinical care.* Nat Rev Genet, 2012. **13**(6): p. 395-405.

98. BDAMIA, *Standards for medical identifiers, codes, and messages needed to create an efficient computer-stored medical record. American Medical Informatics Association.* J Am Med Inform Assoc, 1994. **1**(1): p. 1-7.

99. HL7. [cited 2016 Aug 3]; Available from: http://www.hl7.org.

100. Henry, W.L., et al., *Report of the American Society of Echocardiography Committee on Nomenclature and Standards in Two-dimensional Echocardiography.* Circulation, 1980. **62**(2): p. 212-7.

101. CHI, *Consolidated Health Informatics: Standards Adoption Recommendation, Multimedia Information in Patient Records*

102. WHO. *ICD*. [cited 2016 Aug 3]; Available from: http://www.who.int/classifications/icd/en/.

103. IHTSDO. *SNOMED CT*. [cited 2016 Aug 3]; Available from: http://www.ihtsdo.org/*SNOMED-CT*.

104. IHTSDO, *SNOMED CT® User Guide*. 2013.

105. UMLS. *SNOMED Clinical Terms® To Be Added To UMLS® Metathesaurus®*. [cited 2016 Aug 3]; Available from: https://www.nlm.nih.gov/research/umls/SNOMED/SNOMED_announcement.html.

106. Richesson, R.L., J.E. Andrews, and J.P. Krischer, *Use of SNOMED CT to represent clinical research data: a semantic characterization of data items on case report forms in vasculitis research.* J Am Med Inform Assoc, 2006. **13**(5): p. 536-46.

107. Spackman, K.A. and K.E. Campbell, *Compositional concept representation using SNOMED: towards further convergence of clinical terminologies.* Proc AMIA Symp, 1998: p. 740-4.

108. IHTSDO, *SNOMED CT® Starter Guide*. 2014.

109. Cornet, R. and N. de Keizer, *Forty years of SNOMED: a literature review.* BMC Med Inform Decis Mak, 2008. **8 Suppl 1**: p. S2.

110. Lee, D., et al., *Literature review of SNOMED CT use.* Journal of the American Medical Informatics Association, 2014. **21**(e1): p. e11-e19.

111. Giannangelo, K. and J. Millar, *Mapping SNOMED CT to ICD-10.* Stud Health Technol Inform, 2012. **180**: p. 83-7.

112. Nadkarni, P.M. and J.A. Darer, *Migrating existing clinical content from ICD-9 to SNOMED.* J Am Med Inform Assoc, 2010. **17**(5): p. 602-7.

113. Wade, G. and S.T. Rosenbloom, *Experiences mapping a legacy interface terminology to SNOMED CT.* BMC Med Inform Decis Mak, 2008. **8 Suppl 1**: p. S3.

114. Wade, G. and S.T. Rosenbloom, *The impact of SNOMED CT revisions on a mapped interface terminology: terminology development and implementation issues.* J Biomed Inform, 2009. **42**(3): p. 490-3.

115. Day, M.J., et al., *Histopathological standards for the diagnosis of gastrointestinal inflammation in endoscopic biopsy samples from the dog and cat: a report from the*

World Small Animal Veterinary Association Gastrointestinal Standardization Group. J Comp Pathol, 2008. **138 Suppl 1**: p. S1-43.

116. Washabau, R.J. *Report from: WSAVA Gastrointestinal Standardization Group*. 2005. [cited 2016 Aug 3]; Available from: http://www.wsava.org/sites/default/files/GI%20Report%202005.pdf.

117. Washabau, R., et al., *Endoscopic, biopsy, and histopathologic guidelines for the evaluation of gastrointestinal inflammation in companion animals.* Journal of veterinary internal medicine, 2010. **24**(1): p. 10-26.

118. Kelton, D.F., K.D. Lissemore, and R.E. Martin, *Recommendations for recording and calculating the incidence of selected clinical diseases of dairy cattle.* J Dairy Sci, 1998. **81**(9): p. 2502-9.

119. Berendt, M., et al., *International veterinary epilepsy task force consensus report on epilepsy definition, classification and terminology in companion animals.* BMC Vet Res, 2015. **11**: p. 182.

120. Zimmerman, K.L., et al., *SNOMED representation of explanatory knowledge in veterinary clinical pathology.* Vet Clin Pathol, 2005. **34**(1): p. 7-16.

121. *VTSL*. [cited 2016 Aug 3]; Available from: http://vtsl.vetmed.vt.edu/.

122. Koopman, B., et al., *Towards semantic search and inference in electronic medical records: An approach using concept--based information retrieval.* Australas Med J, 2012. **5**(9): p. 482-8.

123. Arnot-Smith, J. and A.F. Smith, *Patient safety incidents involving neuromuscular blockade: analysis of the UK National Reporting and Learning System data from 2006 to 2008.* Anaesthesia, 2010. **65**(11): p. 1106-13.

124. Benoit, S.R., et al., *Automated surveillance of Clostridium difficile infections using BioSense.* Infect Control Hosp Epidemiol, 2011. **32**(1): p. 26-33.

125. Liaw, S.T., Chen, H. Y., Maneze, D., Taggart, J., Dennis, S., Vagholkar, S., & Bunker, J. , *Health reform: is routinely collected electronic information fit for purpose?* Emergency Medicine Australasia, 2012. **24(1)**: p. 57-63.

126. Matheny, M.E., et al., *Detection of infectious symptoms from VA emergency department and primary care clinical documentation.* Int J Med Inform, 2012. **81**(3): p. 143-56.

127. Nguyen, A., et al., *Classification of pathology reports for cancer registry notifications.* Stud Health Technol Inform, 2012. **178**: p. 150-6.

128. Zuccon, G., et al., *The impact of OCR accuracy on automated cancer classification of pathology reports.* Stud Health Technol Inform, 2012. **178**: p. 250-6.

129. Strauss, J.A., Chao, C. R., Kwan, M. L., Ahmed, S. A., Schottinger, J. E., & Quinn, V. P., *Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm.* Journal of the American Medical Informatics Association, 2013. **20(2)**: p. 349-355.

130. Schulz, S., et al., *SNOMED reaching its adolescence: ontologists' and logicians' health check.* Int J Med Inform, 2009. **78 Suppl 1**: p. S86-94.

131. Heiler, S., *Semantic interoperability.* ACM Computing Surveys (CSUR), 1995. **27**(2): p. 271-273.

132. Spackman, K.A., *Normal forms for description logic expressions of clinical concepts in SNOMED RT.* Proc AMIA Symp, 2001: p. 627-31.

133. Campbell, W.S., et al., *Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings.* J Am Med Inform Assoc, 2014. **21**(5): p. 885-92.

134. *AVMA. SNOMED, HL7, LOINC the official informatics standards for veterinary medicine*. [cited 2016 Aug 3]; Available from: https://www.avma.org/News/JAVMANews/Pages/s020601o.aspx.

135. *AAHA. AAHA Problem and Diagnosis Terms*. [cited 2016 Aug 3]; Available from: https://www.aaha.org/professional/resources/diagnostic_terms.aspx.

136. Santamaria, S.L. and K.L. Zimmerman, *Uses of informatics to solve real world problems in veterinary medicine.* J Vet Med Educ, 2011. **38**(2): p. 103-9.

137. Sarkar, I.N., *Leveraging biomedical ontologies and annotation services to organize microbiome data from Mammalian hosts.* AMIA Annu Symp Proc, 2010. **2010**: p. 717-21.

138. Rosse, C. and J.L. Mejino, Jr., *A reference ontology for biomedical informatics: the Foundational Model of Anatomy.* J Biomed Inform, 2003. **36**(6): p. 478-500.

139. Bodenreider, O. and S. Zhang, *Comparing the representation of anatomy in the FMA and SNOMED CT.* AMIA Annu Symp Proc, 2006: p. 46-50.

140. Zaninelli, M., et al., *The O3-Vet project: integration of a standard nomenclature of clinical terms in a veterinary electronic medical record for veterinary hospitals.* Comput Methods Programs Biomed, 2012. **108**(2): p. 760-72.

141. Smith-Akin, K.A., et al., *Toward a veterinary informatics research agenda: an analysis of the PubMed-indexed literature.* international journal of medical informatics, 2007. **76**(4): p. 306-312.

142. Cimino, J.J., *Desiderata for controlled medical vocabularies in the twenty-first century.* Methods Inf Med, 1998. **37**(4-5): p. 394-403.

143. Datamark, *Unstructured Data in Electronic Health Record (EHR) Systems: Challenges and Solutions*. 2013.

144. Branavan, S., et al., *Learning document-level semantic properties from free-text annotations.* Journal of Artificial Intelligence Research, 2009. **34**: p. 569-603.

145. McDonald, C.J. and W.M. Tierney, *Computer-stored medical records: their future role in medical practice.* Jama, 1988. **259**(23): p. 3433-3440.

146. Fries, J.F., *Alternatives in medical record formats.* Medical care, 1974: p. 871-881.

147. McDonald, C.J. and W.M. Tierney, *Computer-stored medical records. Their future role in medical practice.* JAMA, 1988. **259**(23): p. 3433-40.

148. Whiting-O'Keefe, Q.E., et al., *A computerized summary medical record system can provide more information than the standard medical record.* JAMA, 1985. **254**(9): p. 1185-92.

149. Wilson, G.A., C.J. McDonald, and G.P. McCabe, Jr., *The effect of immediate access to a computerized medical record on physician test ordering: a controlled clinical trial in the emergency room.* Am J Public Health, 1982. **72**(7): p. 698-702.

150. Gouveia-Oliveira, A., et al., *Longitudinal comparative study on the influence of computers on reporting of clinical data.* Endoscopy, 1991. **23**(6): p. 334-7.

151. Mann, R. and J. Williams, *Standards in medical record keeping.* Clin Med (Lond), 2003. **3**(4): p. 329-32.

152. Johnson, S.B., et al., *An electronic health record based on structured narrative.* J Am Med Inform Assoc, 2008. **15**(1): p. 54-64.

153. Kemper, H.-G., *Management Support with Structured and Unstructured Data - An Integrated Business Intelligence Framework.* Information Systems Management, 2008. **25**(2): p. 132-148.

154. Fonseca, F., Max Egenhofer, Clodoveu Davis, Gilberto Câmara, *Semantic granularity in ontology-driven geographic information systems.* 2002. **36(1-2)**: p. 121-151.

155. Rosenbloom, S.T., et al., *Data from clinical notes: a perspective on the tension between structure and flexible documentation.* J Am Med Inform Assoc, 2011. **18**(2): p. 181-6.

156. Montoliu-Fornas, G. and L. Marti-Bonmati, *Magnetic resonance imaging structured reporting in infertility.* Fertil Steril, 2016. **105**(6): p. 1421-31.

157. Allin, S., et al., *Evaluation of Automated Fracture Risk Assessment Based on the Canadian Association of Radiologists and Osteoporosis Canada Assessment Tool.* J Clin Densitom, 2016. **19**(3): p. 332-9.

158. Ryu, K.H., et al., *Cervical Lymph Node Imaging Reporting and Data System for Ultrasound of Cervical Lymphadenopathy: A Pilot Study.* AJR Am J Roentgenol, 2016. **206**(6): p. 1286-91.

159. McKinley, M. and M. Newman, *Observations on the application of the Papanicolaou Society of Cytopathology standardised terminology and nomenclature for pancreaticobiliary cytology.* Pathology, 2016. **48**(4): p. 353-6.

160. Christopher, M.M. and C.S. Hotz, *Cytologic diagnosis: expression of probability by clinical pathologists.* Vet Clin Pathol, 2004. **33**(2): p. 84-95.

161. Gouveia-Oliveira, A., et al., *Longitudinal comparative study on the influence of computers on reporting of clinical data.* Endoscopy, 1991. **23**(06): p. 334-337.

162. Schweitzer, W., Christine Bartsch, Thomas D. Ruder, Michael J. Thali. , *Virtopsy approach: structured reporting versus free reporting for PMCT findings.* 2014. **2(1)**: p. 28-33.

163. Amin, M.B., *The 2009 version of the cancer protocols of the college of american pathologists.* Arch Pathol Lab Med, 2010. **134**(3): p. 326-30.

164. Zarbo, R.J., *Interinstitutional assessment of colorectal carcinoma surgical pathology report adequacy. A College of American Pathologists Q-Probes study of practice patterns from 532 laboratories and 15,940 reports.* Arch Pathol Lab Med, 1992. **116**(11): p. 1113-9.

165. Kempson, R.L., *The time is now. Checklists for surgical pathology reports.* Arch Pathol Lab Med, 1992. **116**(11): p. 1107-8.

166. Rosai, J., *Standardized reporting of surgical pathology diagnoses for the major tumor types. A proposal. The Department of Pathology, Memorial Sloan-Kettering Cancer Center.* Am J Clin Pathol, 1993. **100**(3): p. 240-55.

167. Leslie, K.O. and J. Rosai, *Standardization of the surgical pathology report: formats, templates, and synoptic reports.* Semin Diagn Pathol, 1994. **11**(4): p. 253-7.

168. Gephardt, G.N. and P.B. Baker, *Lung carcinoma surgical pathology report adequacy: a College of American Pathologists Q-Probes study of over 8300 cases from 464 institutions.* Arch Pathol Lab Med, 1996. **120**(10): p. 922-7.

169. Farrow, G. and M.B. Amin, *Protocol for the examination of specimens from patients with carcinomas of renal tubular origin, exclusive of Wilms tumor and tumors of urothelial origin: a basis for checklists. Cancer Committee, College of American Pathologists.* Arch Pathol Lab Med, 1999. **123**(1): p. 23-7.

170. Oberhuber, G., G. Granditsch, and H. Vogelsang, *The histopathology of coeliac disease: time for a standardized report scheme for pathologists.* Eur J Gastroenterol Hepatol, 1999. **11**(10): p. 1185-94.

171. Prevention, C.f.D.C.a., *Report on the Reporting Pathology Protocols for Colon and Rectum Cancers Project*. 2005.

172. Powsner, S.M., J. Costa, and R.J. Homer, *Clinicians are from Mars and pathologists are from Venus: clinician interpretation of pathology reports.* Archives of pathology & laboratory medicine, 2000. **124**(7): p. 1040-1046.

173. Christopher, L., *Current status of discrete data capture in synoptic surgical pathology and cancer reporting.* Pathology and Laboratory Medicine International, 2015.

174. Srigley, J.R., et al., *Standardized synoptic cancer pathology reporting: a population-based approach.* J Surg Oncol, 2009. **99**(8): p. 517-24.

175. Idowu, M.O., et al., *Adequacy of surgical pathology reporting of cancer: a College of American Pathologists Q-Probes study of 86 institutions.* Archives of pathology & laboratory medicine, 2010. **134**(7): p. 969-974.

176. Lindley, S.W., E.M. Gillies, and L.A. Hassell, *Communicating diagnostic uncertainty in surgical pathology reports: Disparities between sender and receiver.* Pathology-Research and Practice, 2014. **210**(10): p. 628-633.

177. Cowan, D., *Quality assurance in anatomic pathology. An information system approach.* Archives of pathology & laboratory medicine, 1990. **114**(2): p. 129-134.

178. King, S., M. Dimech, and S. Johnstone, *Structured pathology reporting improves the macroscopic assessment of rectal tumour resection specimens.* Pathology, 2016. **48**(4): p. 349-52.

179. Leslie, K.O. and J. Rosai. *Standardization of the surgical pathology report: formats, templates, and synoptic reports*. in *Seminars in diagnostic pathology*. 1994.

180. Scolyer, R.A., Thompson, J. F., Stretch, J. R., Sharma, R., & McCarthy, S. W, *Pathology of melanocytic lesions: new, controversial, and clinically important issues.* Journal of surgical oncology, 2004. **86(4)**: p. 200-211.

181. Scolyer, R., et al. *Collaboration between clinicians and pathologists: a necessity for the optimal management of melanoma patients*. in *Cancer Forum*. 2005. The Cancer Council Australia.

182. Miller, P.R., *Inpatient diagnostic assessments: 2. Interrater reliability and outcomes of structured vs. unstructured interviews.* Psychiatry Res, 2001. **105**(3): p. 265-71.

183. Miller, P.R., et al., *Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews.* Psychiatry Res, 2001. **105**(3): p. 255-64.

184. Miller, P.R., *Inpatient diagnostic assessments: 3. Causes and effects of diagnostic imprecision.* Psychiatry Res, 2002. **111**(2-3): p. 191-7.

185. Leib, M.S., et al., *Endoscopic aspiration of intestinal contents in dogs and cats: 394 cases.* J Vet Intern Med, 1999. **13**(3): p. 191-3.

186. Louwerens, M., et al., *Feline lymphoma in the post-feline leukemia virus era.* J Vet Intern Med, 2005. **19**(3): p. 329-35.

187. Carreras, J.K., et al., *Feline epitheliotropic intestinal malignant lymphoma: 10 cases (1997-2000).* J Vet Intern Med, 2003. **17**(3): p. 326-31.

188. Russell, K.J., et al., *Feline low-grade alimentary lymphoma: how common is it?* J Feline Med Surg, 2012. **14**(12): p. 910-2.

189. Bridgeford, E.C., et al., *Gastric Helicobacter species as a cause of feline gastric lymphoma: a viable hypothesis.* Veterinary immunology and immunopathology, 2008. **123**(1): p. 106-113.

190. Bertone, E.R., L.A. Snyder, and A.S. Moore, *Environmental tobacco smoke and risk of malignant lymphoma in pet cats.* Am J Epidemiol, 2002. **156**(3): p. 268-73.

191. Gieger, T., *Alimentary lymphoma in cats and dogs.* Veterinary Clinics of North America: Small Animal Practice, 2011. **41**(2): p. 419-432.

192. Hartmann, K., *Clinical aspects of feline retroviruses: a review.* Viruses, 2012. **4**(11): p. 2684-710.

193. Barrs, V.R. and J.A. Beatty, *Feline alimentary lymphoma: 1. Classification, risk factors, clinical signs and non-invasive diagnostics.* J Feline Med Surg, 2012. **14**(3): p. 182-90.

194. Pohlman, L.M., et al., *Immunophenotypic and histologic classification of 50 cases of feline gastrointestinal lymphoma.* Vet Pathol, 2009. **46**(2): p. 259-68.

195. Moore, P.F., et al., *Characterization of feline T cell receptor gamma (TCRG) variable region genes for the molecular diagnosis of feline intestinal T cell lymphoma.* Vet Immunol Immunopathol, 2005. **106**(3-4): p. 167-78.

196. Gabor, L.J., P.J. Canfield, and R. Malik, *Haematological and biochemical findings in cats in Australia with lymphosarcoma.* Aust Vet J, 2000. **78**(7): p. 456-61.

197. Schnabel, L.V., et al., *Primary alimentary lymphoma with metastasis to the liver causing encephalopathy in a horse.* J Vet Intern Med, 2006. **20**(1): p. 204-6.

198. Smith, A.L., et al., *Perioperative complications after full-thickness gastrointestinal surgery in cats with alimentary lymphoma.* Vet Surg, 2011. **40**(7): p. 849-52.

199. Taylor, S.S., et al., *Serum protein electrophoresis in 155 cats.* J Feline Med Surg, 2010. **12**(8): p. 643-53.

200. Frank, J.D., et al., *Clinical outcomes of 30 cases (1997-2004) of canine gastrointestinal lymphoma.* J Am Anim Hosp Assoc, 2007. **43**(6): p. 313-21.

201. Couto, C.G., et al., *Gastrointestinal lymphoma in 20 dogs.* Journal of Veterinary Internal Medicine, 1989. **3**(2): p. 73-78.

202. Ragaini, L., et al., *Inflammatory bowel disease mimicking alimentary lymphosarcoma in a cat.* Vet Res Commun, 2003. **27 Suppl 1**: p. 791-3.

203. Cook, A.K., et al., *Prevalence and prognostic impact of hypocobalaminemia in dogs with lymphoma.* Journal of the American Veterinary Medical Association, 2009. **235**(12): p. 1437-1441.

204. Willard, M.D., *Feline inflammatory bowel disease: a review.* J Feline Med Surg, 1999. **1**(3): p. 155-64.

205. Lalor, S., et al., *Cats with Inflammatory Bowel Disease and Intestinal Small Cell Lymphoma Have Low Serum Concentrations of 25-Hydroxyvitamin D.* Journal of Veterinary Internal Medicine, 2014. **28**(2): p. 351-355.

206. Jergens, A.E., et al., *A clinical index for disease activity in cats with chronic enteropathy.* J Vet Intern Med, 2010. **24**(5): p. 1027-33.

207. Evans, S.E., et al., *Comparison of endoscopic and full-thickness biopsy specimens for diagnosis of inflammatory bowel disease and alimentary tract lymphoma in cats.* J Am Vet Med Assoc, 2006. **229**(9): p. 1447-50.

208. Roth, L., et al., *Comparisons between Endoscopic and Histologic Evaluation of the Gastrointestinal-Tract in Dogs and Cats - 75 Cases (1984-1987).* J Am Vet Med Assoc, 1990. **196**(4): p. 635-638.

209. Willard, M.D., et al., *Effect of sample quality on the sensitivity of endoscopic biopsy for detecting gastric and duodenal lesions in dogs and cats.* J Vet Intern Med, 2008. **22**(5): p. 1084-9.

210. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.

211. Al-Ghazlat, S., C.E. de Rezende, and J. Ferreri, *Feline small cell lymphosarcoma versus inflammatory bowel disease: diagnostic challenges.* Compend Contin Educ Vet, 2013. **35**(6): p. E1-5; quiz E6.

212. Wilson, H.M., *Feline alimentary lymphoma: demystifying the enigma.* Top Companion Anim Med, 2008. **23**(4): p. 177-84.

213. Kiupel, M., et al., *Diagnostic algorithm to differentiate lymphoma from inflammation in feline small intestinal biopsy samples.* Vet Pathol, 2011. **48**(1): p. 212-22.

214. Moore, P.F., A. Rodriguez-Bertos, and P.H. Kass, *Feline gastrointestinal lymphoma: mucosal architecture, immunophenotype, and molecular clonality.* Vet Pathol, 2012. **49**(4): p. 658-68.

215. Zwingenberger, A.L., et al., *Ultrasonographic evaluation of the muscularis propria in cats with diffuse small intestinal lymphoma or inflammatory bowel disease.* J Vet Intern Med, 2010. **24**(2): p. 289-92.

216. Waly, N.E., et al., *Immunohistochemical diagnosis of alimentary lymphomas and severe intestinal inflammation in cats.* J Comp Pathol, 2005. **133**(4): p. 253-60.

217. Johnston, S.A., D.H. Thamm, and J.B. Legutki, *The immunosignature of canine lymphoma: characterization and diagnostic application.* BMC Cancer, 2014. **14**: p. 657.

218. Burke, K.F., et al., *Evaluation of fecal alpha1-proteinase inhibitor concentrations in cats with idiopathic inflammatory bowel disease and cats with gastrointestinal neoplasia.* Vet J, 2013. **196**(2): p. 189-96.

219. Stewart, W.F., et al., *Bridging the inferential gap: the electronic health record and clinical evidence.* Health Aff (Millwood), 2007. **26**(2): p. w181-91.

220. Lipkin, M. and J.D. Hardy, *Mechanical correlation of data in differential diagnosis of hematological diseases.* J Am Med Assoc, 1958. **166**(2): p. 113-25.

221. Ledley, S., L.B. Lusted, and R.S. Ledley. *Reasoning foundations of medical diagnosis*. in *Science*. 1959. Citeseer.

222. Shortliffe, E.H., B.G. Buchanan, and E.A. Feigenbaum, *Knowledge engineering for medical decision making: A review of computer-based clinical decision aids.* Proceedings of the IEEE, 1979. **67**(9): p. 1207-1224.

223. Pople, H.E., J. Myers, and R. Miller. *DIALOG: A Model Of Diagnostic Logic For Internal Medicine*. in *IJCAI*. 1975. Citeseer.

224. Blois, M., M. Tuttle, and D. Sherertz. *RECONSIDER: A Program for Generating Differential Diagnoses**. in *Proceedings/the... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care*. 1981. American Medical Informatics Association.

225. Barnett, G.O., et al. *DXplain: Experience with knowledge acquisition and program evaluation*. in *Proceedings of the Annual Symposium on Computer Application in Medical Care*. 1987. American Medical Informatics Association.

226. Willems, J.L., et al., *The diagnostic performance of computer programs for the interpretation of electrocardiograms.* N Engl J Med, 1991. **325**(25): p. 1767-73.

227. Floyd, C.E., et al., *Prediction of breast cancer malignancy using an artificial neural network.* Cancer, 1994. **74**(11): p. 2944-2948.

228. Furundzic, D., M. Djordjevic, and A.J. Bekic, *Neural networks approach to early breast cancer detection.* Journal of systems architecture, 1998. **44**(8): p. 617-633.

229. Abbass, H.A., *An evolutionary artificial neural networks approach for breast cancer diagnosis.* Artificial intelligence in Medicine, 2002. **25**(3): p. 265-281.

230. Prasad, S., L.M. Bruce, and J.E. Ball. *A multi-classifier and decision fusion framework for robust classification of mammographic masses*. in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2008. IEEE.

231. Dheeba, J. and S. Tamil Selvi, *An improved decision support system for detection of lesions in mammograms using Differential Evolution Optimized Wavelet Neural Network.* J Med Syst, 2012. **36**(5): p. 3223-32.

232. Setiono, R., *Extracting rules from pruned neural networks for breast cancer diagnosis.* Artificial Intelligence in Medicine, 1996. **8**(1): p. 37-51.

233. Jimbo, M., et al., *What is lacking in current decision aids on cancer screening?* CA: a cancer journal for clinicians, 2013. **63**(3): p. 193-214.

234. Abbod, M.F., et al., *Application of artificial intelligence to the management of urological cancer.* The Journal of urology, 2007. **178**(4): p. 1150-1156.

235. Sawka, A.M., et al., *Decision aid on radioactive iodine treatment for early stage papillary thyroid cancer: update to study protocol with follow-up extension.* Trials, 2015. **16**: p. 302.

236. Shen, Y., et al., *Emerging medical informatics with case-based reasoning for aiding clinical decision in multi-agent system.* J Biomed Inform, 2015. **56**: p. 307-17.

237. Hills, R.L., P.A. Kulbok, and M. Clark, *Evaluating a Quality Improvement Program for Cervical Cancer Screening at an Urban Safety Net Clinic.* Health Promot Pract, 2015. **16**(5): p. 631-41.

238. Qiu, Y., et al., *Knowledge discovery for pancreatic cancer using inductive logic programming.* IET systems biology, 2014. **8**(4): p. 162-168.

239. Julia-Sape, M., et al., *Multicentre evaluation of the INTERPRET decision support system 2.0 for brain tumour classification.* NMR Biomed, 2014. **27**(9): p. 1009-18.

240. Al-Kadi, O.S., *A multiresolution clinical decision support system based on fractal model design for classification of histological brain tumours.* Comput Med Imaging Graph, 2015. **41**: p. 67-79.

241. Hutchings, M., *How does PET/CT help in selecting therapy for patients with Hodgkin lymphoma?* ASH Education Program Book, 2012. **2012**(1): p. 322-327.

242. Zelic, I., et al., *Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries.* J Med Syst, 1997. **21**(6): p. 429-44.

243. Evans, R.S., et al., *A computer-assisted management program for antibiotics and other antiinfective agents.* New England Journal of Medicine, 1998. **338**(4): p. 232-238.

244. Lewis, G., et al., *Computerized assessment of common mental disorders in primary care: effect on clinical outcome.* Fam Pract, 1996. **13**(2): p. 120-6.

245. Cannon, D.S. and S.N. Allen, *A comparison of the effects of computer and manual reminders on compliance with a mental health clinical practice guideline.* Journal of the American Medical Informatics Association, 2000. **7**(2): p. 196-203.

246. Schriger, D.L., et al., *Enabling the diagnosis of occult psychiatric illness in the emergency department: a randomized, controlled trial of the computerized, self-administered PRIME-MD diagnostic system.* Annals of emergency medicine, 2001. **37**(2): p. 132-140.

247. Pozen, M.W., et al., *A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease: a prospective multicenter clinical trial.* New England Journal of Medicine, 1984. **310**(20): p. 1273-1278.

248. Selker, H.P., et al., *Use of the acute cardiac ischemia time-insensitive predictive instrument (ACI-TIPI) to assist with triage of patients with chest pain or other symptoms suggestive of acute cardiac ischemia: a multicenter, controlled clinical trial.* Annals of Internal Medicine, 1998. **129**(11_Part_1): p. 845-855.

249. Wellwood, J., S. Johannessen, and D. Spiegelhalter, *How does computer-aided diagnosis improve the management of acute abdominal pain?* Annals of the Royal College of Surgeons of England, 1992. **74**(1): p. 40.

250. Ferreira, D., A. Oliveira, and A. Freitas, *Applying data mining techniques to improve diagnosis in neonatal jaundice.* BMC medical informatics and decision making, 2012. **12**(1): p. 1.

251. Schmickl, C.N., et al., *Decision support tool for differential diagnosis of Acute Respiratory Distress Syndrome (ARDS) vs Cardiogenic Pulmonary Edema (CPE): a prospective validation and meta-analysis.* Crit Care, 2014. **18**(6): p. 659.

252. Verma, L., S. Srivastava, and P.C. Negi, *A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data.* J Med Syst, 2016. **40**(7): p. 178.

253. Wagholikar, K.B., et al., *Clinical decision support with automated text processing for cervical cancer screening.* J Am Med Inform Assoc, 2012. **19**(5): p. 833-9.

254. Al-Omari, F.A., et al., *An intelligent decision support system for quantitative assessment of gastric atrophy.* Journal of clinical pathology, 2011. **64**(4): p. 330-337.

255. Simões, P.W., et al., *Classification of images acquired with colposcopy using artificial neural networks.* Cancer informatics, 2014. **13**: p. 119.

256. Wu, W.-J., S.-W. Lin, and W.K. Moon, *An Artificial Immune System-Based Support Vector Machine Approach for Classifying Ultrasound Breast Tumor Images.* Journal of digital imaging, 2015. **28**(5): p. 576-585.

257. Arevalo, J., et al., *An unsupervised feature learning framework for basal cell carcinoma image analysis.* Artif Intell Med, 2015. **64**(2): p. 131-45.

258. Zeldis, D. and S. Prescott, *Fish disease diagnosis program—problems and some solutions.* Aquacultural Engineering, 2000. **23**(1): p. 3-11.

259. Faunt, K., E. Lund, and W. Novak, *The power of practice: harnessing patient outcomes for clinical decision making.* Veterinary Clinics of North America: Small Animal Practice, 2007. **37**(3): p. 521-532.

260. Dórea, F.C., et al., *Exploratory analysis of methods for automated classification of laboratory test orders into syndromic groups in veterinary medicine.* PLoS One, 2013. **8**(3): p. e57334.

261. Faria, A.R., et al., *High-throughput analysis of synthetic peptides for the immunodiagnosis of canine visceral leishmaniasis.* PLoS Negl Trop Dis, 2011. **5**(9): p. e1310.

262. Wan, L. and W. Bao. *Animal disease diagnoses expert system based on SVM*. in *International Conference on Computer and Computing Technologies in Agriculture*. 2009. Springer.

263. Parkhi, O.M., et al. *Cats and dogs*. in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. 2012. IEEE.

264. Parkhi, O.M., et al. *The truth about cats and dogs*. in *2011 International Conference on Computer Vision*. 2011. IEEE.

Chapter 2 - Development of Structured Histopathological Diagnoses for Gastrointestinal Biopsies

2.1 Abstract

The World Small Animal Veterinary Association (WSAVA) Gastrointestinal International (GI) Standardization Group has proposed standards for recording the microscopic findings of endoscopic biopsies for dogs and cats. To date, WSAVA has not proposed standard means by which such assessments, called histopathological diagnoses are recorded. The work here describes a semantic model and a limited terminology that can be combined to allow for consistent structured recording of histopathological diagnoses. An information model (IM) based on attribute-value pairs was created to express the unstructured diagnoses in a structured syntax. Attributes and value lists (terms) were derived from an assessment of GI biopsy reports from client-owned dogs and cats presented to our institution's hospital. Direct 1:1 maps were created between the reported unstructured diagnoses and the values of our structured diagnoses attributes against the Systematized Nomenclature of Medicine - Clinical Terms (*SNOMED-CT)*. Our IM was composed of "Finding site" and "Abnormal morphology" attributes that are essential to be recorded for each histopathological diagnosis and three other attributes called option. Strategic instance testing confirmed expected placement of instances (composed of findings) as diagnoses class members. Direct 1:1 mapping of 114 unique unstructured diagnoses to *SNOMED-CT* revealed 44 % valid representation. Ninety-one percent of our values for "Abnormal morphology" attribute and 49% of our values for "Finding site" attribute were mapped directly to *SNOMED-CT* hierarchies. The proposed structured format of the histopathological diagnoses is amenable to automation, machine-learning, comparison, and retrieval. The values of our IM attributes were mapped to *SNOMED-CT*, which with additions or refinements can eventually serve as an acceptable source of external standardized terminology to support the recording and retrieval of HDX.

Keywords: structured diagnoses; standard terminology; histopathology; gastrointestinal biopsies

2.2 Introduction

Histopathological review of endoscopic biopsies has long been the 'reference standard' for the diagnosis of Gastrointestinal (GI) diseases in companion animals. [1, 2] The histopathological

findings (HF) and their interpretation into histopathological diagnoses (HDX) are conveyed in reports written by pathologists. These reports are then used to transfer the message from pathologists to clinicians and to represent pathology information in medical records.

The increased adoption of electronic health records and exponential increase in the volume of health information repositories, motivated the use of these health records in secondary activities, such as quality control, research, and public health statistics. These secondary uses of health information are facilitated by structured information recording (recording information in a computable format). However, because secondary uses of records are not top priorities for clinicians and pathologists, several studies have shown incompleteness in pathological data as it is reported in an unstructured (free text) format. [3-6]

Unstructured format has been the traditional format used to record the HF in pathology reports. Because free text descriptions can be incomplete, are not easily converted into a computable format, and don't achieve 100% accuracy when retrieved, clinicians and pathologists are increasingly using structured reporting formats. [8-11] In 2005, the World Small Animal Veterinary Association (WSAVA) GI International Standardization Group took the responsibility of standardizing the histologic evaluation of the gastrointestinal tract of dogs and cats. [12] Three years later, the group proposed standards for recording HF from GI biopsy samples. [13] The standards were developed to improve the quality of data recorded and to minimize variation among pathology reports of HF in the GI tract. The group established criteria for quantifying and ranking key endoscopic histologic observations from the gastric body, antrum, duodenum, and colon. These criteria provided specific numerical intervals to categorize the severity (mild, moderate, and marked) of different observations.

The current WSAVA standardization effort establishes a basis for consistent assessment of tissues by veterinary pathologists. These assessments are descriptions of the microscopic appearance of organs and tissues that may be additionally described by the likely time course of the pathology, its severity, and its distribution within and among the organs from which the biopsies are taken. To date WSAVA has not proposed standard means by which such assessments are reported.

These assessments composed of information either from a single tissue or created as an aggregate of more than one tissue, are often referred to as HDX. Unstructured HDX formats lead to variability in regard to the terms used to represent the diagnosis and the syntax or arrangement

of these terms. This can result in vagueness and ambiguity in the diagnosis and hampers efforts to abstract pathology information from electronic medical records, generate data-based discoveries, and compare retrospective or prospective results. [11, 14, 15] Recording HF and HDX in a computable state can be used by evidence-based practice for histopathology-based diagnosis. [16] Ultimately, efficient retrieval of morphological abnormalities within HDX, in conjunction with digital images and clinical findings, can help annotate histopathological images and facilitate the use of decision support systems.

Our goal is to establish an information model (IM) to represent the HDX components in a structured format. The IM should have specificity when reporting endoscopic HDX from the GI tract. Research projects may require recording of HF in single tissues. For example, an observation that an eosinophilic infiltrate is present in the gastric mucosa may generate a finding of eosinophilic inflammation of the gastric mucosa. Clinical case reports may require aggregate HDX for a particular patient. For example, evidence of eosinophilic inflammation of the gastric mucosa may generate a HDX of eosinophilic gastritis. Recording requirements may also vary as regards requirements for recording particulars of severity, distribution, and pathological course. Additionally, we believe that an IM is more likely to be used if its information structure can be extended to findings and diagnoses of other body sites.

As a first step, we separated the underlying logical structure of the model from the terminology that is germane to specific body regions and morphologies. This allowed us to consider the characteristics of the activity (examination of tissues by light microscopy) separately from the subject of the activity (endoscopic GI biopsies). Also, users of this information model will be able to adjust the terminology to suit the demands of their particular information recording or retrieval activity. For this project, our specific goal was to establish an information model (IM) to represent, in a structured manner, HF and HDX that result from microscopic assessment of endoscopic biopsies from the GI tract. Our model must support an efficient retrieval and aggregation of HDX. The model has additional value if the HDX can be used to retrieve examples of individual specific tissue observations or HF.

An IM organizes and categorizes information so it can be delivered and reused in variety of ways. [17] However, for an IM to be useful, it has to be associated with a terminology that populates its structure with the needed content. So that our IM can be accepted and integrated into modern electronic health information systems, it may also be necessary to associate it with and

map it to an external standardized terminology. *SNOMED-CT* was selected for this role as it is considered to be the most comprehensive medical terminology and the only external standardized terminology with substantial veterinary content appropriate to the intended applications of our information model. *SNOMED-CT* has also been endorsed by the American Veterinary Medical Association to be used as a standard terminology for veterinary medicine. [18] *SNOMED-CT* is the basis of the American Animal Hospital Association and AAHA Diagnostic Terms subset that was developed through the Veterinary Terminology Services Laboratory (VTSL) at Virginia Tech (https://www.vtsl.vetmed.vt.edu). [19] VTSL also manages a functional extension of *SNOMED-CT* that contains terms of interest in veterinary systems.

2.3 Materials and Methods

The IM developed in this study is a set of attributes common to the recording of HF and HDX. The model imparts structure to the recording activity when the attributes are paired with values (or terms) in what are referred to as attribute-value pairs. For this study, the terms or values were limited to those required to represent the histopathology of GI endoscopic biopsies. Separating the model and its attributes from the subject-specific terminology will allow the model to remain stable when applied to other subject areas.

Attributes and value lists (terms) were derived from an assessment of medical records for 1,028 dogs and cats presented to our institution's hospital between November 1, 2006 and April 29, 2013. Specific case records were eligible for this study if they included at least one endoscopic GI biopsy-based HDX showing altered histopathology. Three hundred and nineteen dogs and 125 cats with GI abnormalities presented in 463 histopathology reports met the criteria. The reports were manually reviewed and 916 unstructured HDX representing a list of 114 unique diagnoses were identified.

For purposes of the IM used in this study, detailed morphologic descriptions of tissues and tissue layers were called HF, e.g., eosinophilic infiltrate of the mucosa of the duodenum. Summary assessments of an organ (gastritis) or combinations of organs (gastroenteritis) were called HDX. Structural features of the model were identified by reviewing both the HF and the HDX. The model was tested for its ability to create sound, logical, and unique definitions for both HDX and the more specific HF associated with actual endoscopic biopsies. It was also evaluated for its ability to correctly associate HF with HDX.

Logical testing was accomplished by creating an ontology using *Protégé*, version 5.0.0 [20] The ontology included object properties representing each attribute in our model. Values for finding site, abnormal morphology, severity, distribution, and pathological course were created. Histopathological diagnoses were created as defined classes. Internal logic of the model was tested by auto-classification using *Hermit 1.3.8* as the reasoner. The ability of the model to maintain correct association between HF and HDX was tested by creating findings instances (*Protégé individuals*) and documenting the HDX class to which they would belong.

The Hermit reasoner gives Protégé the ability to evaluate the logical placement of instances as members of classes. We created a limited number of instances to produce a strategic test of the model logic. Instances were created to confirm correct assignment of HDX class membership given the following conditions: 1) Finding site values of instances were subtypes of the finding site values of the correct class or classes (*an instance of inflammation of the gastric epithelium is a member of the gastritis class*); 2) Abnormal morphology values of instances were subtypes of abnormal morphology values of the correct class or classes *(an instance with eosinophilic infiltrate present in the gastric mucosa is a member of the gastritis class)*; 3) Severity, distribution and pathological course had no effect on classification (*marked, focal, chronic inflammation of the gastric mucosa is a member of the gastritis class*); and 4) Instances composed of two or more HF classify to one or more correct classes. (*An instance with eosinophilic infiltrate present in the gastric mucosa and eosinophilic infiltrate present in the small intestinal mucosa is a member of the gastroenteritis class. Because the subtype relationship is transitive, the instance is also a member of both the enteritis and gastritis classes*).

The terminology we developed was assessed in terms of its alignment with an existing terminology standard (*SNOMED-CT*) and the terminology standard was assessed for the correctness of its representation of our terminology.

Once we were satisfied that the IM was functional and internally consistent, we created direct 1:1 maps between *SNOMED-CT* (January 2016 release) and the unstructured HDX from the case set as well the attributes and value sets used in our ontology. The unstructured HDX were mapped to *SNOMED-CT findings* sub-hierarchy. This sub-hierarchy is composed of concept classes defined or pre-coordinated using the *SNOMED-CT* information model. The list of "Abnormal morphology" concepts classes used in our IM mapped to *SNOMED-CT*

morphologically abnormal structure (49755003) subtypes and the list of "Finding site" concept classes were mapped to *SNOMED-CT body structure (123037004)* subtypes.

The VTSL browser (https://www.vtsl.vetmed.vt.edu) was used to identify *SNOMED-CT* concepts during the mapping process. The *SNOMED-CT* concepts and terms mapped to our IM were reviewed by a pathologist (KZ) to ensure semantic equivalence with unstructured HDX. Microsoft Access 2016 was utilized as a database management system to store and manage our IM.

2.4 Results

Figure 2.1 shows the IM developed and an example of using attribute-value pairs based on the model and the associated terminology. Information attributes that appeared in all HDX were the "Finding site" (macroscopic and/or microscopic body structure) and "Abnormal morphology" (microscopic abnormality). Our initial draft model treated the site of the finding and the abnormal morphology as essential attributes as they were consistently represented in HF and HDX. In the final version of the model, grouping was added so that a single finding site could be associated with only one abnormal morphology and vice versa. For instance, in order to define "gastroenteritis" using the model, four attribute-value pairs were required (Gastric mucosa and Inflammatory infiltrate + Small intestinal mucosa and Inflammatory infiltrate).

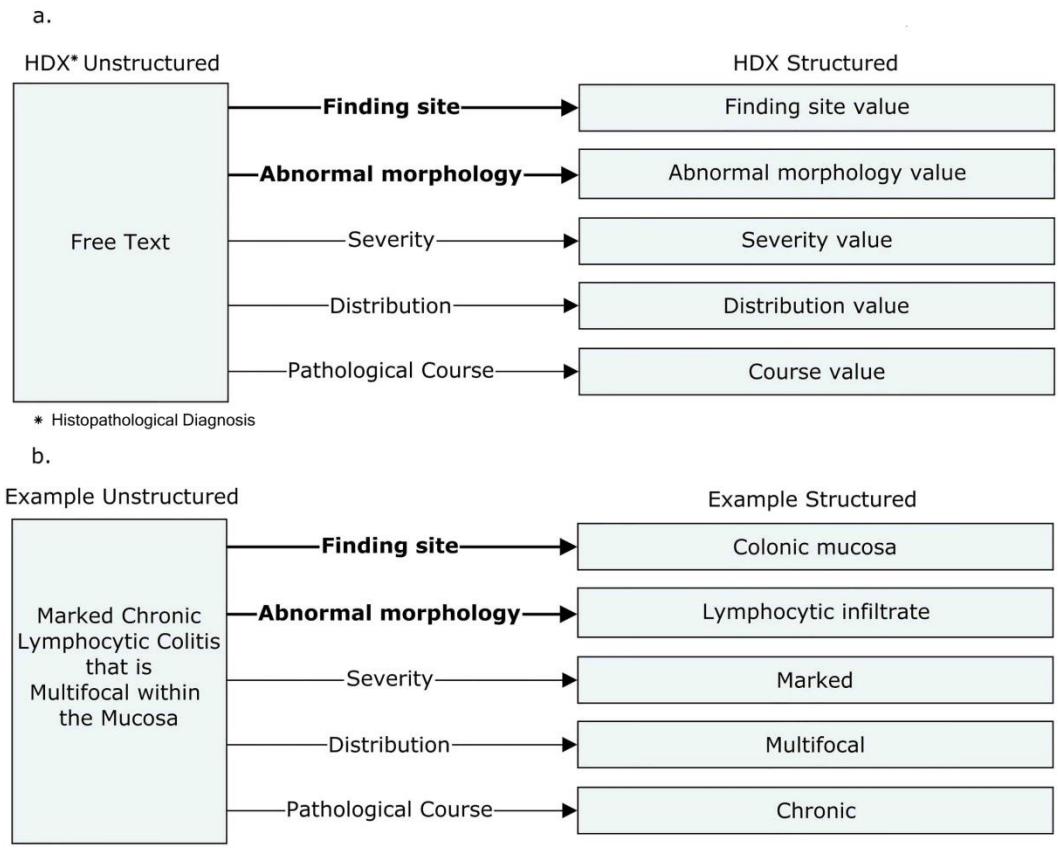

Figure 2.1 (a) A graphical depiction of the information model (IM) developed to represent the unstructured histopathological diagnoses (HDX) in a structured manner. (b) Shows an example of representing one of the unstructured HDX.

Typical definitions of HF and HDX within the ontology are shown in **Figure 2.2 a & b**. The logic of our IM did require the creation of a rule and storage structure that prevents free combination of attributes to create a single HF. If, for example more than one microscopic structure and more than one morphologic abnormality are combined without certain rules, it becomes impossible to determine which morphology is present in which structure. This does not create a requirement for creating a new instance, but it does require that HF be given a particular structure. In our ontology this was represented by an extra attribute and value pair (Is about + HF) as shown in **Figure 2.2c**.

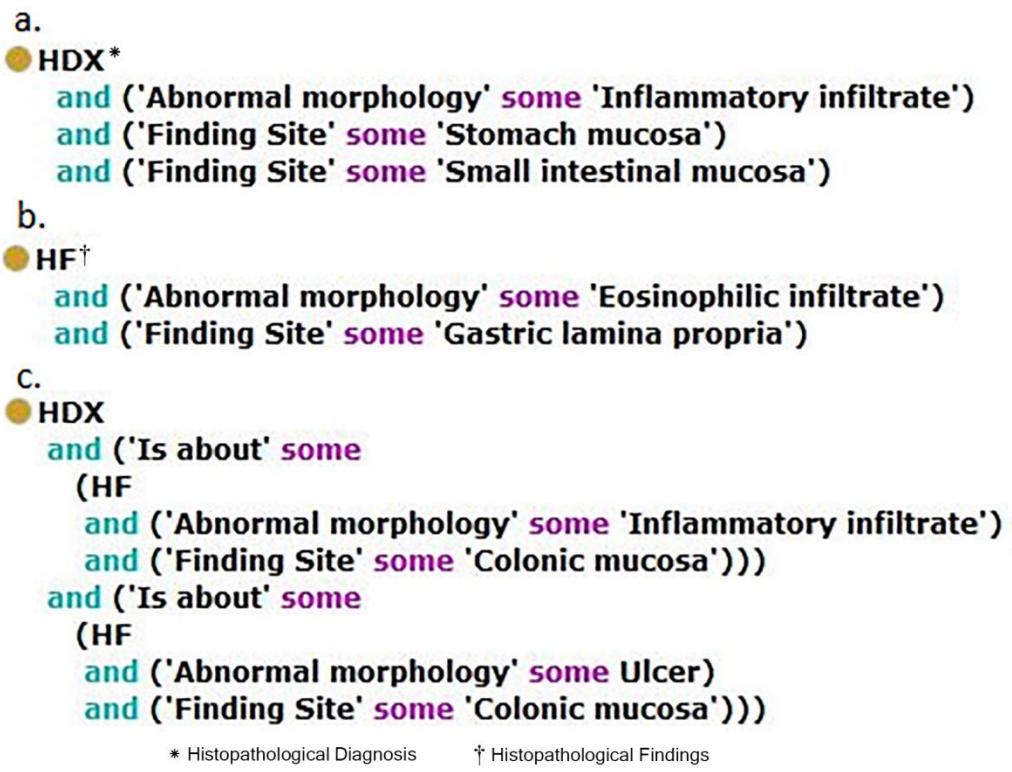

Figure 2.2 Protégé screenshot showing description logic representation of (a) Histopathological diagnoses (HDX) class "Gastroenteritis", (b) Histopathological findings (HF) instance created by the presence of eosinophilic infiltrate in the gastric lamina propria, and (c) HDX "Ulcerative colitis", an example of representing two morphologies within the same site using HF as a grouping attribute.

Table 2.1 shows a relational storage representation of the HDX "Ulcerative colitis" as described by the Protégé description logic representation (**Figure 2.2c**). Attribute-value pairs are grouped by the use of similar HF IDs.

Table 2.1 Storing histopathological findings (HF) for one of the histopathological diagnoses (HDX, "Ulcerative colitis") using HF grouping attribute in a relational table. HF ID shows the identification number used to group HF of each of the HDX (similar values group instances). HDX are concept classes. Attribute shows the attribute name and Attribute value shows the selected value term from each attribute.

HF ID	HDX	Attribute	Attribute value
1	Ulcerative Colitis	Abnormal morphology	Inflammatory morphology
1	Ulcerative Colitis	Finding site	Colonic Mucosa
2	Ulcerative Colitis	Abnormal morphology	Ulcer
2	Ulcerative Colitis	Finding site	Colonic Mucosa

In the records of clinical cases, HDX occasionally included references to severity, distribution, and pathological course. "Severity," "Distribution," and "Pathological course" were treated as optional attributes in the final version of the model. These attribute value pairs are included in the restriction that only a single severity, distribution, or pathological course could be associated with an existing HF (one "Finding site" + one "Abnormal morphology") group.

From the 114 diagnoses, 33 unique morphologies and 41 finding sites were identified (**Tables 2.2 & 2.3**). Values for severity, distribution and pathological course are listed in **Table 2.4**. Although pathologists tended to use "mild," "moderate," and "severe" in unstructured descriptions of HF or HDX, we elected to follow the WSAVA convention and represent "severe" as "marked". All HDX were then expressed in our IM in a structured format using the five previously listed data attributes; two were essential for each HDX and three were optional, as shown in **Figure 2.1**. Except for HF that included lymphocytic infiltrates, strategic instance testing confirmed expected placement of instances as HDX class members. For example, marked eosinophilic infiltrate of the gastric lamina propria (HF) was classified as gastritis (HDX class) which is defined by gastric mucosa and inflammatory morphology. By contrast, the presence of a lymphocytic infiltrate was inadequate to distinguish between inflammation and lymphoma.

Table 2.2 Model abnormal morphology concepts are the terms developed for the "Abnormal morphology" attribute of our information model (IM). SNOMED identifiers are *SNOMED-CT* concept identifiers that were selected for each IM term. SNOMED Synonyms are *SNOMED-CT* terms associated with each concept identifier.

Model Abnormal Morphology Concept	*SNOMED identifiers*	*SNOMED* Synonyms
Necrosis	6574001	Necrosis
Atrophy	13331008	Atrophy
Degeneration	33359002	Degeneration
Dilatation	25322007	Dilatation
Hypertrophy	56246009	Hypertrophy
Lymphangiectasia	308061000009100	Lymphangiectasia
Fibrosis	112674009	Fibrosis
Dysplasia	25723000	Dysplasia
Hyperplasia	76197007	Hyperplasia
Metaplasia	17665002	Metaplasia
Hyperplastic polyp	62047007	Hyperplastic polyp
Abscess	44132006	Abscess
Inflammatory infiltrate	23583003	Inflammatory infiltrate
Eosinophilic infiltrate	35261000009106	Eosinophilic infiltrate
lymphocytic infiltrate	-	-
Plasmacytic infiltrate	26246006	Plasmacytic infiltrate
Neutrophilic infiltrate	41034006	Neutrophilic infiltration*
Small cell lymphocytic infiltrate	-	-
Large cell lymphocytic infiltrate	-	-
Edema	79654002	Edema
Hemorrhage	50960005	Hemorrhage
Perforation	36191001	Perforation
Stricture	27551008	Stricture
Ulcer	56208002	Ulcer
Adenocarcinoma	35917007	Adenocarcinoma
Adenoma	32048006	Adenoma
Carcinoma in situ	68956006	Carcinoma in situ
Carcinoma	68453008	Carcinoma
Ganglioneuroma	53801007	Ganglioneuroma
Leiomyosarcoma	51549004	Leiomyosarcoma
Plasmacytoma	415113000	Plasmacytoma
Round cell tumor	45511000009106	Round cell tumor
Stromal tumor	128752000	Stromal tumor
* In *SNOMED-CT*, neutrophilic infiltration is a synonym associated with the concept *Acute inflammation.*		

Table 2.3 Model finding site concepts are the terms developed for the "Finding site" attribute of our information model (IM). SNOMED identifiers are *SNOMED-CT* concept identifiers that were selected for each IM term. SNOMED Synonyms are *SNOMED-CT* terms associated with each concept identifier.

Model Finding Site Concept	*SNOMED identifier*	*SNOMED* Synonyms
Gastric mucosa	78653002	Gastric mucous membrane structure
Gastric villi	-	-
Gastric epithelium	64977002	Gastric epithelium
Gastric lacteals	-	-
Gastric lamina propria	-	-
Duodenal mucosa	77763002	Duodenal mucous membrane structure
Duodenal villi	-	-
Duodenal epithelium	66036002	Duodenal epithelium
Duodenal crypts	-	-
Duodenal lacteals	-	-
Duodenal lamina propria	76582004	Duodenal lamina propria
Jejunal mucosa	88636004	Jejunal mucous membrane structure
Jejunal villi	-	-
Jejunal epithelium	57300004	Jejunal epithelium
Jejunal crypts	-	-
Jejunal lacteals	-	-
Jejunal lamina propria	14128009	Jejunal lamina propria
Ileal mucosa	85458007	Ileal mucous membrane structure
Ileal villi	-	-
Ileal epithelium	47046000	Ileal epithelium
Ileal crypts	-	-
Ileal lacteals	-	-
Ileal lamina propria	30751000	Ileal lamina propria
Small intestinal mucosa	4369007	Small intestine mucous membrane structure
Small intestinal villi	-	-
Small intestinal epithelium	45480009	Small intestine epithelium
Small intestinal crypts	52688008	Small intestinal crypt of Lieberkühn
Small intestinal Lacteals	-	-
Small intestinal lamina propria	63588008	Small intestine lamina propria
Colonic mucosa	68502009	Colonic mucous membrane structure
Colonic villi	-	-
Colonic epithelium	42978003	Colonic epithelium
Colonic crypts	19447003	Colonic crypt of Lieberkühn
Colonic lacteal	-	-
Colonic lamina propria	113284008	Colonic lamina propria
Rectal mucosa	15079008	Rectal mucous membrane structure
Rectal villi	-	-
Rectal epithelium	-	-
Rectal crypts	-	-
Rectal lacteals	-	-
Rectal lamina propria	-	-

Table 2.4 Model concepts are the terms developed for the "Severity", "Pathological Course", and "Distribution" attributes of our information model (IM). SNOMED identifiers are *SNOMED-CT* concept identifiers that were selected for each IM term. SNOMED Synonyms are *SNOMED-CT* terms associated with each concept identifier.

Model Severity Concept	***SNOMED identifiers***	***SNOMED* Severity Synonyms**
Mild	255604002	Mild
Moderate	6736007	Moderate
Marked	24484000	Marked
Model Distribution Concept	***SNOMED identifiers***	***SNOMED* Distribution Synonyms**
Diffuse	19648000	Diffuse
Focal	87017008	Focal
Multifocal	524008	Multifocal
Model Pathological Course Concept	***SNOMED identifiers***	***SNOMED* Course Synonyms**
Acute	-	-
Subacute	19939008	Subacute
Chronic	90734009	Chronic

Direct 1:1 mapping of the unstructured HDX to *SNOMED-CT* revealed that only 50 out of the 114 diagnoses (44 %) were represented by concepts pre-coordinated in *SNOMED-CT findings* sub-hierarchy. A higher portion of values to be associated with model attributes could be mapped although this varied by attribute. Values that we associated with the "Abnormal morphology" attribute were mapped to subtypes of *SNOMED-CT morphologically abnormal structure (49755003)* as shown in **Table 2.2**. Ninety-one percent (30/33) of these values were mapped. On the other hand, only 49% (20/41) of the IM values for use with the "Finding site" attribute mapped to *SNOMED-CT*, specifically to subtypes of *body structure (123037004)* (**Table 2.3**). The version of *SNOMED-CT* we used for mapping was inconsistent in its representation of microscopic features of specific organs. Mucosa and epithelium were represented for stomach and all intestinal subtypes. Microscopic anatomy including lamina propria, villi and crypts were inconsistently present. For example, *SNOMED-CT* includes a representation for "Intestinal Crypts of Lieberkühn" but does not represent Intestinal crypts for each organ.

All values for the "Severity" and "Distribution" attributes were mapped directly to subtypes of *Severities (272141005)* and *Distributions (255464007)* respectively. In attempting to map values of the "Pathological Course" attribute to *SNOMED-CT Courses* (288524001) subtypes, we identify a single exception. We did not map our value "Acute" to the *SNOMED-CT* value *Acute onset (373933003),* see **Table 2.4**.

2.5 Discussion

The IM developed in this project was shown to be adequate for representing GI HDX from clinical cases based on recording the HF that define them. The internal logic of the model allowed us to create an ontology of GI HDX classes based on actual histopathology reports. The IM itself is not determined by the presence or absence of any particular set of attributes. That any particular attributes should be either essential (required) or optional ultimately depends on the purpose to which the model is put. At least one finding site and one morphology were required for representing each HDX from the test set of cases but required only occasional use of attributes for severity, distribution and pathological course. On this basis, we made two attributes mandatory and three attributes optional in our testing ontology. All HDX classes were modeled using two attribute value pairs; "Finding site" and "Abnormal morphology." The other three attributes were treated as optional and no HDX class definitions depended on them. On the other hand, it is not hard to imagine that a research project might set a requirement that severity or distribution or pathological course are recorded; our IM can support any such requirement.

The ontology self-assembled and auto-classified a logical subtype hierarchy. Additionally, the ontology properly classified HF (instances) of greater specificity than the HDX in our test ontology. Inclusion of optional attribute value pairs (such as severity = mild) had no effect on correct classification of instances. An instance of mild gastritis will be correctly retrieved as a member of the gastritis class. Applied to a system for standardizing HDX recording and reporting, this model has the potential to improve retrieval and aggregation of case records based on the HDX.

Developing a terminology to serve as values for the "Abnormal morphology" attribute presented some interesting challenges. The terms that typically represent abnormal morphologies are either named for the process known to have created them (e.g., fibrosis), for the appearance of a well-recognized lesion (e.g. adenocarcinoma), or for a particular cellular pattern (e.g., eosinophilic infiltrate). The stated subtypes in the morphology hierarchy combined with the stated subtypes in the body structure hierarchy control the automatic classification of instances we created for testing purposes. For example, a HF of eosinophilic infiltrate in the gastric mucosa is made a member of the HDX class gastritis because eosinophilic infiltrate is a subtype of inflammatory infiltrate and the gastric mucosa is a stomach structure (subtype). The abnormal

morphology hierarchy also creates the distinction between neoplasia HDX and those of inflammatory conditions.

Light microscopy HF associated with alimentary lymphoma (ALA) and chronic lymphocytic inflammation do not readily distinguish between these two conditions. A HDX of ALA is not always based on a single visual morphology as it is for other neoplasms, such as adenocarcinoma. Infiltration of lymphocytes into gastrointestinal mucosa is a feature of both ALA and chronic lymphocytic infiltration. In order to reach an HDX of ALA, a pathologist evaluates multiple histopathological features of the specimens submitted, which may include such things as epithelial effacement, epithelial injury, and villous stunting. However, none of these histopathologic features confirms the presence of ALA. Whereas the abnormal morphology adenocarcinoma is more easily thought of as a distinct light microscopy HF, an abnormal morphology of ALA is a conclusion based on multiple HF that are variably present.

The difficulty of modeling lymphoma can be solved in multiple ways, none of which is entirely satisfactory. The model could be augmented with an attribute for process by which a pathologist would assert a conclusion based on an assessment of multiple morphologic features. We decided not to alter the model to accommodate a single HDX class. We believe that the set of features that distinguish alimentary lymphoma from inflammatory bowel disease characterized by the presence of lymphocytes needs to be declared and the terminology needs further improvement in order for our model to be used as a decision support tool for HF-based HDX.

We acknowledge that making a diagnosis is always associated with some degree of uncertainty. However, the inability to represent this uncertainty in our IM is one of its limitations and one shared by most terminologies and ontologies. Another limitation is the inability to represent a case that does not show any abnormal morphologies. We followed Cimino's desiderata for medical terminologies and we chose to reject the use of "Not Elsewhere Classified" terms. [21] Terms for "normal" were not included for the same reason; which is that these terms can never have a formal definition other than one dependent on knowledge of all other concepts in the terminology. Therefore, our model does not provide a way of recording "no abnormal morphology." We believe that cases with "no abnormal morphology" or "uncertainty" should be identified by other means at the system level.

As we have stated, the final version of the model we present here imposes a restriction that each combination of a single finding site and a single abnormal morphology be recorded as a distinct HF. The definition of ulcerative colitis, for example is presented in **Table 2.1**. Although both morphologies apply to the colon, the definition is written such that the finding site is duplicated. We considered an alternate approach that would have allowed multiple morphologies to be associated with a single finding site. This would have meant that one less attribute value pair would have been needed to represent ulcerative colitis, but a rule would still be required to preclude combining more than one finding site and more than one abnormal morphology in the same definition.

Some form of grouping is critical for correct retrieval and aggregation of structured HF. The model must be structured so as to maintain associations between attributes and values as they were recorded. This is most easily demonstrated when one considers optional attributes such as severity. It is reasonable to assert that a patient with gastroenteritis might have marked inflammation of one finding site and moderate inflammation of the other. No matter what the combination of severities, the case should still classify as gastroenteritis. This is not possible if free combination of the variables is allowed. In the end, we opted for a single consistent rule. Each HF "group" may include one and only one value to be associated with each attribute and an HDX may include as many HF as necessary to complete its logical definition.

The long term objective of this work is to produce a consistent and structured way to record HDX from endoscopic biopsies. This objective could be achieved by selecting a standardized structured terminology like *SNOMED-CT* provided that it contained the necessary content. *SNOMED-CT* provides a list of concept classes organized by an information model very similar to ours. Unfortunately, the unstructured GI HDX in our case sample set could only be partially mapped to existing *SNOMED-CT finding* concept classes (49 % 1:1 mapping). Results in this study are consistent with a previous one showing that *SNOMED-CT* coverage of veterinary clinical pathology terms in general is limited. [22]

By contrast, the terminology values required to support our model were successfully mapped to values in *SNOMED-CT* with a few specific exceptions. The *SNOMED-CT body structure* hierarchy did not consistently include terms for specific tissue layers and microscopic structures for each organ of the GI tract. As *SNOMED-CT* is an open terminology, a request can be made to the International Health Terminology Standards Development Organisation (IHTSDO)

to add the missing terms. Difficulty in mapping lymphocytic infiltrates is not unexpected. The only concept class in *SNOMED-CT* that seems applicable is *Chronic lymphocytic inflammation (54727009)* which has a synonym *lymphocytic infiltrate*. While we agree that certain instances of lymphocytic infiltration are in fact associated with chronic inflammation (so that this concept exists), it is not true that all lymphocytic infiltrates are produced by inflammation and these terms are not synonyms. The microscopic presentation of ALA suggests that this is simply not the case. Therefore, we decided not to map our concept "lymphocytic infiltrate" to the more specific *SNOMED-CT* concept *Chronic lymphocytic inflammation (54727009).* Instead, as shown in Table 2, the concept class "lymphocytic infiltrate" was provided.

The results of mappings between the terminology we used in our model and *SNOMED-CT* suggest that all but one of our attributes is represented in *SNOMED-CT* and we used them in essentially the same way. This is another way to say that our IM and the *SNOMED-CT* IM are in virtual alignment. We believe that our "Pathological course" attribute is not equivalent to the *SNOMED-CT* attribute *Clinical course*, which creates an exception to this alignment. The "Pathological course" attribute in our model refers only to the duration of the process that produced the HF and the *Clinical course* attribute in *SNOMED-CT* has values of both duration and onset. This also meant that we did not map our term "Acute" to the *SNOMED-CT* quality value *Acute onset*, the closest similar term. We did retain maps to the values subacute and chronic and would use them provisionally. These values may mean the same thing in two different contexts or we may revise our map in the future to represent a different understanding. There was agreement between *SNOMED-CT* and our terminology concerning values for severity and distribution (**Table 2.3**).

Our IM was able to represent 100% of HDX and yielded computationally traceable structured HDX of GI diseases in dogs and cats. Our IM also aligns reasonably well with *SNOMED-CT*'s IM and many of the values required for recording the HDX in our sample set map directly to *SNOMED-CT*. In this way, our model becomes a useful tool to inform the creation of additional *SNOMED-CT* content and its eventual use in medical information systems (**Appendix A** shows concept classes identified by our IM to be added to *SNOMED-CT*).

Although direct application of our IM and our terminology in laboratory information systems is possible, doing so would create a number of problems for sharing information between

institutions. Given the alignment that already exists, we suggest it is preferable to use our IM and terminology to improve *SNOMED-CT* coverage.

The IM we created to represent unstructured HDX in a structured format facilitates the identification of sets of HF that are associated with various HDX. The morphology concepts of our model and *SNOMED-CT morphologically abnormal structure* sub-hierarchy that we used are not limited to the GI tract, thus allowing this approach for reporting of structured HDX to be used at other anatomic sites. In addition, our model can be extended to capture additional information, such as "causative agent". Finally, our IM-based method is amenable to prospective data standardization. Using the concepts classes (**Tables 2.1, 2.2, and 2.3**) as a terminology in the syntax suggested by our IM provides one possible method for standardized reporting of GI HDX.

To conclude, members of the College of American Pathologists in several studies have shown various problems with unstructured pathology reporting formats and have recommended the use of a standard report. [3, 4, 6, 23] Here we describe an IM and a terminology for structuring the endoscopic biopsy reports by using a test set of cases from our institution. An IM was created to organize the standardized histopathology terminology in a way that is amenable to automation, machine-learning, comparison, and retrieval.

Our work demonstrates that HDX can be represented in a structured form. Further, we have demonstrated that logical alignment exists between specific HF and HDX. Finally, we show that with proper additions and refinements, *SNOMED-CT* can eventually serve as an acceptable source of external standardized terminology to support the recording and retrieval of HDX. Taken together the proposed WSAVA standards for recording HF and our semantic model for recording the HDX of the GI tract should lead to more consistent recording of pathology information and eventual integration with electronic medical records. Our approach provides a structured representation of veterinary pathology information that is amenable to integration with information from human pathology (using *SNOMED-CT*), therefore, bringing a new source of knowledge for pathology discoveries. As a future direction, the accuracy of our ontology to record HF and HDX from new cases should be validated.

2.6 Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

2.7 References

1. Leib, M.S., et al., *Endoscopic aspiration of intestinal contents in dogs and cats: 394 cases.* J Vet Intern Med, 1999. **13**(3): p. 191-3.

2. Roth, L., et al., *Comparisons between Endoscopic and Histologic Evaluation of the Gastrointestinal-Tract in Dogs and Cats - 75 Cases (1984-1987).* J Am Vet Med Assoc, 1990. **196**(4): p. 635-638.

3. Gephardt, G.N. and P.B. Baker, *Lung carcinoma surgical pathology report adequacy: a College of American Pathologists Q-Probes study of over 8300 cases from 464 institutions.* Arch Pathol Lab Med, 1996. **120**(10): p. 922-7.

4. Idowu, M.O., et al., *Adequacy of surgical pathology reporting of cancer: a College of American Pathologists Q-Probes study of 86 institutions.* Arch Pathol Lab Med, 2010. **134**(7): p. 969-74.

5. McDonald, C.J. and W.M. Tierney, *Computer-stored medical records. Their future role in medical practice.* JAMA, 1988. **259**(23): p. 3433-40.

6. Zarbo, R.J. and C.M. Fenoglio-Preiser, *Interinstitutional database for comparison of performance in lung fine-needle aspiration cytology. A College of American Pathologists Q-Probe Study of 5264 cases with histologic correlation.* Arch Pathol Lab Med, 1992. **116**(5): p. 463-70.

7. Miller, P.R., et al., *Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews.* Psychiatry Res, 2001. **105**(3): p. 255-64.

8. Miller, P.R., *Inpatient diagnostic assessments: 2. Interrater reliability and outcomes of structured vs. unstructured interviews.* Psychiatry Res, 2001. **105**(3): p. 265-71.

9. Scolyer, R.A., et al., *Pathology of melanocytic lesions: new, controversial, and clinically important issues.* J Surg Oncol, 2004. **86**(4): p. 200-11.

10. Scolyer, R., et al. *Collaboration between clinicians and pathologists: a necessity for the optimal management of melanoma patients*. in *Cancer Forum*. 2005.

11. van der Meijden, M.J., et al., *Development and implementation of an EPR: how to encourage the user.* Int J Med Inform, 2001. **64**(2-3): p. 173-85.

12. Washabau, R.J. *Report from: WSAVA Gastrointestinal Standardization Group*. 2005. [cited 2016 Aug 3]; Available from: http://www.wsava.org/sites/default/files/GI%20Report%202005.pdf.

13. Day, M.J., et al., *Histopathological standards for the diagnosis of gastrointestinal inflammation in endoscopic biopsy samples from the dog and cat: a report from the World Small Animal Veterinary Association Gastrointestinal Standardization Group.* J Comp Pathol, 2008. **138 Suppl 1**: p. S1-43.

14. Cimino, J.J., *Review paper: coding systems in health care.* Methods Inf Med, 1996. **35**(4-5): p. 273-84.

15. Awaysheh, A., et al., *Evaluation of supervised machine-learning algorithms to distinguish between inflammatory bowel disease and alimentary lymphoma in cats.* J Vet Diagn Invest, 2016. **28**(6): p. 679-687.

16. Santamaria, S.L. and K.L. Zimmerman, *Uses of informatics to solve real world problems in veterinary medicine.* J Vet Med Educ, 2011. **38**(2): p. 103-9.

17. Hackos, J.T., *What is an information model and why do you need one.* The Gilbane Report, 2002. **10**(1).

18. AVMA. *SNOMED, HL7, LOINC the official informatics standards for veterinary medicine*. [cited 2016 Aug 3]; Available from: https://www.avma.org/News/JAVMANews/Pages/s020601o.aspx.

19. AAHA. *AAHA Diagnostic Terms*. [cited 2016 Aug 3]; Available from: https://www.aaha.org/professional/resources/diagnostic_terms.aspx.

20. Noy, N.F., et al., *Protege-2000: an open-source ontology-development and knowledge-acquisition environment.* AMIA Annu Symp Proc, 2003: p. 953.

21. Cimino, J.J., *Desiderata for controlled medical vocabularies in the twenty-first century.* Methods Inf Med, 1998. **37**(4-5): p. 394-403.

22. Zimmerman, K.L., et al., *SNOMED representation of explanatory knowledge in veterinary clinical pathology.* Vet Clin Pathol, 2005. **34**(1): p. 7-16.

23. Amin, M.B., *The 2009 version of the cancer protocols of the college of american pathologists.* Arch Pathol Lab Med, 2010. **134**(3): p. 326-30.

Chapter 3 - Identifying free-text features to improve automated classification of structured histopathology reports

3.1 Abstract

The histopathological evaluation of gastrointestinal (GI) biopsies is the reference standard for diagnosis of a variety of GI diseases, e.g., inflammatory bowel disease (IBD) and alimentary lymphoma (ALA). The histopathology report conveys the pathologists' observed microscopic findings and their interpretation as a diagnosis. In companion animals, the World Small Animal Veterinary Association (WSAVA) Gastrointestinal International Standardization Group proposed a reporting standard for GI biopsies consisting of a defined set of semi-quantifiable microscopic features. To our knowledge there has been no formal assessment of information loss in comparison between microscopic descriptions represented as free-text and those in the WSAVA format. In this study, we compare the machine classification accuracy of free-text microscopic findings with those in the WSAVA format for the diagnosis of IBD and ALA in cats. The goal of this comparison is to determine if there are uncaptured free-text histopathological features which provide additional discriminating information to the existing WSAVA list of microscopic features. Retrospectively, 60 cats, with chronic GI disease diagnosed histologically from a duodenal biopsy (endoscopic), as having: IBD (n=20), ALA (n=20), or otherwise normal (n=20), were identified. The microscopic descriptions from these reports were natively expressed in a free-text format, i.e., unstructured. The microscopic duodenal sample from each case was then scored following the WSAVA guidelines by a single pathologist (kz) who was blinded to the original diagnosis. These newly created WSAVA style reports represented the structured text used for comparison with the unstructured originals. The unstructured report diagnosis was used to assign the group category for the newly created structured report. Three supervised machine learning algorithms (naive Bayes, C4.5 decision tree, and artificial neural networks) were trained on the structured and then separately on the unstructured datasets. Diagnosis classification accuracy for each algorithm was measured and compared within and between the structured and unstructured reports. Unstructured information-based models using naïve Bayes and neural networks achieved higher accuracy (0.90 and 0.88, respectively) in predicting the diagnosis when compared to the structured information-based models (0.74 and 0.72, respectively). Results suggest discriminating diagnostic information

is lost when using the current WSAVA microscopic guideline features. Free-text features (such as the number of plasma cells) increased WSAVA auto-classification performance. Inclusion of this feature to the WSAVA GI biopsy reports improved auto-classification accuracy of computer-based diagnostic models built using WSAVA style reports. The methodologies reported in this study represent a way of extracting and selecting candidate features to be used when structuring or classifying reports

Keywords: Histopathology Report; Structured Report; Text Mining; Machine Learning.

3.2 Introduction

The pathology report is an important communication link between pathologists and clinicians. Different pathology reporting formats have shown to have an effect on the message transferred. A study conducted in 1992 by the College of American Pathologist [1] showed that, standardized or checklist reporting practice was the only factor significantly associated with increased likelihood of providing complete or adequate pathology information. The use of different formats in pathology reporting was re-called in 2000, [2] in which authors uncovered a communication gap between pathologists and surgeons as a result of the unfamiliarity with different reporting formats, and they emphasized the need for more complete, clear, and standardized reporting. Researchers also highlighted the need for a controlled vocabulary to improve communication and improve the quality of data reported. A 2004 survey of reports from 96 veterinary clinical pathologists [3] showed that 68 unique terms were used to express probability or likelihood of cytological diagnoses. Another study conducted recently [4] found expressions of uncertainty in 35% of 1500 human surgical pathology reports. Traditional unstructured (free-text) reports can be more detailed, explicit, and representative of real world findings, but they can be incomplete, unclear, and not easily converted into a computable format. [5] Therefore, in healthcare, to increase message clarity, standardize healthcare practice, and increase data interoperability across different systems, clinicians and pathologists are increasingly using structured reporting formats. [6-10]

In 2005, the World Small Animal Veterinary Association (WSAVA) Gastrointestinal (GI) International Standardization Group took the responsibility of standardizing the histologic evaluation of the gastrointestinal tract of cats and dogs. [11] Three years later, [12] the group proposed standards for reporting microscopic findings from GI biopsy samples. The proposed standards were developed to help minimize variation among pathologists' determination of microscopic severity of changes and to ensure consistent reporting of a standardized set of variables. Moreover, using structured recording of microscopic features will provide a basis for the development of a standardized and evidence-based diagnosis.

A 2010 study conducted by the College of American Pathologist [13] showed that institutions that used checklists reported all required pathological elements at a higher rate than those did not use the checklists (88 % versus 34%). This suggests that the use of WSAVA checklist variables to record microscopic findings will ensure capturing most of the required information.

However, to our knowledge there has been no formal assessment of the coverage of WSAVA structured reports of the information needed to make a diagnosis.

The aim of this study was to identify free-text histopathological features not currently expressed in the WSAVA format that may provide evidence for distinguishing between inflammatory bowel disease (IBD) and alimentary lymphoma (ALA) in cats. This aim was examined by comparing auto-classification accuracy of structured and unstructured histopathology reports into IBD and ALA classes using a variety of machine learning algorithms. We hypothesized that WSAVA-based structured reports will have higher classification accuracy of these disorders in comparison to the use of unstructured format.

3.3 Materials and Methods

This retrospective study examined free-text histopathology reports from three groups of 20 (60 total) client-owned cats presented at the Virginia Tech Veterinary Teaching Hospital (VTH) from 2008 to 2015. All cats were patients presented to the VTH with chronic GI disease of weeks or more and had undergone an upper GI endoscopic biopsy procedure. All cases were classified histopathologically as normal, IBD, or small cell ALA by a veterinary pathologist at the VTH.

Only cats with duodenal lymphocytic or plasma cellular inflammation were placed in the IBD group. None of the IBD cats had lymphoma at any other biopsy site. All cats in the ALA group had small cell lymphoma based upon the 2008 World Health Organization (WHO) histopathological classification standards. [14] Cats with ALA may have had a similar or a different diagnosis at another biopsy site.

The original hematoxylin and eosin duodenal biopsy slides for the unstructured cases were retrieved and randomized. The slides were reevaluated and reported in a structured format using the WSAVA guidelines [12] by a single board-certified pathologist (kz) who was blinded to the diagnosis.

Data Mining

Three supervised machine learning models (naïve Bayes, C4.5 decision tree, and artificial neural networks) were developed for each reporting format (structured and unstructured) using the Waikato Environment for Knowledge Analysis (WEKA, version 3.7) data mining software. [15]

This open source software provided tools for data pre-processing, classification, and visualization. Naïve Bayes, J48, and multiple perceptrons are the three classification algorithms in WEKA that implemented the naïve Bayes, C4.5 decision tree, and artificial neural networks algorithms, respectively.

Structured (WSAVA) Data Transformation

Following the WSAVA standards, nine variables were evaluated for each duodenal section including: "villous stunting," "villous epithelial injury," "crypt distension," "lacteal dilation," "mucosal fibrosis," "intraepithelial lymphocytes," "lamina propria lymphocytes and plasma cells," "lamina propria eosinophils," "lamina propria neutrophils." Each variable was recorded as: normal, mild increase, moderate increase, or marked increase. Only cats with moderate or marked duodenal lymphocytic or plasma cellular inflammation were placed in the IBD group. These ordinal values were transformed numerically into 0, 1, 2, and 3, respectively.

Unstructured (Free-Text) Data Transformation

The free-text description of each unstructured report was transformed into a word occurrences vector using the "bag of words" method. [3, 16] In this approach, every document was represented by a set of words (called features) that were extracted from its text. These features were tokenized from the text using the WEKA "AlphabeticTokenizer" algorithm, in which non-alphabetical elements were excluded. A lower case transformation factor was applied to convert all text letters into lower case. Irrelevant, non-informative, stop words (such as "the" and "of") were excluded from the list of tokens by setting WEKA "useStoplist" variable to "True". The words "Neoplasm" and "Neoplasia" were excluded from the text tokens to prevent any bias introduced by using the classification class (diagnosis) as a predictor of it-self.

The data type of the word occurrences vector was defined as a quantitative data type by adjusting WEKA "outputWordCounts" setting to "True". Then, two transformation factors that have shown to improve documents categorization [17] were applied to create a quantifiable, weighted representation of words.

Term frequency transformation factor was applied using WEKA "TFTransform" setting, in which word frequencies were transformed into"$\log(1 + fij)$", where "fij" is the frequency of word "i" in document "j". This weighting factor assumes that the more often a word occurs in a

document, the more representative it is of the text content, so its weight should be of a high power. [16]

Inverse document frequency is the second weighting factor that was used as defined by WEKA "IDFTransform". This setting transforms word frequencies in a document into:

$$fij * \log\left(\frac{number\ of\ all\ documents}{number\ of\ documents\ with\ word\ i}\right)$$

where "fij" is the frequency of word "i" in document "j". This weighting factor considers words that appear in many documents as words that have little discriminating power when used in classification, thus should be less weighted.

Using WEKA "minTermFreq" option, we chose to include the words that occurred at least three times, and this is enforced on a per-class basis. Word frequencies were normalized with the document length using WEKA "Normalize all data" setting. All other parameters' settings were left at their default values unless declared other than that. Table 1 summarizes WEKA parameters' settings used in converting each document's string into a word vector.

Structured and Unstructured Feature Selection

Only subsets of the structured and unstructured extracted features were selected to be used in classification, on the attempt to improve models' performance, minimize processing time, and to avoid "overfitting" (the problem of being so specific to the examined dataset that the training output just describes the one dataset in detail as opposed to characterizing the main generalizable features of the training subset as a whole), as has been shown in previous studies. [16, 18]

The "Best First" [19] searching method was used to create different subsets of features, and "wrapper" [20] was used as an evaluator of each subset. In wrapper, three algorithms were used to evaluate each subset: naïve Bayes, C4.5 decision tree, and artificial neural networks. As a result, three subsets of features (one by each algorithm) were selected from each data format (structured and unstructured).

New Candidate Features for Structured Reports

New candidate features which might improve the classification performance of the structured models were identified by searching for unstructured features not currently part of the

structured reports. These new features were quantified in all the tissue specimens, by averaging results from five different microscopic fields of the lamina propria at 1000x (1.25 oil lens).

Classification Models

Naïve Bayes, C4.5 decision tree, and artificial neural networks algorithms were utilized to create the classification models. The three models were trained on the selected structured, and independently the unstructured features. Then, the same structured data models were retrained after adding new candidate features extracted from the unstructured reports.

To train and test the generated models, a 10-fold cross-validation technique was performed as previously described. [21] In this technique, 10 different datasets were created; in every one of them a different 10% data partition was held out for testing, and the rest of the 90% data was used for training; this results in the testing set never being a part of the training set in any of the ten cases.

In order to exclude any distinction that can be due to an artifact division of training and testing sets, 10 random repeats of 10-fold cross-validation technique were performed. The parameters' settings used in WEKA for classification are shown in **Table 3.1**.

Table 3.1 The parameters' settings of "StringToWordVector" filter in WEKA* used to convert the free-text histopathological strings into a word vector.

Parameter for "StringToWordVector" filter	Setting
IDFTransform	True
TFTransform	True
attributeIndices	1†
attributeNamePrefix	Plain (no entry)
doNotOperatePerClassBasis	False
invertSelection	False
lowerCaseTokens	True
minTermFreq	3
normalizeDocLength	Normalize all data
outputWordCounts	True
periodicPruning	-1.0
Stemmer	NullStemmer
Stopwords	Weka-3-7
Tokenizer	AlphabeticTokenizer
Usestoplist	True
WordsToKeep	1000

* Waikato Environment for Knowledge Analysis
† This setting is the attribute number that carried the string to be converted, which in our case was "1".

As in the literature, we use the term "sensitivity" to mean "classification accuracy." Average sensitivity and confidence intervals for every classifier were considered to assess and compare models' performances. Differences in models performances would be statistically significant at a two-tailed value of $P < 0.05$. (JMP Pro 11 Program, SAS Inc., Cary, NC).

3.4 Results

Features Selected from the Structured Reports

From the original nine features of the structured (WSAVA) reports, selected features were the following, ***using the naïve Bayes***: "villous epithelial injury," "crypt distension," "lacteal dilation," "mucosal fibrosis," "intraepithelial lymphocytes," "lamina propria lymphocytes and plasma cells," "lamina propria neutrophils," ***using the decision tree and the artificial neural networks***: "lamina propria lymphocytes and plasma cells" was the only selected feature

Features Selected from the Unstructured Reports

From the unstructured reports, initial word vector tokenization resulted in the extraction of 74 unique words, see **Table 3.2**. From these words, 18 were selected in the feature selection

approach using the three algorithms, as shown in **Table 3.3**. Frequency of occurrence for the 18 words in association with the class diagnosis (normal, IBD, and ALA) is shown in **Table 3.4**.

Table 3.2 - Shows word features extracted from the free-text histopathological descriptions using "bag of words" methodology before applying any feature selection.

abundant	clusters	expands	inflammatory	mucosa	observed	score
amounts	Cytoplasm	fields	lamina	mucosal	occasional	sections
amphophilic	Dense	figures	large	multifocal	plasma	sheets
approximately	diameter	glands	lymphocytes	muscularis	population	sized
arranged	diffusely	glandular	markedly	neutrophils	present	small
array	distinct	grade	medium	normal	propria	surface
basophilic	eosinophils	increased	mitotic	noted	rare	ten
borders	epithelial	indistinct	moderate	nuclei	roth	villi
cells	epithelium	infiltrate	moderately	nucleoli	round	
central	expand	infiltrating	monomorphic	numbers	scant	
chromatin	expanded	infiltration	monotonous	numerous	scattered	

Table 3.3 - Shows the distribution of words selected across the three algorithms.

All Words Selected	Selected by Naïve Bayes	Selected by C4.5	Selected by Neural Networks
cells	√		
lamina		√	
plasma	√		√
numbers		√	
small		√	
moderate		√	√
round	√	√	√
figures		√	
population	√		
normal	√		√
present		√	
mitotic			√
large	√		√
inflammatory		√	√
expands	√	√	√
medium	√		√
numerous	√		
surface		√	

Table 3.4 Lists words frequency across the three categories of reports (Normal, IBD, and ALA).

Words Selected	Number of Reports with the Word			
	Normal Cases (20)	IBD* Cases (20)	ALA† Cases (20)	Total (60)
cells	6	16	15	37
lamina	10	13	13	36
plasma	5	18	7	30
numbers	4	8	5	17
small	2	3	11	16
moderate	0	8	6	14
round	0	0	11	11
mitotic	0	0	10	10
figures	0	0	9	9
population	0	0	7	7
normal	7	0	0	7
present	2	2	3	7
large	0	1	4	5
inflammatory	0	1	4	5
expands	0	4	0	4
medium	0	3	0	3
numerous	0	0	3	3
surface	0	0	3	3

* Inflammatory Bowel Disease † Alimentary Lymphoma

New Candidate Features Identified

From the unstructured reports, the word "plasma" appeared in 25 % (5/20) of normal cases, in 90 % (18/20) of IBD cases, and in 35 % (7/20) of ALA cases. In the seven ALA cases with plasma cells mentioned, three described *few* and four described *rare* numbers.

The number of plasma cells was a candidate feature that was not independently recorded by the structured (WSAVA) format. Quantifying the number of plasma cells per 5 microscopic fields revealed that the average number of cells was 4.25 (2.22 – 6.28, 95 % Confidence Intervals, CI) in normal cases, 19.1 (14.66 – 23.54, 95 % CI) in cases with IBD, and 7.70 (5.59 – 9.81) in ALA cases.

The words: "round," "population," "surface," "numerous," "figure," "mitosis" were associated with ALA 100% of the times and never appeared in any normal or IBD reports (**Table 3.4**). "Round" described the shape of the cells, and "population" described some distributional characteristics at large such as monotonous or homogeneous population. "Surface" referenced the *epithelium surface* in three reports, with two mentioning *infiltration on the surface*. "Numerous" was used in association with lymphocytes. In all cases "figures" appeared in context with "mitotic" describing cellular mitotic activity. In all reports that recorded "mitotic" (appeared in ALA only), 70% (7/10) of them described *no mitotic activity observed*, 20% recorded *low mitotic*, and one report recorded *rare mitotic activity*.

The feature "small" was associated with ALA (55%, 11/20) more than it was associated with normal (10%, 2/20) or IBD (15%, 3/20) cases. The 11 ALA and two of the three IBD cases described size of the lymphocytes, the rest described small quantities of cells or tissues. For "large", in four ALA cases: one described *large nodules* formed between glands, one described *slightly large nuclei*, and two cases talked about large numbers of lymphocytes, and the only one IBD case with "large" described *large lymphocytes* and *large nucleus*.

Classification Models' Performance

For classification performance, **Table 3.5** shows performance sensitivity rates resulting from the three classification models using structured and unstructured reporting formats. Two-tailed t-tests showed that sensitivities achieved by naïve Bayes and neural network classifiers using the unstructured reports (0.898 and 0.883, respectively) were higher than using the structured reports (0.735 and 0.717, respectively, $P < .0001$). **Table 3.6** shows sensitivity rates resulting from the classification models after adding the "lamina propria plasma cells" feature to the structured

reports. Two-tailed t-tests showed that sensitivity rates increased using the structured reports after adding the "lamina propria plasma cells" feature (0.792, 0.770, and 0.782 compared to 0.735, 0.717, and 0.717, respectively for naïve Bayes, decision tree, and neural networks, $P < .05$).

Table 3.5 Shows sensitivity (classification accuracy) when applying the three classifiers on the structured and unstructured datasets.

	Structured Reports (WSAVA*)					
	Naïve Bayes		C4.5 Decision Tree		Artificial Neural Networks	
	Sensitivity	95 % Confidence	Sensitivity	95 % Confidence	Sensitivity	95 % Confidence
Normal	0.845	(0.797 – 0.893)	0.750	(0.696 – 0.804)	0.750	(0.696 – 0.804)
IBD†	0.660	(0.590 – 0.730)	0.650	(0.582 – 0.718)	0.650	(0.582 – 0.718)
ALA‡	0.700	(0.635 – 0.765)	0.750	(0.689 – 0.812)	0.750	(0.689 – 0.812)
Average	**0.735**§	**(0.696 – 0.774)**	**0.717**	**(0.678 – 0.755)**	**0.717**‖	**(0.678 – 0.755)**
	Unstructured Reports (free-text)					
Normal	0.850	(0.802 - 0.898)	0.845	(0.795 – 0.895)	0.840	(0.789 – 0.891)
IBD	0.950	(0.920 - 0.980)	0.695	(0.628 – 0.762)	0.990	(0.970 – 1.00)
ALA	0.895	(0.850 – 0.940)	0.725	(0.661 – 0.789)	0.820	(0.764 – 0.876)
Average	**0.898**§	**(0.873 – 0.924)**	**0.755**	**(0.722 – 0.788)**	**(0.883)**‖	**(0.858 – 0.908)**

* World Small Animal Veterinary Association †Inflammatory Bowel Disease ‡ Alimentary Lymphoma §‖ each sign represents a statistically different pair (two-tailed t-test, $P < .0001$).

Table 3.6 Shows sensitivity (classification accuracy) when applying the three classifiers on the structured datasets after adding the "plasma cells" feature into the learning set.

	Structured Reports (WSAVA*) + "Plasma cells" Feature					
	Naïve Bayes		C4.5 Decision Tree		Artificial Neural Networks	
	Sensitivity	95 % Confidence	Sensitivity	95 % Confidence	Sensitivity	95 % Confidence
Normal	0.850	(0.794 – 0.906)	0.890	(0.844 – 0.936)	0.855	(0.804 – 0.906)
IBD†	0.845	(0.793 – 0.897)	0.680	(0.613 – 0.750)	0.780	(0.712 – 0.845)
ALA‡	0.680	(0.618 – 0.742)	0.740	(0.680 – 0.800)	0.710	(0.655 – 0.765)
Average	**0.792**§	**(0.759 – 0.824)**	**0.770**‖	**(0.738 – 0.802)**	**0.782**¶	**(0.751 – 0.812)**
	Recalling Values from Table 5 Structured Reports (without "Plasma cells" Feature)					
Average	**0.735**§	**(0.696 – 0.774)**	**0.717**‖	**(0.678 – 0.755)**	**0.717**¶	**(0.678 – 0.755)**

* World Small Animal Veterinary Association †Inflammatory Bowel Disease ‡ Alimentary Lymphoma §‖¶ each sign represents a statistically different pair (two-tailed t-test, $P < .05$).

3.5 Discussion

The models developed in our study to classify Normal, IBD, and ALA classes using the unstructured histopathology reports showed that models were able to achieve very good performance (sensitivity ranging from 76 % to 90 %). The classification models developed using data from WSAVA structured reports demonstrated lower performance (sensitivity ranging from 72% to 74%).

Frequency of occurrence analysis on the features extracted from the unstructured reports showed that the term "plasma" was more commonly associated with IBD (90 % of cases, 18/20) than normal (25 % of cases, 5/20) or ALA cases (35 %, 7/20). The seven ALA finding descriptions with "plasma" described either few or rare number of plasma cells. Moreover, our study showed that recording the number of plasma cells in conjunction with some of the nine WSAVA variables improved the classification accuracy of the three models (**Table 3.6**). These findings are consistent with the literature, which shows that ALA is characterized by the infiltration of lymphocytes, [22, 23] unlike IBD which is represented by the infiltration of lymphocytes and plasma cells. [24, 25] The current WSAVA standards do not distinguish between the numbers of lymphocytes and the number of plasma cells when reporting; one variable called "lamina propria lymphocytes and plasma cells" is being used to represent the infiltration of any of the two cell types.

The mitotic activity was another feature selected as a predictor variable to distinguish between the three groups. However, all reports described *low* or *rare* mitosis and while reviewing the tissue samples no mitotic activity was identified; this finding is not surprising given that all of the ALA cases were of small cell type and likely an indolent form of lymphoma. Although the mitotic activity did not prove a good candidate in our study, we believe that recording of this variable may have importance in other cases, such as large cell type lymphoma, where the mitotic activity may be more notable.

The Naïve Bayes, decision trees, and artificial neural networks are common examples of machine learning algorithms that can exploit underlying complex patterns in large datasets to classify cases into related groups. By using them in classification and prediction, these three algorithms have shown to support the decision making process in many areas of medicine. [26-28] Our study showed that, supervised machine learning models developed using naïve Bayes and neural networks achieve performance similar or higher than that achieved using the C4.5 decision tree. The naïve Bayes algorithm is known to assume conditional independence of features, [29] which is not always true. Therefore, the naïve Bayes algorithm doesn't perform as well as other classifiers when the training features are dependent on each other. Our results showed that the naïve Bayes outperformed the decision tree classifier when learning from the free-text, suggesting a reasonable independency assumption on free-text pathological features. A previous study of diagnostic models [30] has shown similar results in which the conditional independency assumption of the naïve Bayes did not have any negative effects on the performance.

Despite the relative similarities in the performance of the three algorithms, some other factors should be considered when comparing algorithms. As previously mentioned, the naïve Bayes assumes independency between features, this assumption makes the algorithm simple and computationally inexpensive (which was clearly the motivation for the development of the naïve version of the Bayes algorithm). [29, 31] However, while the decision tree model output is easy to understand and implement, it can be inefficient in solving complex problems. On the other hand, the multilayer artificial neural network algorithm has the ability of easily capturing the complexity of the relationships between features during its leaning phase at a higher computational cost and with final results that make the underlying logic harder to understand. [16, 32]

In features selection, our results showed that only 18 of the 74 extracted words from the free-text were needed to classify the reports into the three classes. This finding is consistent with a recent feature selection study [33] which showed that performances of classification models either improve or remain unchanged despite the removal of 75% of the features. Our study recalls the importance of selecting the subset of features needed for classification, which leads to an improvement in the models' performance while minimizing the training time, data dimensionality, and the application complexity.

The reports selected in this study represent a corpus of simple cases for which no diseases other than IBD and ALA were considered. This selection represents a good example of testing the ability to classify histopathology reports into two of the most common GI diseases in companion animals. However, a larger corpus needs to be examined in the future to validate the results of this study and extract other features that can be noteworthy in other diseases. In this study the free-text reports were created by 14 pathologists, and goodness of fit testing indicated they were not randomly associated with the three classes of reports. The bias that can be introduced by the reporting style of one or more pathologists might be one limitation of this study that can be overcome by utilizing a larger number of cases as well.

Although our study focused on histopathology reports for GI diseases, we believe that this work can be extended to help assist in the creation of standardized histopathology reports involving other body structures and disorders. The methodologies reported in this study represent a way of extracting and selecting candidate features to be used when structuring or classifying reports. It can also serve as a quality assurance process to highlight any information loss when changing reporting formats.

Finally, a previous study [34] has shown that structured reporting formats provide information that is considered to be simple and computable. Another study [1] found that structured pathology reports provide complete or adequate pathology information. A more recent study [2] has shown a communication gap between clinicians and pathologists, and then concluded that the old, less structured, reporting format is associated with less misinterpretation rate by clinicians and further improvement of the structured reports is essential. Consistent with the Powsner study, our results showed that machine learning models achieved a higher accuracy in classifying unstructured reports into the actual diagnosis when compared with structured reports. These results suggest that discriminating information is lost when using just those features currently listed in WSAVA's GI biopsy reporting format. Therefore, further recording of some of the features identified from the unstructured reports, such as the number of plasma cells, in conjunction with the WSAVA features will improve auto-classification for reports of this type.

3.6 Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

3.7 Funding

The author(s) declared that they received no financial support for their research and/or authorship of this article.

3.8 References

1. Zarbo, R.J. and C.M. Fenoglio-Preiser, *Interinstitutional database for comparison of performance in lung fine-needle aspiration cytology. A College of American Pathologists Q-Probe Study of 5264 cases with histologic correlation.* Arch Pathol Lab Med, 1992. **116**(5): p. 463-70.

2. Powsner, S.M., J. Costa, and R.J. Homer, *Clinicians are from Mars and pathologists are from Venus.* Arch Pathol Lab Med, 2000. **124**(7): p. 1040-6.

3. Christopher, M.M. and C.S. Hotz, *Cytologic diagnosis: expression of probability by clinical pathologists.* Vet Clin Pathol, 2004. **33**(2): p. 84-95.

4. Lindley, S.W., E.M. Gillies, and L.A. Hassell, *Communicating diagnostic uncertainty in surgical pathology reports: disparities between sender and receiver.* Pathol Res Pract, 2014. **210**(10): p. 628-33.

5. Branavan, S.R.K., et al., *Learning Document-Level Semantic Properties from Free-Text Annotations.* Journal of Artificial Intelligence Research, 2009. **34**: p. 569-603.

6. Gephardt, G.N. and P.B. Baker, *Lung carcinoma surgical pathology report adequacy: a College of American Pathologists Q-Probes study of over 8300 cases from 464 institutions.* Arch Pathol Lab Med, 1996. **120**(10): p. 922-7.

7. Miller, P.R., *Inpatient diagnostic assessments: 2. Interrater reliability and outcomes of structured vs. unstructured interviews.* Psychiatry Res, 2001. **105**(3): p. 265-71.

8. Miller, P.R., et al., *Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews.* Psychiatry Res, 2001. **105**(3): p. 255-64.

9. Scolyer, R., et al. *Collaboration between clinicians and pathologists: a necessity for the optimal management of melanoma patients*. in *Cancer Forum*. 2005.

10. Scolyer, R.A., et al., *Pathology of melanocytic lesions: new, controversial, and clinically important issues.* J Surg Oncol, 2004. **86**(4): p. 200-11.

11. Washabau, R.J. *Report from: WSAVA Gastrointestinal Standardization Group*. 2005. [cited 2016 Aug 3]; Available from: http://www.wsava.org/sites/default/files/GI%20Report%202005.pdf.

12. Day, M.J., et al., *Histopathological standards for the diagnosis of gastrointestinal inflammation in endoscopic biopsy samples from the dog and cat: a report from the World Small Animal Veterinary Association Gastrointestinal Standardization Group.* J Comp Pathol, 2008. **138 Suppl 1**: p. S1-43.

13. Idowu, M.O., et al., *Adequacy of surgical pathology reporting of cancer: a College of American Pathologists Q-Probes study of 86 institutions.* Arch Pathol Lab Med, 2010. **134**(7): p. 969-74.

14. Jaffe, E.S., *The 2008 WHO classification of lymphomas: implications for clinical practice and translational research.* Hematology Am Soc Hematol Educ Program, 2009: p. 523-31.

15. Hall, M., et al., *The WEKA data mining software: an update.* ACM SIGKDD explorations newsletter, 2009. **11**(1): p. 10-18.

16. Sebastiani, F., *Machine learning in automated text categorization.* Acm Computing Surveys, 2002. **34**(1): p. 1-47.

17. Jouhet, V., et al., *Automated classification of free-text pathology reports for registration of incident cases of cancer.* Methods Inf Med, 2012. **51**(3): p. 242-51.

18. Blum, A.L. and P. Langley, *Selection of relevant features and examples in machine learning.* Artificial Intelligence, 1997. **97**(1-2): p. 245-271.

19. Korf, R.E., *Linear-space best-first search.* Artificial Intelligence, 1993. **62**(1): p. 41-78.

20. Kushmerick, N., *Wrapper induction for information extraction*. 1997, University of Washington.

21. Kohavi, R., *A study of cross-validation and bootstrap for accuracy estimation and model selection*, in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. 1995. p. 1137-1145.

22. Barrs, V.R. and J.A. Beatty, *Feline alimentary lymphoma: 1. Classification, risk factors, clinical signs and non-invasive diagnostics.* J Feline Med Surg, 2012. **14**(3): p. 182-90.

23. Carreras, J.K., et al., *Feline epitheliotropic intestinal malignant lymphoma: 10 cases (1997-2000).* J Vet Intern Med, 2003. **17**(3): p. 326-31.

24. Jergens, A.E., et al., *Idiopathic inflammatory bowel disease in dogs and cats: 84 cases (1987-1990).* J Am Vet Med Assoc, 1992. **201**(10): p. 1603-8.

25. Willard, M.D., *Feline inflammatory bowel disease: a review.* J Feline Med Surg, 1999. **1**(3): p. 155-64.

26. Abbass, H.A., *An evolutionary artificial neural networks approach for breast cancer diagnosis.* Artif Intell Med, 2002. **25**(3): p. 265-81.

27. Al-Omari, F.A., et al., *An intelligent decision support system for quantitative assessment of gastric atrophy.* J Clin Pathol, 2011. **64**(4): p. 330-7.

28. Zhang, X., et al., *Ontology driven decision support for the diagnosis of mild cognitive impairment.* Comput Methods Programs Biomed, 2014. **113**(3): p. 781-91.

29. Lewis, D.D., *Naive (Bayes) at forty: The independence assumption in information retrieval*, in *Machine learning: ECML-98*. 1998, Springer. p. 4-15.

30. Zelic, I., et al., *Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries.* J Med Syst, 1997. **21**(6): p. 429-44.

31. Zhang, H., *The optimality of naive Bayes*, in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*. 2004. p. 3.

32. Rezaei-Darzi, E., et al., *Comparison of two data mining techniques in labeling diagnosis to Iranian pharmacy claim dataset: artificial neural network (ANN) versus decision tree model.* Arch Iran Med, 2014. **17**(12): p. 837-43.

33. Noureldien, N.A., R.A. Hussain, and A. Khalid, *The Effect of Feature Selection on Detection Accuracy of Machine Learning Algorithms.* International Journal of Engineering Research & Technology 2013. **2**(11).

34. Leslie, K.O. and J. Rosai, *Standardization of the surgical pathology report: formats, templates, and synoptic reports.* Semin Diagn Pathol, 1994. **11**(4): p. 253-7.

Chapter 4 - The use of supervised machine learning algorithms to distinguish between inflammatory bowel disease and alimentary lymphoma in cats

4.1 Abstract

Inflammatory bowel disease (IBD) and alimentary lymphoma (ALA) are common gastrointestinal diseases in cats. They present very similar clinical signs and histopathologic features that make the distinction between them hard. We tested the use of supervised machine learning algorithms to differentiate between the two diseases using data generated from non-invasive diagnostic tests. Three prediction models were developed using three machine learning algorithms which were: naïve Bayes, decision trees, and artificial neural networks. The models were trained and tested on data from Complete Blood Count (CBC) and Serum Chemistry (SC) results for three groups of client-owned cats that were: Normal, inflammatory bowel disease (IBD), or alimentary lymphoma (ALA). Naïve Bayes and artificial neural networks achieved higher classification accuracy (sensitivities of 70.8 % and 69.2 %, respectively) than the decision tree algorithm (63 %, $P < .0001$). The area under the receiver operating characteristic (ROC) curves for classifying cases into the 3 categories was 83 % by the naïve Bayes, 79% by the decision tree, and 82 % by the artificial neural networks. Prediction models using machine learning provided a method for distinguishing between ALA – IBD, ALA – Normal, and IBD - Normal. The naïve Bayes and artificial neural networks classifiers used 10 and 4 CBC/SC variables, respectively, to outperform the C4.5 decision tree which used 5 CBC/SC variables in classifying cats into the three classes. These models can provide another non-invasive diagnostic tool to assist clinicians with differentiating between IBD and ALA, and between diseased and non-diseased cats.

Key words: Diagnosis; Inflammatory Bowel Disease; Lymphoma; Machine Learning.

4.2 Introduction

Inflammatory bowel disease (IBD) and alimentary lymphoma (ALA) are common gastrointestinal (GI) disorders in cats. [1-5] The clinical signs of the two disorders are very similar, with the most common symptoms being anorexia, weight loss, vomiting and/or diarrhea. [1, 3, 4, 6] The chance of developing intestinal lymphoma is increased with a previous diagnosis of IBD, as previously seen in humans, dogs, and cats. [2, 7, 8] Currently, the diagnosis of IBD and ALA relies on the histopathological examination of tissue biopsy specimens taken from the small intestine. [9-11]

Histologically, IBD in most cats is characterized by the accumulation of lymphocytes and plasma cells, and less often neutrophils, eosinophils, or macrophages, in the lamina propria of the stomach and small intestine. [2, 3] Similarly, ALA can be associated with the infiltration of lymphoid cells into these same anatomic locations. However, unlike IBD, the infiltrating lymphoid cells in ALA are not limited to the lamina propria of the mucosa but can also involve and efface the epithelial lining, submucosa, tunica muscularis, and serosa. [1, 4, 6]

Histopathological evaluation of endoscopic and full thickness biopsy samples has been the reference standard for diagnosing of IBD and ALA in cats. However, in some cases it can be challenging for the pathologists to differentiate between small cell lymphoma and lymphoplasmacytic enteritis/gastritis, particularly with endoscopic samples. [10, 12-14] This is due to the similar, relevant clinical signs and the overlapping histologic features between the two diseases, especially when neoplastic cells do not extend beyond the mucosa. [15] To some extent, this challenge can be overcome by collecting a full-thickness intestinal wall tissue sample. However, this approach requires a more invasive procedure in which the patient is exposed to a higher morbidity risk due to surgical complications and is associated with higher financial cost for the owner. [15, 16]

For GI diseases, and specifically IBD and ALA, previous studies [17-19] have shown some common hematology and biochemistry abnormalities. Hypoalbuminemia, the most common serum biochemical abnormality in cats with lymphoma, [7, 20] is thought to be due to the disruption of the intestinal wall integrity. Therefore, albumin is one possible analyte that can help discriminate between IBD and ALA. We speculate that there are more complex data patterns involving blood hematology and blood serum chemistry results that can be used to differentiate between these two disorders.

Data mining is the process of using computational techniques to convert raw data into useful information for knowledge discovery. The main goal is to identify patterns and detect hidden pieces of information. System models that employ classification algorithms are one of the most common tools in data mining. Each of these models takes a set of instances as an input, in which every instance belongs to a particular class. The model then outputs a classifier schema that can predict the class of a new instance given its attributes' values with a particular accuracy in a process called machine learning. [21-23]

Naïve Bayes, decision trees, and artificial neural networks are examples of common machine learning algorithms that can exploit underlying complex patterns in large datasets to classify cases into related groups. Naïve Bayes, decision trees, and artificial neural networks have been applied in many areas of medicine [24-26] to distinguish between different disorders and have shown good classification accuracy. These tools are initially trained using case controlled data representing the disorders of interest; they discover the underlying unique pattern for each disorder. After training they use the experiential knowledge to classify new cases into the appropriate disorder category.

Naïve Bayes classifier is one of the best known simple classifiers. [24] It is derived from Bayes' theorem, but unlike Bayesian classifier the naïve Bayes allows for a computationally inexpensive learning technique while keeping an accurate classification under the assumption that the random variables (attributes) are conditionally independent from each other.

Decision trees have proven to be an effective method for classification. [23, 24] They are among the few classifiers that have incorporated visualization and user interaction for an easier classification at the point of practice. Decision trees do not require any assumption of independence between features (laboratory variables). Every decision tree is composed of nodes; the first tree node is called the root node, and children nodes are referred to as internal nodes. Each of these internal nodes carries a particular test used to classify instances, e.g. "Is the albumin level high?" The nodes at the end of the tree are termed leaf nodes and they are used to identify the instances class which will be the target (diagnosis). [21, 23]

Artificial neural networks are computational models that mimic the structure and the function of biological neural networks. In this approach, knowledge is acquired through a learning process called backpropagation and stored within the interconnection strength between nodes (neurons). An artificial neural networks algorithm using a multilayer perceptron is a well-known

classification algorithm that has been shown [26-29] to perform very well in many areas including medicine. Moreover, a previous review of artificial intelligence applications to manage urological cancer [30] found that an artificial neural network consisting of a multilayer perceptron (comprising 3 neural layers) was the most successful model for improving diagnosis.

The objective of this study is to model the influence of IBD and ALA on different complete blood count (CBC) and serum chemistry (SC) variables and to help distinguish between the two diseases using naïve Bayes, decision trees, and artificial neural networks classification algorithms.

4.3 Materials and Methods

This retrospective study examined 3 groups of 40 (120 total) client-owned cats presented at the Virginia Tech Veterinary Teaching Hospital (VTH) from 2008 to present. The groups were identified as normal, IBD, and ALA. The normal cats were a subset of clinically normal patients which had been used to establish normal reference interval CBC[a] and SC[b] results at the Virginia Tech Animal Laboratory Services (ViTALS), Clinical Pathology Service. The IBD and ALA cats were patients presented to the VTH with chronic GI disease and had undergone an upper GI endoscopic biopsy procedure. Duodenal endoscopic biopsy results were used to sort these patients into the IBD or ALA group if they met the following inclusion criteria. Only cats with lymphocytic or plasma cellular inflammation of moderate or marked severity were included in the IBD group. None of these cats had lymphoma identified in any other biopsy sites. Based on the 2008 World Health Organization (WHO) classification standards for lymphomas [31] the ALA group of cats had either a small or large cell lymphoma diagnosis at the duodenal site and may have had a similar or other diagnosis at another sample site. The original attending pathologist's diagnosis was used to classify these cats into their appropriate group. The duodenal biopsy samples were randomized and reviewed by a single board certified pathologist (author KZ) who was blinded to the final diagnosis. The samples were scored by the pathologist following the GI endoscopic biopsy reporting guidelines of the World Small Animal Veterinary Association (WSAVA) GI International Standardization Group. [14]

Twelve CBC and 19 SC independent variables (31 total, all quantitative variables) were evaluated in all 120 cats. Distribution normality of the CBC and SC variables was assessed across different groups by Shapiro–Wilk test. The non-parametric Kruskal-Wallis[c] test without any distributional assumptions was used for group comparison. The analysis was followed by a non-

parametric comparison for each diagnosis pair using Wilcoxon[c]. Results were considered statistically significant at a two-tailed value of $P < 0.05$.

Data Mining

Three classification algorithms were examined using the data mining software WEKA[d]. This open source software [21] provides tools for data pre-processing, classification, regression, clustering, and visualization. *Naïve Bayes*, *J48*, and *multiple perceptron* are three classification algorithms in WEKA that implement the naïve Bayes, C4.5 decision tree, and artificial neural networks algorithms, respectively.

Attribute Selection

A subset of data attributes (CBC and SC variables) was selected for use with each of the classifier algorithms. The aim of the attribute selection was to remove variables with minimal diagnostic weight from the models to minimize any confusion that could be introduced by them, simplify the generated models, minimize the over-fitting of the classifiers to our dataset (make the classification models less specific so they perform well in new examples), and save classification cost with regard to time and computation load. [32] Before the creation of classification models, the list of CBC and SC variables was preprocessed and filtered using WEKA *BestFirst* searching method (to create different attribute subsets) and WEKA *Wrapper* evaluator (to evaluate the performance of every subset). In *Wrapper*, three subsets of the CBC and SC variables were selected using the 3 previously mentioned classification algorithms as part of the subsets performance evaluation method.

The parameters' settings used in WEKA for attribute selection using *Wrapper* are shown in **Table 4.1**. All other parameters' settings in this study were left at their default values unless declared other than that.

Table 4.1 The parameters' settings of Wrapper attribute selection evaluators in Waikato Environment for Knowledge Analysis (WEKA). Three data subsets were created using the same settings but with different classifiers in evaluation.

Parameter for *Wrapper* attribute selection	**Setting**
-IRClassValue	Blank
-Classifier	*(1) Naïve Bayes (2) J48 (3) multiple perceptron**
-Classifier Settings	Default
-evaluationMeasure	Default
-folds	10
-seed	1
-threshold	0.01

* Three classifiers were used to estimate the accuracy of subsets, one at a time. The subset with the higher accuracy was selected with each classifier (total of three subsets).

Classification Algorithms

Naïve Bayes, C4.5 decision tree, and artificial neural networks classification algorithms were utilized to model the data. Three models for diagnosis prediction were trained using the selected attributes. The parameters' settings used in WEKA for classification are shown in **Table 4.2**.

Table 4.2 The parameters' settings of the three classification models (Naïve Bayes, J48 decision tree, and Multilayer Perceptron artificial neural networks) in Waikato Environment for Knowledge Analysis (WEKA).

Model - Parameter	**Setting**	**Model - Parameter**	**Setting**
Naïve Bayes	False	Multilayer Perceptron (artificial neural networks)	
Debug	False	Graphical User Interface	False
displayModelOldFormat	False	autoBuild	True
useKernelEstimator	False	debug	False
useSupervisedDiscretization		decay	False
J48 (C4.5 decision tree)	False	hiddenLayers	a
binarySplits	0.25	learningRate	0.3
confidenceFactor	False	momentum	0.2
debug	2	nominalToBinaryFilter	True
minNumObj	3	normalizeAttributes	True
numFolds	False	normalizeNumericClass	True
reducedErrorPruning	False	reset	True
saveinstanceData	1	seed	0
seed	True	trainingTime	500
subtreeRaising	False	validationSetSize	0
unpruned	False	validationThreshold	20
useLaplace			

To train and test the generated models, a 10-fold cross validation technique was performed as previously described. [33] In this technique, the 120 instances were split into training (90%) and testing sets (10%), in which the testing set was not part of the training set. In the 10-fold cross validation, 10 different datasets were created; in every one of them a different 10% data partition

was held out for testing, and the rest of the 90% data was used for training. The models' performance was assessed using values of sensitivity, specificity, and area under the receiver-operating characteristic curve (AUC).

In order to exclude any distinction that could be due to an artifact division of training and testing sets, 10 random repeats of 10-fold cross validation technique were performed. Average sensitivity and confidence intervals for every classifier were considered and used for classifiers' performance comparisons.

4.4 Results

The distribution of the cats in the groups according to their sex and disease was: 16 males and 24 females in the normal group, 21 males and 19 females in the IBD group, and 28 males with 12 females in the ALA group.

The median age was 6 years (range, 1-15 years) in the normal cats, 10 years (range, 1-17 years) for cats with IBD, and 13 years (range, 4-18 years) for cats with ALA. Non-parametric comparisons' tests revealed that the median age was significantly higher in cats with ALA (Median=12.5 years, σ=2.51 years) compared to the IBD group (Median=10 years, σ=4.41 years, P =.0043), and significantly higher in the IBD compared to the normal cats (Median=6 years, σ=4.57 years, P = .0057).

Nineteen of the 31 CBC and SC variables were not normally distributed on the normal group, along with 22 of the IBD variables, and 16 of the ALA variables. A non-parametric Kruskal-Wallis test revealed 16 CBC and SC variables statistically different across the three groups (normal, IBD, and ALA). **Table 4.3** shows median and range values for the variables that were statistically different across different groups. In **Table 4.4**, a non-parametric each-paired Wilcoxon test shows the CBC and SC variables that were different across each diagnosis pair (Normal - IBD, Normal - ALA, and IBD - ALA). Variables were selected using the *Wrapper* attribute selection method, using naïve Bayes, decision tree, and artificial neural networks as evaluators, as shown in **Table 4.4**.

Table 4.3 Cohort comparison of complete blood count and serum chemistry data; only statistically different variables are shown

Variable	Unit	Median			Range			Non-parametric Wilcoxon
		Normal	IBD*	ALA†	Normal	IBD	ALA	
Hematocrit	L/L	0.40	0.36	0.34	0.33 – 0.47	0.17 – 0.45	0.19 – 0.45	P < .0001
Hemoglobin	g/L	143	121	118	121 – 166	53 – 149	65 – 148	P < .0001
Red Blood Cells Count	$\times 10^{12}$/L	9.32	7.76	7.53	6.86 – 11.27	2.60 – 9.98	4.05 – 10.26	P < .0001
White Blood Cells Count	$\times 10^{9}$/L	9.44	10.92	14.32	3.83 – 15.39	3.77 – 28.55	5.4 – 49.83	P = .0010
Mean Corpuscular Volume	fL	43.45	46.05	46.65	38.40 – 51.10	34.60 – 64.60	37.00 – 56.50	P = .0041
Mean Corpuscular Hemoglobin Concentration	g/L	355	336	336	339 – 373	297 – 368	302 – 381	P < .0001
Neutrophils	$\times 10^{9}$/L	5.17	7.26	9.62	2.74 – 11.84	2.90 – 25.12	3.38 – 4.86	P < .0001
Lymphocytes	$\times 10^{9}$/L	2.54	1.48	1.61	0.44 – 4.95	0.33 – 9.72	0.10 – 6.30	P = .0001
Monocytes	$\times 10^{9}$/L	0.18	0.33	0.38	0.07 – 0.70	0.08 – 0.91	0.07 – 2.52	P = .0011
Total Protein	g/L	76.0	72.0	70.0	20.0 – 87.0	48.0 – 119.0	48.0 – 96.0	P = .0206
Albumin	g/L	35.0	31.5	30.0	29.0 – 77.0	16.0 – 41.0	20.0 – 41.0	P < .0001
Calcium	mmol/L	2.44	2.38	2.30	2.25 – 3.25	1.88 – 2.85	1.95 – 2.70	P < .0001
Potassium	mmol/L	3.70	3.90	4.10	3.20 – 5.10	2.80 – 6.30	3.00 – 5.40	P = .0096
Sodium	mmol/L	152	151	150	149 – 155	143 – 156	145 – 158	P = .0317
Chloride	mmol/L	119	118	119	116– 150	110 – 124	101 – 135	P = .0037
Cholesterol	mmol/L	3.52	3.42	3.00	2.18 – 6.81	1.27 – 9.56	1.50 – 5.62	P = .0188

* Inflammatory Bowel Disease. † Alimentary Lymphoma.

Table 4.4 Each-paired comparison of complete blood count and serum chemistry data according to the diagnosis groups. Variable is only shown if it was different across all pairs and between any pair, or if it was selected by the attribute selection approach

Variable	Non-parametric each-paired Wilcoxon			Attributes selected using *Wrapper* attribute selection		
	Normal-IBD*	Normal-ALA†	IBD-ALA	Naïve Bayes	Decision Tree	Neural Networks
Hematocrit	P < .0001	P < .0001	-		√	
Hemoglobin	P < .0001	P < .0001	-	√	√	√
Red Blood Cells Count	P < .0001	P < .0001	-			
White Blood Cells Count	-	P = .0002	-			
Mean Corpuscular Volume	P = .0334	P = .0013	-			
Mean Corpuscular Hemoglobin Concentration	P < .0001	P < .0001	-	√		√
Neutrophils	P = .0012	P < .0001	-			
Lymphocytes	P = .0017	P = .0011	-	√		
Monocytes	P = .0043	P = .0010	-		√	√
Total Protein	P = .0383	P = .0066	-			
Albumin	P = .0001	P < .0001	-			
Calcium	P = .0122	P < .0001	-			
Sodium	P = .0237	P = .0234	-	√		
Potassium	-	P = .0028	-	√		
Chloride	P = .0005	-	P = .0302			√
Cholesterol	-	P = .0064	P = .0408	√		
Mean Corpuscular Hemoglobin	-	-	-		√	
Glucose	-	-	-	√		
Phosphorus	-	-	-	√		
Urea Nitrogen	-	-	-	√		
Indirect Bilirubin	-	-	-	√	√	

* Inflammatory Bowel Disease. † Alimentary Lymphoma.

The first run of the 3 models using the 120 instances on a 10-fold cross validation to split data into training and testing sets revealed an AUC of 83.2 % using the naïve Bayes classifier, 79.0 % AUC using the decision tree classifier, and 82.0 % AUC using the artificial neural networks (**Table 4.5**). ROC curves of naïve Bayes, C4.5 decision tree, and artificial neural networks are presented in **Figure 4.1** (A, B, and C, respectively). The decision tree generated by the C4.5 model trained on the 120 instances is shown in **Figure 4.2**.

Table 4.5 Shows sensitivity, specificity, and area under the curves from the receiver-operating characteristic graphs for the three classifiers.

	Naïve Bayes			C4.5 Decision Tree			Artificial Neural Networks		
	Sensitivity	Specificity	AUC*	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
Normal	0.925	0.904	0.954	0.750	0.876	0.863	0.850	0.898	0.943
IBD†	0.375	0.948	0.723	0.500	0.796	0.761	0.775	0.766	0.793
ALA‡	0.900	0.711	0.818	0.650	0.768	0.747	0.450	0.874	0.723
Average	**0.733**	**0.854**	**0.832**	**0.633**	**0.813**	**0.790**	**0.692**	**0.846**	**0.820**

* Area Under the Curve. † Inflammatory Bowel Disease. ‡ Alimentary Lymphoma

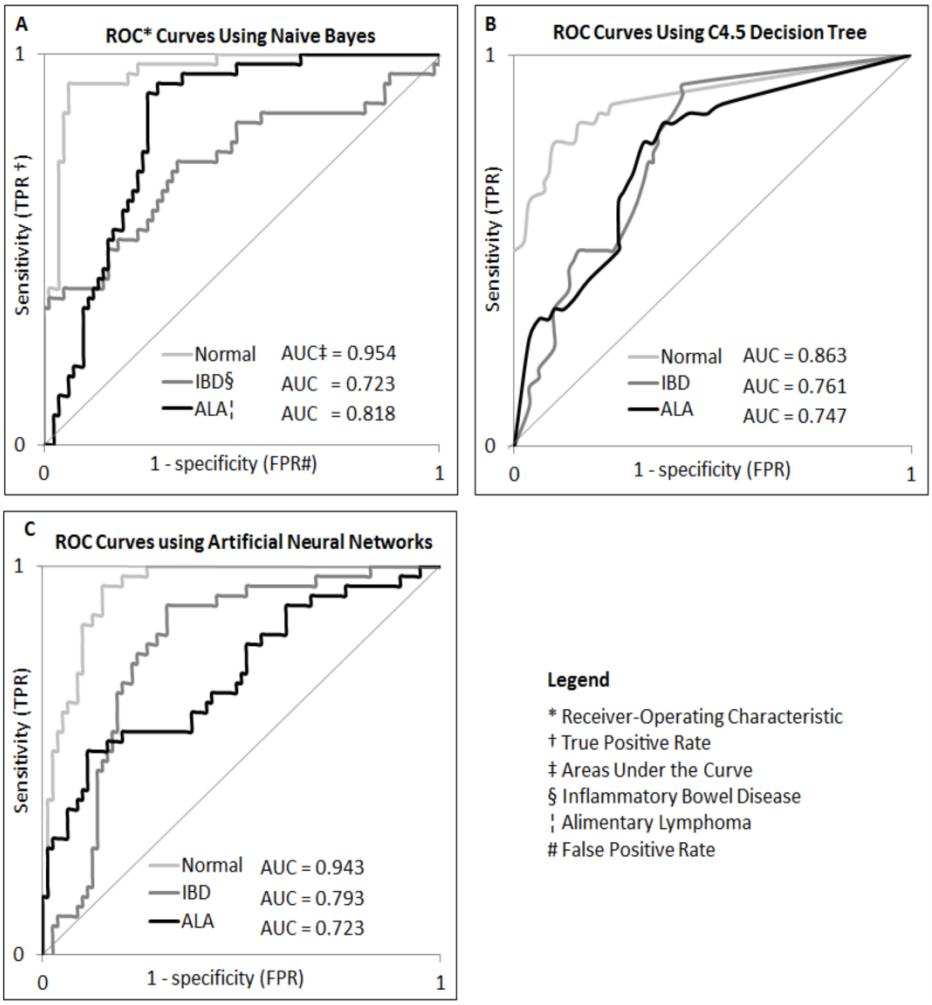

Figure 4.1 Receiver-operating characteristic graphs and areas under the curves from the three classification models: naïve Bayes (A), C4.5 decision tree (B), and artificial neural networks (C) to assess efficacy in classifying normal, inflammatory bowel disease (IBD), and alimentary lymphoma (ALA) cases.

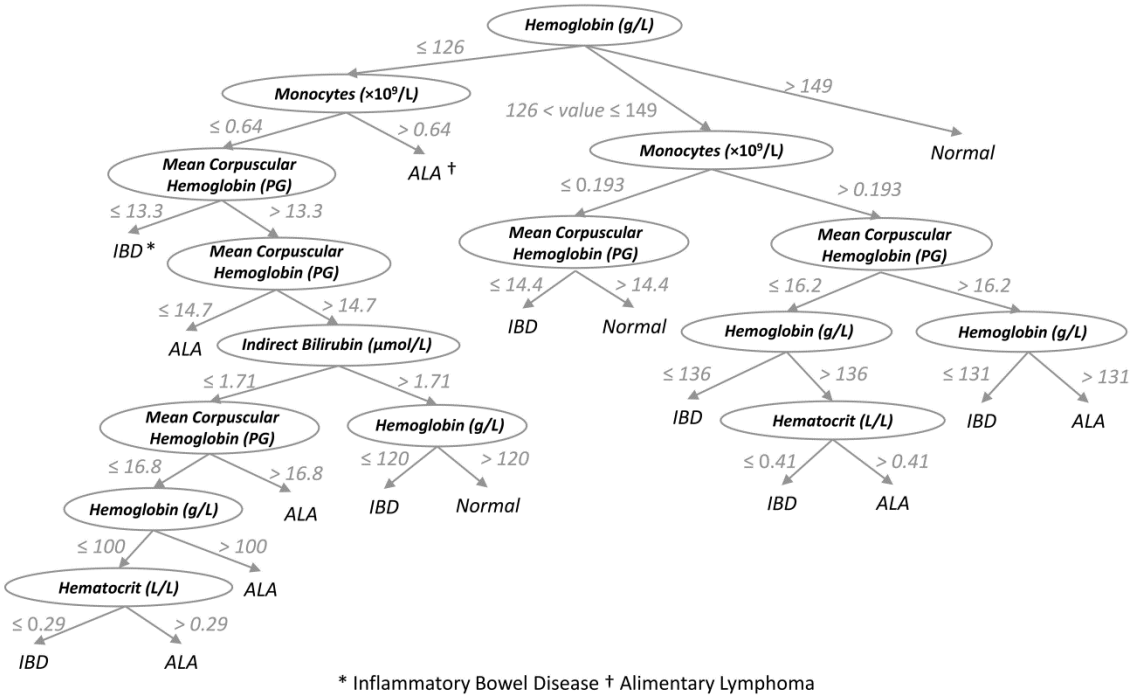

Figure 4.2 A flowchart showing the decision tree generated by the C4.5 classifier. Each node (variable shown in oval shape) represents an attribute and criteria used to classify cases. Numbers indicate values used for branching between nodes. Leaf terms (the ends of branches) represent case predicted class.

The average sensitivity values for every model using 10 repeats of 10-fold cross validation were the following: 70.83 % (68.91 – 72.75% CIs) for the naïve Bayes, 62.67 % (59.59 – 65.75% CIs) for the C4.5 decision tree, and 69.17 % (67.04 – 71.30% CIs) for the artificial neural networks, (**Table 4.6**). Two-tailed t-tests showed that performance sensitivity achieved by naïve Bayes was not different than the sensitivity achieved using the neural network classifier (P = .2586). However, both classifiers achieved a sensitivity that was higher than the one achieved using the decision tree classifier (P < .0001).

Table 4.6 Summary for 10 random repeats of 10-fold cross validation. The performance of naïve Bayes, C4.5 decision tree, and artificial neural networks classifiers.

	Sensitivity	95 % Confidence Intervals
Naïve Bayes	70.83 %*	(68.91 – 72.75) %
C4.5 Decision Tree	62.67 %	(59.59 – 65.75) %
Artificial Neural Networks	69.17 %*	(67.04 – 71.30) %

* The sensitivity rates achieved by naïve Bayes and artificial neural networks were significantly higher than the one achieved using C4.5 decision tree (two-tailed t-test).

4.5 Discussion

Basic statistical analyses in previous studies [4, 6, 34, 35] have shown laboratory differences between normal cats and cats with IBD and ALA. Cats with IBD or ALA tended to have mild non-regenerative anemia attributed to chronic disease and suppression of hematopoiesis as previously found. [34, 35] Consistent with these previous reports, CBC results in this study showed that cats with IBD and ALA had significantly lower red blood cell numbers (7.76×10^6/uL and 7.53×10^6/uL) and cats with ALA had higher white blood cell counts (14.32×10^3/uL) compared to the normal group (9.32×10^6/uL red blood cells, 9.44×10^3/uL white blood cells). Some of the SC abnormalities revealed in this study are also consistent with previous studies,[36] in which plasma total protein and albumin concentrations were significantly lower in IBD and ALA cats compared to the normal group (**Table 4.3** and **4.4**). Unfortunately, few differences were identified as possible discriminators using traditional comparative statistics between the ALA and IBD except for chloride and cholesterol (**Table 4.4**).

However, the naïve Bayes, C4.5 decision tree, and artificial neural networks proved more successful in differentiating between IBD and ALA cases as well as normal animals. Of the three algorithms examined, the naïve Bayes and artificial neural network's overall sensitivities were significantly higher than achieved by the decision tree algorithm (two-tailed values $P<0.0001$ and $P=.0008$, respectively), as shown in **Table 4.6**.

In classifying ALA cases alone, the mean sensitivity achieved by the naïve Bayes was higher ($P < .0001$) than achieved by the decision tree classifier, and both were higher ($P < .0001$) than achieved by the neural networks. For IBD, the mean sensitivity achieved using the neural networks classifier was higher ($P < .0001$) than achieved by the naïve Bayes and the decision tree classifiers.

Naïve Bayes was the top performer for identifying normal cats (but not different from neural networks) and this group had the highest sensitivity in all classifications. Sensitivities achieved by the naïve Bayes and the neural networks were significantly higher ($P < .0001$) than achieved by decision tree.

These findings suggest best practice is to use the naïve Bayes or artificial neural networks classifiers for class prediction. It might be essential to consider the number of attributes used in making the classification; the naïve Bayes achieved its accuracy using 10 CBC and SC variables,

and the artificial neural networks classifier achieved a not significantly different accuracy using only 4 variables as shown in **Table 4.4**.

Several studies have investigated the use of a variety of biomarkers to differentiate feline ALA from IBD and support the diagnosis. In immunophenotyping, one study [37] showed that in 28 T-cell type ALA cases, 82% had >75% of neoplastic lymphocytes labeled for expression of CD3 and 18% of cases had 51-75% of cells expressing CD3. In the same study 50% of cases with lymphoplasmacytic enteritis had the same number of $CD3^+$ or $CD79a^+$ lymphocytes. In another study, [38] immunolabeling for the critical lymphocyte survival factor, Bcl-2, was performed on small intestinal biopsy sections and determined that the expression in ALA was significantly higher than it was on cats with IBD. Another approach to distinguish between IBD and ALA is by clonality testing using polymerase chain reaction (PCR). In one study, [39] 22 out of 28 cats with ALA had clonal rearrangement of the T-cell receptor γ chain gene (TCRG), and polyclonal population of cells was identified in all 9 cats with IBD. Using diseases indicators from feces, [40] it was previously shown that fecal α1-PI concentration is higher in cats with severe IBD or ALA. The same study showed that low serum albumin and total protein concentrations may be common findings in cats with IBD or ALA. Serum thymidine kinase activity [41] has been reported to be a useful tumor marker in humans and animals. The level of thymidine kinase in the serum has shown also to be a prognostic indicator of ALA. However, the area under the ROC curve which has been advocated as an evaluation criterion for comparing different diagnostic tools showed that using this predictor for the diagnosis of ALA is weak (AUC = 0.66), leading to the inability of identifying the cats with ALA. The AUC in this study for the diagnosis of ALA using the naïve Bayes model is considered to be good (**Figure 4.2**). Moreover, the variables used in prediction don't require any extra tests to be performed; we are utilizing routine laboratory values (CBC and SC variables) to make the prediction.

In spite of its "black box" limitation, the accuracy achieved using the naïve Bayes algorithm alone (80.6%) was good in this study. It was the overall best classifier examined. Other medical decision support studies have found similar results using naïve Bayes classifiers. An intelligent heart disease prediction system was developed using the three machine learning techniques, decision tree, naïve Bayes, and neural networks. These techniques were applied on medical profiles data from attributes such as age, sex, blood sugar to predict the likelihood of patients getting a heart disease. The revealed prediction system had a high performance using the

three techniques; however, the naïve Bayes model was shown [42] to outperform the other two with the highest number of correctly classified instances. The naïve Bayes classifier was considered [43] the most useful machine learning classifier that can support physicians' decisions; it achieved the highest accuracy in classifying 20 cases and 10 cases datasets into 30 diagnostic classes of sport injuries using only 118 training instances.

A decision tree for predicting classification of groups can be developed using the C4.5 classifier. The decision tree's advantage over the other algorithms used in this study is that it creates a visual graph depicting selected variables and threshold values used for decision making. This is in direct contrast to the neural network and naïve Bayes algorithms which have been referred to as a "black box," providing no insight as to how classification was arrived at. The decision tree consists of human-readable rules of classification that makes it easy to be applied by clinicians and pathologists at the point of practice. However, the tree can still be hard to understand from a pathophysiologic standpoint, with some of the decision variables making little intuited sense.

To conclude, clinically it is difficult to distinguish between IBD and ALA in cats. We examined the use of three decision support algorithms to retrospectively differentiate between normal, IBD and ALA cats. Compared to statistical-based descriptive analyses, prediction models using machine learning provided a method for distinguishing between the 2 disorders, and they allow for individualized prediction of disease. The prediction models reported in this study used non-invasive laboratory data, a decided advantage over traditional endoscopic diagnostics. The naïve Bayes and artificial neural networks classifiers used 10 and 4 CBC/SC variables, respectively, to outperform the C4.5 decision tree which used 5 CBC/SC variables. The results show that both the naive Bayes and neural network models are good algorithm choices for constructing prediction models in this medical domain.

4.6 Sources and manufactures

[a] Advia 2120 Hematology Analyzer; Siemens Healthcare Diagnostics, Deerfield, IL, USA.

[b] Beckman Coulter AU480 Analyzer; Beckman Coulter, Krefeld, Germany.

[c] JMP Pro 11 Program, SAS Inc., Cary, NC, USA.

[d] Waikato Environment for Knowledge Analysis, version 3.7, University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/ml/weka.

4.7 Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

4.8 References

1. Leib, M.S., et al., *Endoscopic aspiration of intestinal contents in dogs and cats: 394 cases.* J Vet Intern Med, 1999. **13**(3): p. 191-3.

2. Jergens, A.E., et al., *Idiopathic inflammatory bowel disease in dogs and cats: 84 cases (1987-1990).* J Am Vet Med Assoc, 1992. **201**(10): p. 1603-8.

3. Willard, M.D., *Feline inflammatory bowel disease: a review.* J Feline Med Surg, 1999. **1**(3): p. 155-64.

4. Russell, K.J., et al., *Feline low-grade alimentary lymphoma: how common is it?* J Feline Med Surg, 2012. **14**(12): p. 910-912.

5. Rassnick, K.M., et al., *Efficacy of combination chemotherapy for treatment of gastrointestinal lymphoma in dogs.* J Vet Intern Med, 2009. **23**(2): p. 317-22.

6. Carreras, J.K., et al., *Feline epitheliotropic intestinal malignant lymphoma: 10 cases (1997-2000).* J Vet Intern Med, 2003. **17**(3): p. 326-331.

7. Ragaini, L., et al., *Inflammatory bowel disease mimicking alimentary lymphosarcoma in a cat.* Vet Res Commun, 2003. **27**: p. 791-793.

8. Wasmer, M.L., et al., *Food intolerance mimicking alimentary lymphosarcoma.* J Am Anim Hosp Assoc, 1995. **31**(6): p. 463-6.

9. Wilcock, B., *Endoscopic biopsy interpretation in canine or feline enterocolitis.* Semin Vet Med Surg, 1992. **7**(2): p. 162-171.

10. Washabau, R.J., et al., *Endoscopic, biopsy, and histopathologic guidelines for the evaluation of gastrointestinal inflammation in companion animals.* J Vet Intern Med, 2010. **24**(1): p. 10-26.

11. Golden, D.L., *Gastrointestinal endoscopic biopsy techniques.* Semin Vet Med Surg (Small Anim), 1993. **8**(4): p. 239-44.

12. Evans, S.E., et al., *Comparison of endoscopic and full-thickness biopsy specimens for diagnosis of inflammatory bowel disease and alimentary tract lymphoma in cats.* J Am Vet Med Assoc, 2006. **229**(9): p. 1447-1450.

13. Roth, L., et al., *Comparisons between Endoscopic and Histologic Evaluation of the Gastrointestinal-Tract in Dogs and Cats - 75 Cases (1984-1987).* J Am Vet Med Assoc, 1990. **196**(4): p. 635-638.

14. Day, M.J., et al., *Histopathological standards for the diagnosis of gastrointestinal inflammation in endoscopic biopsy samples from the dog and cat: a report from the World Small Animal Veterinary Association Gastrointestinal Standardization Group.* J Comp Pathol, 2008. **138 Suppl 1**: p. S1-43.

15. Willard, M.D., et al., *Interobserver variation among histopathologic evaluations of intestinal tissues from dogs and cats.* J Am Vet Med Assoc, 2002. **220**(8): p. 1177-1182.

16. Marsilio, S., et al., *Immunohistochemical and Morphometric Analysis of Intestinal Full-thickness Biopsy Samples from Cats with Lymphoplasmacytic Inflammatory Bowel Disease.* J Comp Pathol, 2014. **150**(4): p. 416-423.

17. Gabor, L.J., P.J. Canfield, and R. Malik, *Haematological and biochemical findings in cats in Australia with lymphosarcoma.* Aust Vet J, 2000. **78**(7): p. 456-461.

18. Smith, A.L., et al., *Perioperative complications after full-thickness gastrointestinal surgery in cats with alimentary lymphoma.* Vet Surg, 2011. **40**(7): p. 849-52.

19. Schnabel, L.V., et al., *Primary alimentary lymphoma with metastasis to the liver causing encephalopathy in a horse.* J Vet Intern Med, 2006. **20**(1): p. 204-6.

20. Taylor, S.S., et al., *Serum protein electrophoresis in 155 cats.* J Feline Med Surg, 2010. **12**(8): p. 643-53.

21. Hall, M., I. Witten, and E. Frank, *Data mining: Practical machine learning tools and techniques.* Kaufmann, Burlington, 2011.

22. McDonald, J.M., S. Brossette, and S.A. Moser, *Pathology information systems: data mining leads to knowledge discovery.* Arch Pathol Lab Med, 1998. **122**(5): p. 409-11.

23. Zhang, X.W., et al., *Ontology driven decision support for the diagnosis of mild cognitive impairment.* Comput Meth Prog Bio, 2014. **113**(3): p. 781-791.

24. Xhemali, D., Hinde, C.J., Stone, R.G.,, *Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages.* Comp. Sci, 2009. **4 (1)**: p. 16–23.

25. Quinlan, J.R., *Book Review: C4.5: by J. Ross Quinlan. Inc., 1993*. The Morgan Kaufmann series in machine learning. 1993, San Mateo, Calif.: Morgan Kaufmann Publishers. x, 302 p.

26. Abbass, H.A., *An evolutionary artificial neural networks approach for breast cancer diagnosis.* Artif Intell Med, 2002. **25**(3): p. 265-281.

27. Gabriels, W., et al., *Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks.* Aquat Ecol, 2007. **41**(3): p. 427-441.

28. Gevrey, M., L. Dimopoulos, and S. Lek, *Review and comparison of methods to study the contribution of variables in artificial neural network models.* Ecol Modell, 2003. **160**(3): p. 249-264.

29. Al-Omari, F.A., et al., *An intelligent decision support system for quantitative assessment of gastric atrophy.* J Clin Pathol, 2011. **64**(4): p. 330-337.

30. Abbod, M.F., et al., *Application of artificial intelligence to the management of urological cancer.* J Urol, 2007. **178**(4 Pt 1): p. 1150-6.

31. Jaffe, E.S., *The 2008 WHO classification of lymphomas: implications for clinical practice and translational research.* Hematology Am Soc Hematol Educ Program, 2009: p. 523-31.

32. Blum, A.L. and P. Langley, *Selection of relevant features and examples in machine learning.* Artif Intell, 1997. **97**(1-2): p. 245-271.

33. Kohavi, R., *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.* IJCAI (U S), 1995.

34. Lalor, S., et al., *Cats with inflammatory bowel disease and intestinal small cell lymphoma have low serum concentrations of 25-hydroxyvitamin D.* J Vet Intern Med, 2014. **28**(2): p. 351-5.

35. Jergens, A.E., et al., *A clinical index for disease activity in cats with chronic enteropathy.* J Vet Intern Med, 2010. **24**(5): p. 1027-1033.

36. Dennis, J.S., J.M. Kruger, and T.P. Mullaney, *Lymphocytic plasmacytic gastroenteritis in cats - 14 cases (1985-1990).* J Am Vet Med Assoc, 1992. **200**(11): p. 1712-1718.

37. Briscoe, K.A., et al., *Histopathological and immunohistochemical evaluation of 53 cases of feline lymphoplasmacytic enteritis and low-grade alimentary lymphoma.* J Comp Pathol, 2011. **145**(2-3): p. 187-98.

38. Swanson, C.M., et al., *Expression of the Bcl-2 apoptotic marker in cats diagnosed with inflammatory bowel disease and gastrointestinal lymphoma.* J Feline Med Surg, 2012. **14**(10): p. 741-5.

39. Moore, P.F., et al., *Characterization of feline T cell receptor gamma (TCRG) variable region genes for the molecular diagnosis of feline intestinal T cell lymphoma.* Vet Immunol Immunopathol, 2005. **106**(3-4): p. 167-78.

40. Burke, K.F., et al., *Evaluation of fecal alpha1-proteinase inhibitor concentrations in cats with idiopathic inflammatory bowel disease and cats with gastrointestinal neoplasia.* Vet J, 2013. **196**(2): p. 189-96.

41. Taylor, S.S., et al., *Serum thymidine kinase activity in clinically healthy and diseased cats: a potential biomarker for lymphoma.* J Feline Med Surg, 2013. **15**(2): p. 142-7.

42. Palaniappan, S. and R. Awang, *Intelligent heart disease prediction system using data mining techniques.* I C Comp Syst Applic, 2008: p. 108-115.

43. Zelic, I., et al., *Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries.* J Med Syst, 1997. **21**(6): p. 429-44.

Appendix A - Concept classes identified by our IM and not currently expressed in *SNOMED-CT*

Model Abnormal Morphology Concept	***SNOMED identifiers***	***SNOMED* Synonyms**
lymphocytic infiltrate	-	-
Small cell lymphocytic infiltrate	-	-
Large cell lymphocytic infiltrate	-	-
Model Finding Site Concept	***SNOMED identifiers***	***SNOMED* Synonyms**
Gastric villi	-	-
Gastric lacteals	-	-
Gastric lamina propria	-	-
Duodenal villi	-	-
Duodenal crypts	-	-
Duodenal lacteals	-	-
Jejunal villi	-	-
Jejunal crypts	-	-
Jejunal lacteals	-	-
Ileal villi	-	-
Ileal crypts	-	-
Ileal lacteals	-	-
Small intestinal villi	-	-
Small intestinal Lacteals	-	-
Colonic villi	-	-
Colonic lacteal	-	-
Rectal villi	-	-
Rectal epithelium	-	-
Rectal crypts	-	-
Rectal lacteals	-	-
Rectal lamina propria	-	-