# UTILIZATION OF DEEP LEARNING TO ACCURATELY DETERMINE CELL OF ORIGIN IN CANINE FOLLICULAR AND MEDULLARY THYROID CARCINOMAS ON ROUTINELY PROCESSED, H&E-STAINED TISSUE SECTIONS

BY

JILLIAN M. ATHEY

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in VMS-Veterinary Clinical Medicine
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Committee members:

Clinical Assistant Professor Miranda D. Vieson, Advisor
Dr. Keith Bailey, Charles River Laboratories
Dr. Dan Rudmann, Charles River Laboratories
Clinical Associate Professor Wes A. Baumgartner
Associate Professor Kim A. Selting

ABSTRACT

Canine thyroid carcinomas (CTCs) are a common endocrine malignancy that requires histopathologic examination with costly and time-consuming immunohistochemistries for definitive diagnosis, which can delay the time to treatment beyond surgical excision. A significant diagnostic challenge arises in differentiating compact follicular thyroid carcinomas (FTCs, derived from follicular cells) and medullary thyroid carcinomas (MTCs, derived from medullary cells) with routine hematoxylin and eosin (H&E) staining. Literature suggests these have similar clinical outcomes; however, publications often do not distinguish between compact FTCs and MTCs.

The primary objective of this project is to develop and validate an artificial intelligence (AI) deep learning algorithm that can accurately determine the cell of origin (follicular or medullary) in CTCs without the use of ancillary immunohistochemical (IHC) stains. The primary hypothesis is that the algorithm can accurately determine the cell of origin in CTCs on routine H&E-stained histopathology slides. A secondary objective includes reviewing and comparing demographic information between follicular-derived or medullary-derived CTCs and between several follicular subtypes and medullary carcinomas, while tertiary objectives include evaluating the ability of pathologists to correctly identify compact follicular thyroid carcinomas from medullary thyroid carcinomas on H&E alone and comparing their diagnoses to the interpretations of the algorithm's output.

This study confirmed the primary hypothesis that it is feasible to determine the cell of origin for CTCs by an AI model. Additional demographic information with comparisons between the different types of CTCs is provided, and the need for ancillary diagnostics in differentiating compact FTCs and MTCs is re-iterated. For this model, most of the convoluted neural nets are ready for use in conjunction with interpretation by a pathologist. Additional work

is needed on the convoluted neural net that is for differentiating between FTC subtypes and MTCs. The use of this AI model could expedite the workflow for the pathologist and allow for rapid definitive diagnosis between compact FTCs or MTCs in dogs on routine H&E-stained slides which would translate to a decreased financial burden for the client, decreased time to diagnosis for the patient, and ultimately decreased costs of reagents and supplies for the diagnostic lab in future cases. Additionally, a successful algorithm could be applied to prospective or, potentially past studies, with whole slide images (WSIs) of CTCs to establish consistent differentiation between FTCs and MTCs when IHCs are not available. This in turn allows for more reliable interpretations of study results (e.g., a response to treatment or the patient outcome) or for re-evaluation of results and conclusions derived from past studies where FTCs and MTCs were not distinguished. These applications could contribute to elucidating previously obscure differences in demographics, prognoses, effective treatment modalities, and factors contributing to tumorigenesis. Furthermore, rapid and inexpensive methods to determine the neoplastic cell of origin in CTCs will assist in paving the way for more swiftly customized and successful therapies (personalized healthcare, precision medicine), as is currently occurring in human medicine.

## ACKNOWLEDGEMENTS

Firstly, I would like to thank all of my advisors for all of their help and guidance throughout this process. Specifically, I want to thank Dr. Keith Bailey for planting the idea for this project; Dr. Dan Rudmann and Charles River Laboratories for allowing me to explore the fun world of artificial intelligence and pathology; and Dr. Miranda Vieson for providing abundant emotional support, academic guidance, and thoughtful feedback whenever I need it. Dr. Lindsey Smith (Aiforia) deserves special attention for all of her help in teaching me how to apply and train algorithms for histologic images. Finally, I am so grateful for the unwavering support of my husband, family, friends, and, of course, resident mates.

## TABLE OF CONTENTS

**List of Figures**

**List of Tables**

# CHAPTER 1: INTRODUCTION

## Introduction

Canine thyroid carcinomas (CTCs) are a common endocrine malignancy that can be divided into follicular or medullary carcinomas (FTCs or MTCs, respectively) based on their cell of origin (Campos et al., 2014a; Campos et al., 2014b; Carver et al., 1995; Hassan et al., 2020; Jegatheeson et al., 2021; Kiupel et al., 2008; Liptak, 2007; Pineyro et al., 2014; Ramos-Vara, 2002; Rosol and Meuten, 2017; Rosol and Frone, 2016; Soares et al., 2020). FTCs are derived from thyroid follicular cells (thyrocytes) while MTCs are derived from medullary cells (parafollicular cells or C-cells) (Carver et al., 1995; Patnaik and Lieberman, 1991; Pineyro et al., 2014; Rosol and Meuten, 2017; Rosol and Frone, 2016). Currently, differentiation between these entities requires an initial histopathologic examination, and a panel of immunohistochemical (IHC) stains which can include thyroglobulin, calcitonin, synaptophysin, and chromogranin A, among others (Campos et al., 2014b; Carver et al., 1995; Hassan et al., 2020; Soares et al., 2020; Moore et al., 1984). Innately, production of hematoxylin and eosin (H&E) stained slides costs time (few to several days) and money, while the additional need for IHC staining further increases the financial burden and time to a definitive diagnosis. This could potentially delay patient receipt of ancillary treatments (e.g., chemotherapeutics, radioactive iodine administration, and/or radiation therapy) beyond the initial treatment protocol, which is usually surgery (Castillo et al., 2016; Campos et al., 2014b; Campos et al., 2014c; Carver et al., 1995; Jegatheeson et al., 2021; Lee et al., 2020; Liptak, 2007; Moore et al., 1984; Sheppard-Olivares et al., 2020).

Well-differentiated FTCs may be subtyped based on their histologic patterns of growth into follicular, compact, mixed, or papillary (Campos et al., 2014b; Carver et al., 1995; Hassan et al., 2020; Kiupel et al., 2008; Liptak, 2007; Pineyro et al., 2014; Rosol and Meuten, 2017; Rosol

and Frone, 2016). A significant diagnostic challenge arises in differentiating compact FTCs (and MTCs with routine hematoxylin and eosin (H&E) staining, as both neoplasms may grow in solid forms with minimal additional differentiating features (Carver et al., 1995; Patnaik and Lieberman, 1991; Pineyro et al., 2014; Rosol and Meuten, 2017; Rosol and Frone, 2016). At this time, a clear prognostic difference between canine FTCs and MTCs is not documented, which contrasts with what is seen in human FTCs or MTCs (Carver et al., 1995). Treatment in humans is superficially similar to how CTCs are treated, but in human medicine, utilization of therapies that are specific to genetic derangements or molecule expression are increasingly being used (precision medicine) (Antonelli et al., 2018; Cabanillas et al., 2019; Ceolin et al., 2019; Gambardella et al., 2019; Meijer et al., 2013; Turrel et al., 2006; Valerio et al., 2017). It is well known that human FTCs and MTCs tend to have different genetic mutations, and it can therefore be extrapolated that novel treatments specific to these genetic or even biochemical derangements may be developed based on these differences (Campos et al., 2014c; Ceolin et al., 2019; Hassan et al., 2020; Valerio et al., 2017). There are, however, different metastatic tendencies between FTCs and MTCs, where FTCs tend to metastasize to the lungs via intravascular emboli while MTCs tend to metastasize to the anterior cervical lymph nodes; but, this is not exclusive (Hassan et al., 2020; Kiupel et al., 2008; Liptak, 2007; Rosol and Meuten, 2017; Rosol and Frone, 2016). Given the historical diagnostic challenge to differentiate compact FTCs from MTCs, and because IHC staining for calcitonin immunohistochemistry (IHC) was not fully established until somewhat recently, MTCs were likely underreported which might skew older demographic, prognostic, and biologic behavior data (Barber, 2007). Even current (2020 and 2021) clinical literature regarding treatment modalities, still does not often distinguish between FTCs and MTCs, which likely perpetuates the underestimation of MTC prevalence as well as possibly

masks any potential underlying differences in response to treatment or biological behavior (Barber, 2007).

Artificial intelligence (AI) in the context of pathology brings an exciting frontier of applications that are only in the nascent stages of being utilized in veterinary diagnostic pathology. In human medicine diagnostic settings, AI applications, and more specifically deep learning (DL), have been more extensively studied, and data show that AI models can more consistently and quickly perform menial or repetitive tasks with equivalent or increased accuracy as compared to a group of expert panelists (Ching et al., 2018; Coudray et al., 2018; Echle et al., 2021; Laury et al., 2021; Levine et al., 2019; Moxley-Wyles et al., 2020; Polonia et al., 2021; Turner et al., 2020). Utilization of AI or other analytic software to estimate the likelihood that a feature represents a specific disease process (e.g., benign versus malignant) in diagnosing medical conditions is known as computer-aided diagnosis (CAD or CADX) (Castellino, 2005; Chan et al., 2020; Echle et al., 2021; Laury et al., 2021; Levine et al., 2019; Tosun et al., 2020; Zuraw, 2020). This is currently in use in several avenues of human diagnostics to improve the accuracy and efficiency of various diagnostic and treatment processes (Bulten et al., 2021; Chan et al., 2020; Polonia et al., 2021; Sultan et al., 2020; Zuraw, 2020). More advanced uses of AI contribute to knowledge discovery and can provide information that a human observer is unable to provide without ancillary testing (e.g., genetic mutations or identification/prediction of biomarkers) which may assist in rapidly providing clinically actionable data (Castellino, 2005; Chan et al., 2020; Ching et al., 2018; Echle et al., 2021; Laury et al., 2021; Levine et al., 2019; Sultan et al., 2020; Tosun et al., 2020; Zuraw, 2020). Therefore, a successful deep learning algorithm to differentiate compact FTCs and MTCs on H&E-stained whole slide images (WSI)

will facilitate rapid, cost-effective, and highly consistent results and will further advance the acceptance and use of AI models in the veterinary diagnostic setting.

## CHAPTER 2: LITERATURE REVIEW

Canine thyroid carcinomas (CTCs) are the most common endocrine malignancy in dogs, tend to grow rapidly, often invade local tissues and vasculature, and approximately one-third of dogs with CTCs have metastasis at diagnosis (Campos et al., 2014a; Campos et al., 2014b; Carver et al., 1995; Hassan et al., 2020; Kiupel et al., 2008; Liptak, 2007; Nadeau and Kitchell, 2011; Pineyro et al., 2014; Ramos-Vara, 2002; Rosol and Meuten, 2017; Rosol and Frone, 2016; Soares et al., 2020). CTCs arise from one of the two main epithelial constituents of the thyroid gland– follicular cells (thyrocytes) which produce colloid, thyroid hormones (serum $T_3$ and $T_4$), and thyroglobulin (Tg) or medullary cells (C-cells, parafollicular cells) which secrete calcitonin (Lee et al., 2020; Liptak, 2007; Patnaik et al., 1978; Pessina et al., 2016; Pineyro et al., 2014; Ramos-Vara, 2002; Rosol and Meuten, 2017; Rosol and Frone, 2016). Malignant neoplastic transformation of these cell types yields follicular thyroid carcinomas (FTC), which are then classified based on histologic patterns, and medullary thyroid carcinomas (MTC), respectively (Carver et al., 1995; Hassan et al., 2020; Kiupel et al., 2008; Liptak, 2007; Pineyro et al., 2014; Rosol and Frone, 2016). CTCs of follicular origin are considered much more common; historically, MTCs accounted for less than 5% of all canine thyroid neoplasms but recent literature suggests MTCs may be more prevalent with incidence ranges between 16% and 36% (Barber, 2007; Campos et al., 2014a; Campos et al., 2014b; Campos et al., 2014c; Carver et al., 1995; Hassan et al., 2020; Kiupel et al., 2008; Leav et al., 1976; Liptak, 2007; Patnaik and Lieberman, 1991; Pessina et al., 2014; Pineyro et al., 2014; Ramos-Vara, 2002; Rosol and Meuten, 2017; Rosol and Frone, 2016; Soares et al., 2020). At this time the basic treatment modality is surgical excision which may be accompanied by a multitude of other modalities (Castillo et al., 2016; Campos et al., 2014b; Campos et al., 2014c; Carver et al., 1995;

Jegatheeson et al., 2021; Lee et al., 2020; Liptak, 2007; Moore et al., 1984; Sheppard-Olivares et al., 2020). Histologic examination is required for the confirmation of malignancy (namely local invasion or intravascular tumor emboli), while the use of immunohistochemistry is required for determining the cell of origin (FTC versus MTC) (Campos et al., 2014b; Carver et al., 1995; Ramos-Vara et al., 2016; Soares et al., 2020). Beyond the individual patient, routinely and consistently confirming the cell of origin and subsequent appropriate histologic subtyping provides two main benefits. Firstly, this allows for more accurate comparisons between types of CTCs, as would be needed for prospective analyses on survival and prognosis or experimental treatment modalities. Secondly, the current trend in human medicine is for highly customized, molecular-based treatment regimens (known as personalized healthcare or precision medicine) which are based on specific diagnoses, like the presence of specific genetic mutations or overexpression of certain molecules which contribute to tumorigenesis (Al Rasheed and Xu, 2019; Bai et al., 2020; Cabanillas et al., 2019; Nitulescu et al., 2015; Valerio et al., 2017; Wen et al., 2021). Since veterinary medicine tends to lag behind human medicine, the investigation into more specific diagnoses, including genetic or biochemical derangements, is critical to further advance our understanding of tumorigenesis of CTCs and the subsequent development of effective, customized treatments.

## Typical Signalment Features, Presentation, and Etiopathogenesis

It is generally accepted that dogs presenting with any form of thyroid carcinoma (FTC or MTC) fall within a 9- to 10-year mean and/or median age with increased risk associated with advancing age and no sex predispositions (Barber, 2007; Campos et al., 2014a; Hassan et al., 2020; Hayes and Fraumeni, 1975; Liptak, 2007; Rosol and Meuten, 2017; Rosol and Frone,

2016; Soares et al., 2020). Generally, regarding age, FTCs and MTCs are not separated in literature reports, but a few studies show that MTCs may have a mean age of 9.6 years or median age of 9 years with a range of either 4 to 12, 13, or 16 years (Campos et al., 2014b; Carver et al., 1995; Patnaik and Lieberman, 1991). Campos et al. (2014b) report that differentiated FTCs have a median age of 10 years with a range of 4 to 14 years. One older study of 16 dogs with MTCs found males were three times more likely to have MTCs than females (Patnaik and Lieberman, 1991).

While reports are conflicting and variable, predisposed breeds may include boxers, beagles, Siberian huskies, golden retrievers as well as mixed-breed dogs (Hassan et al., 2020; Hayes and Fraumeni, 1975; Liptak, 2007; Rosol and Meuten, 2017; Rosol and Frone, 2016). One study also suggests that Shetland collies (also known as sheltie or Shetland sheepdog), old English sheepdogs, and Cairn terriers have increased risk (Hassan et al., 2020). Hayes and Fraumeni (1975) found that poodles (miniature and toy) may have decreased risk of developing CTC.

CTCs as a whole are most commonly unilateral tumors that arise near the larynx without a reported side predisposition (i.e, right versus left thyroid gland) and are less commonly bilateral neoplasms (Hassan et al., 2020; Leav et al., 1976; Liptak, 2007; Patnaik and Lieberman, 1991; Pessina et al., 2014; Rosol and Meuten, 2017; Rosol and Frone, 2016). Thyroid neoplasms may form anywhere along the cranial neck (including the tongue) to the thoracic inlet or even in the anterior mediastinum, pericardium, heart base, and descending aorta due to ectopic thyroid parenchyma that was presumably entrapped during early development (Liptak, 2007; Rosol and Meuten, 2017).

The most common presentation for CTCs as a whole is a palpable ventral or ventrolateral cervical mass that may be associated with a variety of additional presenting clinical signs or symptoms such as coughing, dyspnea, dysphagia, dysphonia, gagging, retching, regurgitation, Horner's syndrome, and cranial vena caval syndrome due to either tumor compression (mass effect) or invasion (Lee et al., 2020; Liptak, 2007; Rosol and Meuten, 2017). These clinical signs may facilitate either the death of the animal or election of humane euthanasia by the owners.

Clinically distinguishing between FTCs and MTCs can be challenging, but functional neoplasms may be identified by evaluating serum hormone levels. Dogs with FTCs tend to be euthyroid, but hypothyroidism can occur by several mechanisms that may not be specific to the neoplastic process. These include complete bilateral destruction of the thyroid glands, suppression of pituitary derived thyroid-stimulating hormone (TSH; thyrotropin), or the suppressive effects of nonspecific illnesses on circulating thyroid hormone concentrations (euthyroid sick syndrome); mild hyperthyroidism may rarely occur with functional FTCs that secrete sufficient $T_3$ and/or $T_4$ thyroid hormones (Carver et al., 1995; Hassan et al., 2020; Liptak, 2007; Pessina et al., 2014; Rosol and Meuten, 2017; Rosol and Frone, 2016). Dogs with functional FTCs and concurrent hyperthyroidism are clinically similar but less severely affected than cats with clinical hyperthyroidism (Liptak, 2007). Dogs with CTCs tend to experience hyperthyroidism more than humans (Leav et al., 1976). Humans and dogs may have hypercalcitoninemia, which can be used as a diagnostic marker for human MTCs; however, this is not commonly utilized in veterinary diagnostics (Ceolin et al., 2019; Pineyro et al., 2014). Interestingly, there are apparently no reports of nonfunctional MTCs (non-calcitonin secreting MTCs) in the veterinary literature, while in humans this is rarely reported based on serum calcitonin levels and pentagastrin stimulation testing; in these human cases, approximately half

were IHC-positive for calcitonin and carcinoembryonic antigen (CEA) IHCs and most were IHC-positive for chromogranin A (CgA) IHC (Gambardella et al., 2019; Pineyro et al., 2014). MTCs may be accompanied by low to normal levels of serum calcium (Leav et al., 1976; Patnaik et al., 1978; Rosol and Meuten, 2017). Alternatively, MTCs may be associated with hypercalcemia (given the proposed pathogenesis, highlighted below), but this is less useful to distinguish from FTCs, as they may cause humoral hypercalcemia of malignancy (Rosol and Meuten, 2017; Rosol and Frone, 2016).

The etiopathogenesis behind both FTCs and MTCs is largely unknown. For FTCs, iodine deficiency causing chronic thyroid follicular hyperplasia, prolonged lymphocytic thyroiditis with hypothyroidism, and sufficient regional irradiation have all been implicated (Hayes and Fraumeni, 1975; Liptak, 2007). For the former two mechanisms, there is speculation that FTCs may retain TSH sensitivity, and since TSH is the main regulator of thyrocyte differentiation and proliferation, it may then act as a continued growth factor causing chronic overstimulation and subsequent neoplastic transformation (Pessina et al., 2014). Briefly, recent studies have attempted to explore the expression of markers that may contribute to CTC development, and markers include insulin-like growth factor (IFG)-1, vascular endothelial growth factor (VEGF), fibroblast growth factor (FGF)-2, and their receptors, as well as tumor cell interactions with the stroma in the tumor microenvironment (Campos et al., 2014a; Pessina et al., 2016). Campos et al. (2014c) found several genes in both FTCs and MTCs that contribute to increased expression of the phosphatidylinositol-3-kinase (PI3K/AKT) signaling pathway, which essentially promotes cell growth and survival. MTCs may be preceded by hypercalcemia (as may occur with primary hyperparathyroidism or hypercalcemia of malignancy) (Rosol and Meuten, 2017; Rosol and

Frone, 2016). In bulls, a relationship between chronic dietary intake of excessive calcium and MTCs is proposed, but there are no reports evaluating this in dogs (Rosol and Frone, 2016).

There is a paucity of reported familial thyroid cancer in dogs. Some Dutch German longhaired pointers have two recessive deletion mutations within the thyroid peroxidase (TPO; thyroperoxidase) gene, which contributes to familial FTC in this breed (Yu et al., 2021). TPO is important in the production of thyroid hormones (Rosol and Meuten, 2017; Rosol and Frone, 2016). Further investigation is needed for the presence of this derangement in CTCs of other breeds and species. TPO mutations in humans have been previously associated with thyroid carcinoma, while inactivating TPO mutations in both humans and dogs have been shown to cause autosomal recessive congenital goitrous primary hypothyroidism (Yu et al., 2021). Lee et al. (2006) reports the first case of familial MTCs in 3 related dogs, but *RET* mutations were not identified which contrasts what is seen in human hereditary MTCs and multiple endocrine neoplasm (MEN) syndromes (Cabanillas et al., 2019; Campos et al., 2014c; Ceolin et al., 2019; Fuchs et al., 2020; Gambardella et al., 2019; Hayes and Fraumeni, 1975; Martucciello et al., 2012; Meijer et al., 2013; Rosol and Meuten, 2017; Valerio et al., 2017; Yu et al., 2021).

## Histopathologic Classification of CTCs

Canine FTCs can be classified by histologic examination utilizing the World Health Organization (WHO) classification scheme into well-differentiated (with subtypes), poorly differentiated, and undifferentiated (with subtypes, like carcinosarcoma [malignant mixed thyroid tumor]), while MTCs are considered separately as a single entity (Campos et al., 2014b; Kiupel et al., 2008). Subtypes of well-differentiated FTCs are based on histologic patterns and include follicular, follicular-compact (also known as mixed), compact cellular (also known as

compact or solid), and papillary (Carver et al., 1995); Kiupel et al., 2008; Liptak, 2007; Pineyro et al., 2014; Rosol and Meuten, 2017; Rosol and Frone, 2016). Confusingly, some veterinary references use "mixed" to describe carcinosarcoma (malignant mixed thyroid tumors) where malignant follicular cells are mixed with malignant mesenchymal components (e.g., cartilage, bone, or both), while in human thyroid carcinomas there also exists a mixed carcinoma consisting of both neoplastic medullary cells and neoplastic follicular cells (Bais et al., 2020; Cameselle-Teijeiro et al., 2020; Ramos-Vara et al., 2002). For this study, the follicular-compact subtype will hereby be referred to as the "mixed" subtype, and the compact-cellular will be referred to as the "compact" subtype. Overall, the follicular, mixed, and compact subtypes are considered relatively common diagnoses (Pessina et al., 2016). Of these, mixed FTCs may be the most common and are characterized by approximately equal proportions of follicular and compact neoplastic growth, with neoplastic follicles that may be smaller and contain less colloid than would be observed in the follicular pattern (Kiupel et al., 2008; Rosol and Meuten, 2017; Rosol and Frone, 2016). Few studies indicate the compact type may be more common than other subtypes (Campos et al., 2014c, Pessina et al., 2014).

A diagnostic challenge arises with differentiating compact FTCs from MTCs with routine hematoxylin and eosin (H&E) light microscopy and ancillary diagnostic modalities are required (**Figure 1**) (Carver et al., 1995; Kiupel et al., 2008; Leav et al., 1976; Patnaik and Lieberman, 1991; Pineyro et al., 2014; Rosol and Meuten, 2017; Rosol and Frone, 2016). Compact FTCs are described as aggregates to solid sheets of polyhedral cells with finely granulated to vacuolated eosinophilic cytoplasm, which gives little to no attempt at follicle formation and/or colloid secretion (Kiupel et al., 2008; Rosol and Meuten, 2017; Rosol and Frone, 2016). MTCs are similarly described as polyhedral to spindle-shaped cells with lightly eosinophilic to

amphophilic, finely granular cytoplasm, and an oval to elongate vesicular nucleus (Kiupel et al., 2008; Patnaik et al., 1978; Ramos-Vara et al., 2002; Rosol and Meuten, 2017; Rosol and Frone, 2016; Williams, 1966). Histologic features that support FTC include numerous variably sized follicles with periodic acid-Schiff (PAS)-positive colloid; eosinophilic cytoplasm; greater pleomorphism; and hemorrhage into follicles. In contrast, histologic features that support MTC include potentially more intrathyroidal and/or extracapsular invasion; occasional ducts and acini; production of amyloid derived from secreted proteins; solid growth with neuroendocrine (NE) packeting; amphophilic cytoplasm; prominent stroma; few to no follicles; and palisading cells along the periphery of lobules (Leav et al., 1976; Pineyro et al., 2014; Ramos-Vara et al., 2002; Rosol and Meuten, 2017). Interestingly, MTCs may also have variable histologic patterns of growth including tubular (follicular), papillary, small cell, giant cell, clear cell, oncocytic, and mixed variants, however, this distinction is not often pursued in the current diagnostic setting (Kiupel et al., 2008). Amyloid production is rarely observed in canine MTCs and is more commonly found in human or bull medullary carcinomas (Kiupel et al., 2008; Patnaik et al., 1978; Pineyro et al., 2014; Rosol and Meuten, 2017; Rosol and Frone, 2016). Differentiation between MTC and compact FTC is further confounded by the presence of either entrapped, "normal" (non-neoplastic) thyroid follicles surrounded by neoplastic cells or the formation of medullary cell follicles, which are neoplastic medullary cells surrounding free colloid following the destruction and loss of follicular thyrocytes (Rosol and Meuten, 2017; Rosol and Frone, 2016). Interestingly, MTCs tend to entrap thyroid follicles more than FTCs tend to entrap medullary cells (Rosol and Meuten, 2017).

**Differentiation of CTCs by IHC and Histologic and Immunohistochemical Differentials**

Because differentiation of MTCs from compact FTCs with routine H&E-stained slides is challenging, IHC examination is often pursued and required for further characterization (Campos et al., 2014b; Carver et al., 1995; Leav et al., 1976; Moore et al., 1984; Soares et al., 2020). FTCs are diagnosed with positive thyroglobulin (Tg) immunoreactivity in 10% to 100% of the cells and ideally concurrent negative immunohistochemistry for calcitonin and/or NE markers (such as chromogranin A+B, protein gene product 9.5 [PGP9.5], neuron-specific enolase, to name a few) (Carver et al., 1995; Kiupel et al., 2008; Moore et al., 1984; Pineyro et al., 2014; Ramos-Vara et al., 2002). In contrast, MTCs may be diagnosed with variable amounts of cytoplasmic immunoreactivity for calcitonin or NE markers (similar as above and synaptophysin [SYP]) with concurrent negative immunoreactivity for Tg (Carver et al., 1995; Kiupel et al., 2008; Moore et al., 1984; Pineyro et al., 2014; Ramos-Vara et al., 2002). The range in thyroglobulin immunoreactivity may relate to histologic subtypes, as follicular and mixed subtypes appear to more consistently exhibit thyroglobulin immunoreactivity as compared to the compact subtypes (Carver et al., 1995). Less commonly, IHCs for thyroid transcription factor 1 (TTF-1 [NKX2]) or PAX8 can be used to confirm FTCs while calcitonin gene-related peptide (CGRP), napsin A, and carcinoembryonic antigen (CEA) can be used to confirm MTCs; however, these not as readily available for routine diagnostic use (Hassan et al., 2020; Pineyro et al., 2014; Ramos-Vara et al., 2002; Ramos-Vara et al., 2016; Rosol and Frone, 2016). Ramos-Vara et al. (2002) found that TTF-1 cannot be used as a discriminatory IHC, since TTF-staining was detected in both FTCs and MTCs; they instead propose TTF may be used in conjunction with Tg IHC to improve the overall sensitivity of the IHC panel. One study also suggests CGRP may be a more sensitive marker than calcitonin for diagnosing MTCs (Solar Arias et al., 2016). An additional challenge

may arise with either scant or faint IHC staining of either Tg or calcitonin. For Tg, this may be from altered physicochemical hormones or inactivity of neoplastic cells, and for calcitonin, this may result from preferential secretion and concurrent lack of hormone storage with highly functional MTCs (Moore et al., 1984; Rosol and Meuten, 2017). Conversely, inadequate or delayed fixation of CTCs may allow diffusion of Tg into the interstitium, contributing to increased background staining and potentially confounding Tg IHC interpretation (Ramos-Vara et al., 2002). In MTCs, the presence of entrapped, remnant, Tg-positive parenchyma may also interfere with accurate IHC interpretation (Rosol and Meuten, 2017).

In the past, MTCs were routinely underdiagnosed due to the marked histologic similarity and the lack of the confirmatory calcitonin IHC marker (Campos et al., 2014b; Carver et al., 1995). Underdiagnosis of canine MTCs likely contributed to the lack of comparative prognostic studies with canine FTCs, and as previously mentioned, many current clinical studies fail to distinguish between these tumors. This lack of discrimination between FTCs and MTCs may perpetuate skewed data with lower MTC prevalence and may yield unreliable interpretations and conclusions from CTC studies due to the masking of differences between MTCs and FTC.

Interestingly, bulls may get ultimobranchial thyroid neoplasms which have a heterogenous histologic pattern of more typical medullary cells mixed with undifferentiated cells that have IHC positivity to both thyroglobulin and calcitonin (dual immunoreactivity); the ultimobranchial body is thought to be the origin of medullary cells to the thyroid gland in embryological development (Kiupel et al., 2008; Rosol and Frone, 2016). There is speculation that a unique version of mixed thyroid carcinoma in humans (mixed medullary and follicular thyroid carcinoma) could also represent this entity (Bai et al., 2020; Cameselle-Teijeiro et al., 2020; Kiupel et al., 2008; Rosol and Frone, 2016). Ultimobranchial tumors should therefore be

considered tumors of stem cells and may be a consideration for unusual, dual-positive canine tumors (Kiupel et al., 2008; Moore et al., 1984). Canine ultimobranchial tumors have not been previously reported, although incidental ultimobranchial cysts are relatively common (Rosol and Meuten, 2017; Rosol and Frone, 2016).

Recently, Soler Arias et al. (2016) report a calcitonin-negative (nonmedullary) primary neuroendocrine tumor of the thyroid in a dog with negative IHC staining for calcitonin, CEA, Tg, S100 protein, and positivity for synaptophysin and cytokeratin AE1-AE3, which is similar to the rare human tumor known as calcitonin-negative neuroendocrine tumor of the thyroid (CNNET) or "nonmedullary" thyroid tumor. In this case, IHCs to rule out a parathyroid tumor were not possible, but primary hyperparathyroidism had already been ruled out clinically and a parathyroid tumor was therefore considered unlikely (Soler Arias et al., 2016). Other differentials for human calcitonin-negative neuroendocrine tumors of the thyroid gland include paraganglioma, hyalinizing trabecular tumor, metastatic neuroendocrine tumor to the thyroid gland, and intrathyroidal parathyroid adenoma or tumor; some of these may also be viable differentials for dogs (Cameselle-Teijeiro et al., 2020; Soler Arias et al., 2016).

For parathyroid tumors, clinical primary hyperparathyroidism often aids in establishing an initial diagnosis with histopathology and IHC evaluation acting to confirm the diagnosis (Soler Arias et al., 2016). Neoplastic cells for both parathyroid adenomas and carcinomas will have IHC positivity to parathyroid hormone, cytokeratins, neuroendocrine markers (e.g., CgA), and neurofilaments (such as S100) (Kiupel et al., 2008; Rosol and Frone, 2016; Rosol and Meuten, 2017).

Therefore, with a lack of pertinent clinical history beyond the suspicion of a thyroid neoplasm, such as the presence or absence of clinical hyperthyroidism or hyperparathyroidism,

and without a robust IHC panel, definitive diagnosis may be challenging. A robust IHC-panel could include at least two FTC markers (Tg and TTF-1), 2 MTC markers (calcitonin and CGRP), one or more NE markers (SYP or CgA), and parathyroid hormone, for example.

## Differences in Biologic Behavior and Prognosis between FTCs and MTCs and Treatment Modalities

Metastatic disease is frequently present at the time of diagnosis of CTCs, but some recent studies suggest that the progression of metastatic disease is slow (Giannasi et al., 2020). Computed tomography appears to be a more sensitive modality in detecting distant metastases in CTCs as compared to older, different imaging techniques (Giannasi et al., 2021). The likelihood of metastasis increases with the increased size of the primary tumor, evidence of vascular invasion (e.g., tumor thrombi in the cranial thyroid vein), and bilateral disease (Campos et al., 2014b; Hassan et al., 2020; Jegatheeson et al., 2021; Liptak, 2007; Nadeau and Kitchell, 2011). A study on stereotactic body radiation therapy (SBRT) in the treatment of CTCs found the presence of metastasis was not a negative prognostic factor, while another study found that after thyroidectomy the overall survival, disease-free survival, time to metastasis, and time to recurrence were not different between well-differentiated FTCs and MTCs (Campos et al., 2014b; Lee et al., 2020). Reported rates of metastasis at diagnosis include a range from 14-26% for regional metastasis (regional lymph nodes), 20-38% for distant metastasis (pulmonary), and 18-95% for overall metastasis (combining regional and distant) or approximately one-third of dogs (Giannasi et al., 2021; Hassan et al., 2020; Jegatheeson et al., 2021; Nadeau and Kitchell, 2011). These sources largely do not distinguish between FTCs and MTCs. Regional lymph nodes

may include the submandibular, medial retropharyngeal, and parotid lymph nodes (Liptak, 2007).

To underscore the need for accurate differentiation of CTCs, FTCs and MTCs carry different metastatic tendencies which may have undiscovered treatment implications. These patterns are not exclusive. FTCs tend to metastasize to the lung via invasion of the cranial and caudal thyroid veins with intravascular neoplastic emboli, while MTCs tend to metastasize to the anterior cervical lymph nodes (Hassan et al., 2020; Kiupel et al., 2008; Liptak, 2007; Rosol and Meuten, 2017; Rosol and Frone, 2016). Distant metastases may uncommonly be found in the brain, bone, liver, kidney, adrenal gland, liver, heart, and other organs (Lee et al., 2020; Liptak, 2007). Previously, there was speculation that the metastatic rate for MTCs may be lower than FTCs, but one study found that at diagnosis, there was no difference in the incidence of metastasis (Campos et al., 2014b; Carver et al., 1995; Hassan et al., 2020; Rosol and Meuten, 2017). Scintigraphy may be useful in identifying regional lymph node metastasis but not pulmonary metastasis in FTCs which may be linked to the concept that histologic subtypes of FTCs are related to the amount of cellular differentiation (Castillo et al., 2016; Liptak, 2007; Pessina et al., 2014). Furthermore, functional FTCs may have a lower metastatic rate than previously reported CTCs (Scharf et al., 2020).

At this time, both MTCs and all FTC subtypes appear to have similar treatment protocols and prognoses (Campos et al., 2014b). However, Carver et al. (1995) suggests MTCs may have a less biologically aggressive nature than other thyroid carcinomas, given that most MTCs in their study were well circumscribed and resectable. Campos et al (2014b) suggests that while MTCs may be more amenable to complete surgical resection, the post-thyroidectomy outcome is similar to that of well-differentiated FTCs. Since it is suggested that compact FTCs may be less

differentiated than follicular or mixed FTCs, they are more likely to behave aggressively and recur (Castillo et al., 2016; Campos et al., 2014b).

Ki-67, a cellular marker of proliferation, has been evaluated in both canine FTCs and MTCs with no significant differences identified, and it appears to be positively associated with local invasiveness and negatively associated with time to metastasis in both tumor types (Campos et al., 2014b; Soares et al., 2020). In well-differentiated FTCs in humans, increased Ki-67 is associated with higher metastatic rates at diagnosis and decreased disease-free survival (Campos et al., 2014b).

Currently, the main treatment modality is surgery for curative intent for mobile and well-circumscribed CTCs (complete thyroidectomy), but incisional biopsy with or without additional ancillary therapies may be pursued for invasive and non-resectable neoplasms (Castillo et al., 2016; Campos et al., 2014b; Campos et al., 2014c; Carver et al., 1995; Jegatheeson et al., 2021; Lee et al., 2020; Liptak, 2007; Moore et al., 1984; Sheppard-Olivares et al., 2020). Ancillary therapies may include local external beam radiation therapy, chemotherapy (e.g., toceranib phosphate [Palladia], doxorubicin, carboplatin, cisplatin, and adjunctive retinoic acid 9 cis [isotretinoin 9-cis; RA9-cis]), and/or radioactive iodine administration ($^{131}$iodine) (Castillo et al., 2016; Campos et al., 2014b; Campos et al., 2014c; Carver et al., 1995; Jegatheeson et al., 2021; Lee et al., 2020; Liptak, 2007; Moore et al., 1984; Sheppard-Olivares et al., 2020). Only approximately 25-50% of cases are amenable to complete surgical resection due to local invasiveness and proximity of critical anatomic structures (Jegatheeson et al., 2021; Lee et al., 2020). For cases where surgical resection is not possible, radiation therapy and radioiodide therapy may provide median survival times of up to 24 and 30 months, respectively, while the

reported median survival time of dogs with incompletely excised CTCs without ancillary therapies is 10 months (Brearley et al., 1999; Scharf et al., 2020).

Uptake of radioactive iodine administration appears to be related to the degree of neoplastic cell differentiation. This is supported by a preliminary study which found follicular FTCs were associated with hyperthyroidism and increased uptake of scintigraphy agent $^{99m}$Tc, mixed FTCs were associated with euthyroidism and normal uptake of $^{99m}$Tc uptake, and compact FTCs, which may be considered less differentiated, were associated with hypothyroidism and decreased $^{99m}$Tc uptake (Castillo et al., 2016; Jegatheeson et al., 2021; Pessina et al., 2014). CTCs do not need to be functional for abnormal scintigraphy studies, and there is speculation that neoplasms may be sensitive to radioactive iodine regardless of functional status, but consensus, even in human medicine, is lacking (Erdogan et al., 2006; Liptak, 2007; Meijer et al., 2013). There is no reported information regarding iodine uptake and efficacy of radioactive iodine administration on MTCs in dogs. However, Jegatheeson et al. (2021) recently describes the response of CTCs to radioiodine, but study their set were primarily FTCs and they do not describe any methods to rule out MTCs. In human medicine, one study concluded that radioactive iodine treatment may be a valid, locally aimed adjuvant treatment modality in MTCs, even though these cells do not concentrate radioactive iodine, while another multicenter study did not corroborate this and strongly opposed the use of radioactive iodine in MTCs (Erdogan et al., 2006; Meijer et al., 2013). Mechanistically, it is thought that the organification of radioactive iodine isotopes in adjacent thyroid follicular cells could destroy the adjacent medullary cells ("bystander effect") (Erdogan et al., 2006; Meijer et al., 2013). Numerous publications on the role of radiation therapy in fixed CTCs suggest that the primary treatment focus should be on local control of the primary tumor rather than systemic treatments, based on their study results

and that progression of metastatic disease is slow (Giannasi et al., 2020; Jegatheeson et al., 2021; Nadeau and Kitchell, 2011).

Prognostication largely varies with gross and histologic characteristics, such as tumor size or volume, local invasiveness, evidence of distant metastases, or evidence of vascular invasion; however, some of these features may be controversial (Campos et al., 2014b; Carver et al., 1995; Hassan et al., 2020; Liptak, 2007; Soares et al., 2020). Campos et al (2014b) conclude that some of these features had no significant effect on overall survival, disease-free survival, time to distant metastasis, or time to loco-regional occurance but macroscopic and histologic vascular invasion were independent negative predictors for disease-free survival. Prognosis can range from good to excellent; for surgery alone, median survival times range from 7-8 months to over 36 months, while the median survival time for untreated dogs is 3 months (Campos et al., 2014b; Carver et al., 1995; Hassan et al., 2020; Liptak, 2007; Soares et al., 2020). The reported median survival time for incompletely excised CTCs without ancillary therapy is 10 months and is largely due to disease secondary to local neoplastic invasion (Brearley et al., 1999). The control provided by radioiodide appears similar to other non-surgical treatment modalities, alleviates clinical signs, and prolongs survival, but does not significantly reduce tumor size; and, in people with lymph node and distant metastasis, radioactive iodine independently predicts a longer survival (Jegatheeson et al., 2021).

Some of the aforementioned gross or clinical characteristics can be surmised from the clinician's impression before and during surgery; however, the surgeon's interpretation may be inaccurate, so histologic examination is critical for confirmation of the gross findings, cell of origin, and further evaluation of criteria of malignancy (e.g., capsular invasion, vascular invasion, mitotic rate) (Campos et al., 2014b; Carver et al., 1995; Soares et al., 2020). Negative

prognostic indicators include both gross and histologic vascular invasion, increased (delayed) time to presentation, and increased tumor size (before thyroidectomy), while some papers maintain that histologic evidence of capsular or vascular invasion does not correlate with poor survival (Campos et al., 2014b; Liptak, 2007). However, some of this data was from dogs that were examined for necropsy or had inoperable tumors, and therefore may not accurately reflect the situation in dogs with operable tumors (Campos et al., 2014b). Studies on operable thyroid tumors found that bilateral disease and histologic grade of malignancy may be helpful as prognostic indicators, although a histologic grading scheme has not been previously published nor accepted (Campos et al., 2014b). Another paper says the prognosis is good for completely excised mobile thyroid tumors and irradiated fixed thyroid carcinomas (Liptak, 2007). As mentioned above, the median survival time of dogs with untreated thyroid carcinoma is 3 months, while in patients with resectable tumors and no metastasis, the median survival time with surgery alone is 7-8 months to over 36 months (Castillo et al., 2016; Hassan et al., 2020). While the use of chemotherapy or surgery tends to improve the prognosis, there does not appear to be a difference in median survival time when chemotherapy and surgery are used in combination as compared to surgery alone (Castillo et al., 2016; Giannasi et al., 2021; Liptak, 2007; Nadeau and Kitchell, 2011).

Several papers have attempted to correlate World Health Organization TNM staging (tumor, lymph nodes, metastasis) system to prognosis in CTCs, but results are largely conflicting at this time (Giannasi et al., 2020; Jegatheeson et al., 2021; Turrel et al., 2006).

**Biochemical and Genetic Derangements and Applications**

Altered growth factors, such as insulin-like growth factor (IGF)-1, vascular endothelial growth factor (VEGF), and fibroblast growth factor (FGF-2) and their receptors are confirmed or thought to contribute to the progression of both human and canine thyroid cancers (Campos et al., 2014a; Pessina et al., 2016). In addition to staining of neoplastic cells, tumor-associated fibroblasts and/or endothelial cells of compact FTCs were found to have increased IHC expression of IGF-1, VEGF, FGF-2, and retinoic acid receptor (RAR)α which may contribute to tumor progression; in this study, mixed FTCs had intermediate expression levels which could further reflect the amount of differentiation present (Pessina et al., 2016). Another study has identified Tg and TSH-R in tumor-associated fibroblasts and that proliferating cell nuclear antigen (PCNA; a marker for cell mitogenesis) was expressed in neoplastic follicular cells and fibroblasts while TTF-1 was restricted only to neoplastic follicular cells (Pessina et al., 2014). The significance and implications of some of these findings remain unclear (Pessina et al., 2014).

In humans, IGF-1/IGF-1R is overexpressed in some types of thyroid carcinomas and is correlated with poor prognosis (Liu et al., 2013; Pessina et al., 2016). VEGF contributes to tumor-induced angiogenesis controlled by neoplastic, stromal, and tumor-infiltrating cells, and one study correlated increased VEGF expression with poorer outcomes in compact FTCs (Pessina et al., 2016). FGF-2 is implicated in abnormal human thyroid growth as a mitogen and stimulator for endothelial cell growth but reports on levels FGF-2 levels in human thyroid carcinomas are contradictory and may depend on the degree of differentiation (Pessina et al., 2016).

Specifically, both canine FTCs and MTCs tend to have overexpression of VEGF while MTCs tend to have overexpression of cyclooxygenase-2 (cox-2) and P-glycoprotein (P-gp)

(Campos et al., 2014a). These overexpressed molecules may be therapeutic targets in the future. For example, a preliminary study utilizing multitargeted tyrosine kinase inhibitor (TKI) found promising results of remission in four of fifteen dogs and stable disease in eight of fifteen dogs with thyroid carcinoma (Campos et al., 2014a). P-glycoprotein (*ABCB1* gene; P-gp) is an efflux pump involved in multi-drug resistance and has been found in human chemotherapy-resistant MTC cell lines; targeting and inhibiting this molecule may improve chemotherapeutic efficacy in otherwise resistant tumors (Campos et al., 2014a). Cox-2 expression may be linked to tumor development, and studies in human thyroid cancer have found that increased cox-2 expression has a direct causal relationship with P-gp; meaning, cox-inhibitors may decrease P-gp expression and therefore improve tumor sensitivity to chemotherapeutic agents (Campos et al., 2014a).

Another study has found significant upregulation of *VEGFR-1*, *VEGFR-2*, *PDPK1*, *AKT1*, and *AKT2* in canine FTCs and of *EGFR*, *VEGFR-1*, and *PIK3CA* in canine MTCs which are all genes involved in the PI3K/AKT signaling pathway and could be candidates for therapeutic targets (Campos et al., 2014c). This pathway regulates numerous cell processes, including cell proliferation, differentiation, and survival (Nitulescu et al., 2015). In humans, copy number gains in these receptor tyrosine kinase (RTK) genes are particularly associated with activation of the PI3K/AKT pathway and are involved in the pathogenesis of human FTC (Campos et al., 2014c). In this study, it was concluded that activation of this pathway could contribute to the pathogenesis of canine thyroid carcinomas via promoting cell proliferation, resistance to apoptosis (via derangement of cox-2 expression), and malignant transformation, but more research is still needed (Campos et al., 2014c). Interestingly, the most common mutations in human thyroid carcinomas (*RAS [N, K, and H]*, *PIK3CA*, *BRAF*, *RET*, and *PTEN*) are considered rare in dogs (Al Rasheed and Xu, 2019; Campos et al., 2014c; Valerio et al., 2017;

Varricchi et al., 2019). Human FTCs tend to have mutations in the *RAS*, *PTEN*, *PIK3CA,* and *BRAF* genes, while the main mutation associated with human MTC is aberrant activation of *RET* (and less commonly *RAS* mutations), which signals through the PI3K/AKT and MAPK pathways (Campos et al., 2014c; Ceolin et al., 2019; Hassan et al., 2020; Valerio et al., 2017). It should be noted that up to 30% of human MTCs are hereditary due to germline alterations in the *RET* proto-oncogene that gives rise to MEN type 2 syndrome A and B, while between 30-50% of sporadic MTCs have somatic activating *RET* mutations, but this was not found in the first reported case of familial canine MTCs (Cabanillas et al., 2019; Campos et al., 2014c; Ceolin et al., 2019; Fuchs et al., 2020; Gambardella et al., 2019; Hayes and Fraumeni, 1975; Lee et al., 2006; Martucciello et al., 2012; Meijer et al., 2013; Rosol and Meuten, 2017; Valerio et al., 2017; Yu et al., 2021). Hassan et al. (2020) corroborate patterns of expression of some of these genes (FTC expression of *AKT2* and *PIK3CA* and MTC expression of *RET*) and provides additional differentially expressed genes on an mRNA level between normal canine thyroid glands, FTCs, MTCs, and a canine thyroid adenocarcinoma cell line (CTAC). Encouraging results have been identified from human studies using PI3K/AKT signaling inhibitors as well as preliminary canine studies using toceranib phosphate (multitargeted TKI that targets VEGFR-2) (Campos et al., 2014c).

Human epidermal growth factor (HER)-2 immunohistochemical expression has also been evaluated in canine CTCs, with nearly 50% of cases having moderate to strong expression (Yoshimoto et al., 2004). In human thyroid carcinomas, increased HER-2 expression is correlated with worse prognostic indicators, but no correlation has been identified in CTCs (Yoshimoto et al., 2004). This study does not appear to distinguish between FTCs and MTCs.

These specific molecules, receptors, genes, and pathways are mentioned as they may represent therapeutic targets that may eventually play a role in precision medicine for canine patients. Furthermore, both human and canine FTCs and MTCs appear to have genetic differences that appear to contribute to tumorigenesis, which supports the argument of differentiating these tumors rather than consolidating them under the broad category of CTCs. As explored more below, artificial intelligence (AI) has been shown to accurately predict gene expression or mutations (image-based or morphological profiling) in several human neoplasms. Training an AI model to predict genetic mutation or expression between canine FTCs and MTCs could allow for rapid and cost-effective identification of similar features and in turn promote the practice of precision medicine in the veterinary domain (Ching et al., 2018; Echle et al., 2021; Levine et al., 2019; Sultan et al., 2020). This concept is not limited to CTCs and could be applied to a multitude of neoplasms.

## Additional Corollaries with Human Medicine

Similar to dogs, thyroid cancer is the most common endocrine malignancy in humans and is a leading cause of death among endocrine cancers, usually due to invasion into surrounding tissues or metastasis (Hassan et al., 2020; Pessina et al., 2016; Varrichi et al., 2019). From a demographic perspective, older women have a higher risk of developing thyroid cancer while men tend to experience more aggressive cancer (Barber, 2007; Hassan et al., 2020; Hayes and Fraumeni, 1975; Varricchi et al., 2019). A predilection for the left or right thyroid gland is not reported. Additional risk factors in humans include having a history of goiter or thyroid nodules, a family history of thyroid carcinoma, a low-iodine diet, radiation exposure, obesity and there is

an association between chronic inflammation and the development of thyroid carcinoma (Varricchi et al., 2019).

The WHO classification scheme used for CTCs is superficially similar to the scheme for human thyroid neoplasms (Campos et al., 2014a; Campos et al., 2014b). Differences include that the human classification scheme incorporates pertinent molecular and genetic features, it contains numerous more specific subtypes, variants, and categories of thyroid neoplasms (e.g., Hürthle [oncocytic] cell tumor, mucoepidermoid carcinoma, etc.), and it is more regularly and recently updated (most recently in 2017, 4$^{th}$ edition), as compared to the scheme in domestic animals (originally from the 1970s with the most current, 2$^{nd}$ edition released in 2008) (Bai et al., 2020; Kiupel et al., 2008). Additionally, differentiated thyroid carcinomas in humans appear to be restricted to either papillary thyroid carcinoma (PTC, derived from follicular cells), follicular thyroid carcinoma (derived from follicular cells), medullary thyroid carcinoma (derived from medullary cells), and Hürthle cell carcinoma (thought to be derived from metabolically altered follicular cells with increased numbers of mitochondria) (Al Rasheed and Xu, 2019; Bai et al., 2020; Varricchi et al., 2019).

The incidence of various diagnoses also differs between humans and canids. In humans, PTC predominates (estimated 75-90% of all differentiated thyroid carcinomas) and carries a favorable prognosis with a low incidence of metastasis; however, several described variants carry a worse prognosis (Al Rasheed and Xu, 2019; Antonelli et al., 2018; Bai et al., 2020; Cabanillas et al., 2019; Carver et al., 1995; Chmielik et al., 2018; Hassan et al., 2020; Moore et al., 1984; Yu et al., 2021). Differences in prognosis based on subtype or variant present are in direct opposition to what is currently seen and accepted in dogs (Carver et al., 1995). Similar to CTCs, histologic vascular invasion is an independent predictor of cancer-related mortality in

human thyroid carcinomas, and additional prognostic factors include tumor size, tumor stage, and histologic grade (Campos et al., 2014b).

Regardless of these differences, it is currently believed that canine FTCs (including follicular, compact, mixed, and papillary subtypes) of dogs have overall similar histologic and biologic behavior compared to differentiated human thyroid carcinomas, lending support to the idea of dogs being used as an animal model for human thyroid cancer (Campos et al., 2014c; Chmielik et al., 2018; Haddad et al., 2018; Meijer et al., 2013; Valerio et al., 2017; Yu et al., 2021). Both human FTCs and MTCs are relatively indolent neoplasms with high 10-year-survival rates in humans (Campos et al., 2014c; Chmielik et al., 2018; Haddad et al., 2018; Meijer et al., 2013; Valerio et al., 2017; Yu et al., 2021). However, some consider human MTCs as more aggressive with a less favorable prognosis than differentiated FTCs, while for dogs, there is speculation that MTCs are less malignant (Carver et al., 1995; Fuchs et al., 2020).

Canine MTC is similar to human MTC concerning morphology, cytochemical, and IHC features, and human MTCs are estimated to compose approximately 5-10% or less of all differentiated thyroid carcinomas, which is similar to the originally reported prevalence of canine MTCs (Antonelli et al., 2018; Cabanillas et al., 2019; Campos et al., 2014c; Ceolin et al., 2019; Gambardella, et al., 2019; Hassan et al., 2020; Leav et al., 1976).

Ki-67 can assist in differentiating more poorly differentiated human thyroid neoplasms from well-differentiated or anaplastic thyroid carcinomas, and recently one study proposed a grading scheme for human MTCs using Ki-67, mitotic figures, and necrosis which is thought to accurately predict overall survival (Bai et al., 2020; Fuchs et al., 2020). Unfortunately, there is controversy surrounding the routine use of Ki-67, as other studies have found no prognostic association (Fuchs et al., 2020).

Superficially, therapeutic methods between canine and human patients with thyroid neoplasms are similar for both FTCs and MTCs, with the first step being surgery for curative intent (total thyroidectomy), if possible, followed by chemotherapy (e.g., tyrosine kinase inhibitors, *BRAF* inhibitors, immune checkpoint inhibitors, etc.; singly or in combination), radiation therapy, and even radioiodine therapy for higher-risk patients or those with unresectable neoplasms; for MTCs, lymphadenectomy may be considered to account for nodal metastases (Antonelli et al., 2018; Cabanillas et al., 2019; Ceolin et al., 2019; Gambardella et al., 2019; Meijer et al., 2013; Turrel et al., 2006; Valerio et al., 2017). Treatment with thyrotropine hormone-suppressive levothyroxine may also be used in humans (Valerio et al., 2017). One glaring difference between veterinary medicine and human medicine is that in human thyroid cancer, it is becoming routine to pursue genetic subtyping for common mutations (e.g., $BRAF^{V600E}$) which then permits the use of a customized and highly targeted therapeutic regimen (also known as personalized health care or precision medicine) for the patient in question, especially in cases with advanced disease (e.g., utilization of dabrafenib/trametinib combination therapy for *BRAF*-mutated cancer [Food and Drug Administration approved for anaplastic thyroid carcinoma]) (Al Rasheed and Xu, 2019; Bai et al., 2020; Cabanillas et al., 2019; Cabanillas et al., 2018; Ceolin et al., 2019; Haddad et al., 2018; Valerio et al., 2017; Varricchi et al., 2019). The recent update to the human WHO thyroid neoplasm classification scheme includes data on molecular and genetic derangements which promotes the ability of clinicians to practice precision medicine (Bai et al., 2020). Identification of specific genetic mutations includes the use of next-generation sequencing or single-point mutation testing with Sanger sequencing (gold standard), IHC examination for mutated protein expression, or liquid biopsy (detecting neoplastic cells, their fragments, and DNA in the bloodstream) (Cabanillas et al.,

2019). An additional possible target includes cancer-associated fibroblasts, which are activated fibroblasts that promote cancer cell survival, invasion, and metastasis, induce angiogenesis, and attenuate sensitivity to chemotherapeutics (Wen et al., 2021). Unsurprisingly, veterinary medicine lags human medicine in these aspects, but this may be what the future of cancer treatment looks like for veterinary patients as well. The possible use and efficacy of checkpoint inhibitors in canine thyroid carcinoma are not yet reported, but in a study of dogs with pulmonary metastatic oral malignant melanoma treated with anti-programmed cell death-ligand (PD-L) 1 monoclonal antibody, overall survival time was increased (Maekawa et al., 2021). This same study surveyed several malignant canine cancers for the presence of PD-L1 expression, but CTCs were not included (Maekawa et al., 2021).

MEN, also known as multiple endocrine adenomatosis, MEA, is an inherited human disorder with several subtypes that results in multiple neoplasms in several endocrine organs (Hayes and Fraumeni, 1975; Patnaik et al., 1978; Rosol and Meuten, 2017). MEN type 1 is associated with hyperplasia, adenomas, and/or carcinomas of the thyroid glands, adrenal cortex, and pituitary gland, while MEN type 2 is associated with pheochromocytomas and MTCs (Hayes and Fraumeni, 1975). A similar MEN-like syndrome (type 2) is well-described in bulls, but MEN-like syndromes are less commonly reported in other domestic animals, including dogs (Hayes and Fraumeni, 1975; Rosol and Meuten, 2017).

Based on the molecular and genetic information in humans, FTCs and MTCs have distinct characteristics, and a similar situation may occur in dogs.

## Application of artificial intelligence in human and veterinary medicine

In Chapter 3, the objective of utilizing an artificial intelligence algorithm that can accurately determine the cell of origin (FTCs or MTCs) in canine thyroid carcinomas without the use of ancillary immunohistochemical stains is discussed.

### *Overview of Artificial Intelligence*

Artificial intelligence (AI) is a broad term encompassing all computer-based decision-making processes and is increasingly being used in both human and veterinary medicine, especially in the field of oncology (Chan et al., 2020; Ching et al., 2018; Levine et al., 2019; Moxley-Wyles et al., 2020; Turner et al., 2020). Machine learning is a subset of AI where computers can analyze and identify patterns without much if any, human programming, such that they can learn and improve their accuracy upon being presented with novel but related data (Cohen, 2021; Turner et al., 2020). Specifically, supervised learning occurs when a training set has been annotated by a human observer to aid in the classification of data (Ching et al., 2018). This is followed by reinforcement learning which allows cumulative improvement to minimize the difference between the actual example (ground truth) and predicted value, functioning similarly to operant conditioning; this can also allow for continual improvement of the model over time during its use (Ching et al., 2018; Cohen, 2021; Turner et al., 2021; Mitchell, 2021; Turner et al., 2020; Zuraw et al., 2020). One common example is artificial neural networks (ANNs), which are modeled after biological brain function, such that information is processed through subsequent layers of neurons which allows the computer to achieve more complex analysis with each step and ultimately teach itself how to learn (Ching et al., 2018; Cohen, 2021; Moxley-Wyles et al., 2020; Turner et al., 2020; Wang et al., 2019a). Deep learning (DL) is a specialized branch of AI that utilizes numerous ANNs and is required to develop pattern

recognition from complex data, such as numerous WSIs that contain a variety of features including a range of potentially highly variable differences in staining quality, shapes, and textures (Moxley-Wyles et al., 2020; Turner et al., 2020; Zuraw et al., 2020). Convoluted neural networks (CNNs) are a type of ANN that is considered the current standard in DL image recognition because they extract salient features of images to output an increasingly complex representation by a series of hidden convolution layers (Cohen, 2021; Turner et al., 2020; Wang et al., 2019a; Wang et al., 2019b; Zuraw et al., 2020). This is not an exhaustive overview of all of the additional types of DL methods, but a few additional examples include recurrent neural networks, transfer learning, and generative adversarial networks (Sultan et al., 2020). To utilize DL, an AI system is given a large initial data set and subsequently learns to autonomously identify patterns to maximally separate classes ("separability") (Ching et al., 2018; Cohen, 2021; Levine et al., 2019; Moxley-Wyles et al., 2020; Turner et al., 2021; Turner et al., 2020). Not only must the training set need to be large for a robust algorithm, but it must be accompanied by a verified reference truth with representative characteristics of the population of interest; this can be costly and/or challenging to acquire, especially for those cases with highly variable lesions (Chan et al., 2020; Ching et al., 2018; Sultan et al., 2020). To corroborate the need for a large data set, large multicenter studies have found DL performance increases with the patient number in the training set and reaches a performance plateau after training on 10,000-15,000 histological WSI (Echle et al., 2021). In humans, publicly available WSIs from the Cancer Genome Atlas have been used for several DL studies and assists in providing a large study set with correlating clinical and molecular data (Coudray et al., 2018; Dolezal et al., 2021; Levine et al., 2019; Tsou et al., 2019). Not only does the training data volume contribute to the effectiveness of conventional ANNs, but their success also largely depends on the expertise of the developers and

the capability of the mathematical formulas or empirical image analysis techniques to translate the image characteristics into numerical values (Chan et al., 2020; Ching et al., 2018; Levine et al., 2019; Sultan et al., 2020; Wang et al., 2019a). In contrast, deep learning (often using CNNs) can automatically extract relevant features from a training image set without manual training; these features are expected to be superior to the conventional version, as they have high selectivity and invariance (Chan et al., 2020). Since having access to only a small training data set may be a challenge in model development, data augmentation can be applied to the images to expand the number and diversity of the image set via augmentation training (e.g., image rotations) and/or adversarial training (e.g., small targeted transformations). Transfer learning, where features from one task are re-used for a slightly different project goal, is another way to circumvent the issue of a small training set (Ching et al., 2018; Wang et al., 2019a). Both data augmentation and transfer learning ultimately reduce overfitting and improve generalizability (Ching et al., 2018; Wang et al., 2019a).

Segmentation refers to the identification of structures within images, which may include nuclei, cells, or additional microscopic to macroscopic structures, depending on the desired outcome; or, said alternatively, this is the automated delineation between tissues and tissue structures (Ching et al., 2018; Levine et al., 2019). Segmentation may be semantic, referring to the segmentation of image parts with different meanings (i.e., tissue regions) or it may be instance segmentation referring to the segmentation of discrete objects regardless of whether they belong to the same category or not (i.e., distinguishing each cell within a tissue region) (Wang et al., 2019a). Training a segmentation neural network is a supervised learning process and requires experts (pathologists) to manually annotate (label) the ground truth (Wang et al., 2019a). The successful and accurate segmentation of images may be inhibited by pre-analytic

artifacts (e.g., tissue quality, fixation, slice thickness, etc.), changes in color and brightness (e.g., samples stained at different laboratories), and can reach a performance ceiling due to features that are challenging to distinguish (i.e., a model can segment out highly distinguishable features easily but will have a harder time segmenting out similar features) (Levine et al., 2019; Wang et al., 2019b). The performance of DL is more robust than earlier image analysis programs because they do not solely rely on staining intensity or hand-crafted (manually defined) features, and they are able to take into account neighborhood structural information (e.g., tissue architecture) (Wang et al., 2019a). The loss function of a DL network quantifies the difference between the neural network output and the desired behavior given the network parameters; the training phase is a process to minimize this loss by adjusting network parameters iteratively (Wang et al., 2019a). An iteration is one forward and one backward propagation construct within one training step; in other words, more iterations allow more training repetitions to improve the algorithm's functional goal (Wang et al., 2019a).

*Artificial Intelligence in the Medical Field*

In recent years, the applications of AI in the medical field have rapidly expanded and currently include the ability to analyze images, but can also extend to the examination of genomics, protein structures, and text data (e.g., electronic health records) (Ching et al., 2018). Current DL-based methods now match or surpass the previous state of the art in a diverse array of tasks in patient and disease categorization, fundamental biological study, genomics, and treatment development (Ching et al., 2018). Work is continuing, especially for AI image analysis where large and diverse sample sets are needed for training and models are developed such that misclassification of a challenging or artifactual sample is minimized (Ching et al., 2018).

The two main categories of DL applications in medicine include:

1) basic applications which aim to simplify workflows; this is also known as automated analysis, computer-assisted (or aided) diagnosis (CAD or CADX), or in the context of pathology, pCAD (CAD for pathologists) which would include tumor detection in a biopsy sample or tumor subtyping by morphology
2) advanced applications (also known as knowledge discovery) which provide information the human observer is unable to provide without ancillary testing (e.g., genetic mutations or identification/prediction of biomarkers; "image-based" or "morphological profiling") (Castellino, 2005; Chan et al., 2020; Echle et al., 2021; Laury et al., 2021; Levine et al., 2019; Tosun et al., 2020; Zuraw, 2020).

Early AI image analysis was readily applied to radiology due to the relative ease of acquiring digital images with minimal data loss, and studies showed that AI models could consistently yield results equivalent to or even superior to radiologists in some, specific tasks; recently, similar findings have been found in the realm of histopathologic studies involving pathologists (Castellino, 2005; Chan et al., 2020; Levine et al., 2019; Sultan et al., 2020; Zuraw, 2020). In 1998, the first commercial CAD system was approved for use by the Food and Drug Administration (FDA) which functions as a second opinion screening in mammography (Castellino, 2005; Chan et al., 2020; Levine et al., 2019; Zuraw, 2020). As of 2018, the only CAD application with widespread clinical use is the detection of breast cancer in screening mammography, although there are more than a dozen FDA-approved DL applications in radiology (Bulten et al., 2021; Castellino, 2005; Chan et al., 2020; Echle et al., 2021; Levine et al., 2019; Zuraw, 2020). When the CAD application for mammography screening is used in conjunction with a radiologist, the overall sensitivity is improved which supports the idea of a

synergistic relationship between expert clinicians and the use of AI models (Bulten et al., 2021; Chan et al., 2020).

A reported potential pitfall with the CAD second opinion screening in mammography includes radiologist over-reliance on CAD with decreased vigilance in their interpretations, while another study recommends that CAD models be built with high specificity to minimize the number of false positives and avoid clinician fatigue during screening (Chan et al., 2020). Additionally, CAD tools require performance standards and acceptance testing before clinical use, which includes quality assurance to monitor the consistency and accuracy of the tool over time and to prevent improper use of the CAD that could negatively impact patients (Chan et al., 2020). A similar situation should be anticipated for the use of CAD programs by pathologists.

CAD applications for various medical imaging modalities that have been investigated include disease detection, characterization, staging, treatment response assessment, prognosis prediction, and risk assessment for various diseases; conventionally, these use feature extraction techniques and image processing to distinguish between various states (e.g., normal versus abnormal, or malignant versus benign) (Chan et al., 2020).

Initially, the use of AI was not practical in pathology due to inadequate computers and hardware, including insufficient computational power, graphics processing units, storage space, and an inability to digitize histopathology slides as WSIs (Ching et al., 2018; Coudray et al., 2018; Laury et al., 2021; Sultan et al., 2020; Turner et al., 2020). Now, technology has improved to the point where WSIs are used daily in some veterinary and human diagnostic labs worldwide, which has opened the door for an abundance of available WSIs that could be used for the development and application of AI image analysis (Levine et al., 2019; Moxley-Wyles et al., 2020). Not only do histology slides carry more pixels than what is found in radiologic images,

but they also carry millions of different cells with their morphologies in the context of their spatial arrangement which yields arguably more information (Echle et al., 2021). With this trove of untapped data, DL models can infer high-level labels which could ultimately help guide oncologic treatment decisions, including prediction of genetic alterations, prediction of survival, and prediction of treatment responses as well as shedding light on specific pathogenic mechanisms (Echle et al., 2021). Humans cannot reliably infer these high-level labels from H&E-stained images and require additional methods to reach the same conclusions (Echle et al., 2021). For instance, in one study where a model was trained to predict glioma outcomes in conjunction with common genomic markers, it was found that some previously overlooked malignant features (e.g., adjacent edema and sparsely infiltrated brain) correlated with a higher risk outcome (Levine et al., 2019).

In the realm of pCAD, DL has achieved performance comparable to pathologists in interpreting WSIs for the detection of tumor regions and lymph node metastases (Wang et al., 2019a). Another potential for pCAD DL applications includes quantification of important features in slides, such as the number of cells or mitotic figures; for mitotic figures, extensive work has been done but this remains challenging given the lack of 3D information (the z-axis) (Levine et al., 2019). Currently, mitotic counts are often manually performed by the pathologist which is time-consuming, highly subjective, and does not allow standardized reporting of mitotic scores across pathology laboratories, although there is a push in veterinary medicine to standardize mitotic figure reporting to an area of 2.37 $mm^2$ (Donovan et al., 2020; Sultan et al., 2020). Benefits of pCAD applications may eventually reshape the diagnostic process by improving accuracy and consistency of diagnosis and reporting which expedites the pathologists' workflow and nets in increased growth, productivity, and profit of the institution (Bulten, et al.,

2021; Ching et al., 2018; Echle et al., 2021; Levine et al., 2019; Moxley-Wyles et al., 2020; Sultan et al., 2020; Turner et al., 2020).

The use of AI in pathology will most likely and most effectively resemble a synergistic relationship, as alluded to in radiology, and will include tasks such as tumor detection, grading, and IHC scoring (Aeffner et al., 2017; Bulten et al., 2021; Ching et al., 2018; Moxley-Wyles et al., 2020; Sultan et al., 2020; Turner et al., 2020). Although both pathologists and AI systems loosely utilize algorithmic decision trees to assist with diagnoses, subtyping, and prognostication, each has their advantages and challenges, which highlights their proposed synergistic relationship (Aeffner et al., 2017; Bulten et al., 2021; Turner et al., 2020). Pathologists suffer from inter-pathologist variability due to inconsistency and inaccuracy when counting large quantities and unintentional bias when interpreting routine slides, special stains, and IHCs; but humans excel at interpreting the whole picture and relying on experience (e.g., identification of normal tissues, atypical tissue patterns, or rare cancer subtypes that an AI model may have never seen before) (Aeffner et al., 2017; Bulten et al., 2021; Turner et al., 2020). In contrast, AI systems tend to follow more repeatable interpretation (especially for counting) and allow for more un-biased (objective) staining assessment but cannot interpret nuances in cell types and tissue architecture without similar instances being presented in the training sets (Aeffner et al., 2017; Bulten et al., 2021; Turner et al., 2020).

Thus, expertly trained AI systems excel at accurately completing repetitive tasks that include detecting mitoses, classifying tissues, and analyzing IHCs with accuracy similar to an expert pathologist interpreting WSIs with unlimited time (Coudray et al., 2018; Echle et al., 2021; Laury et al., 2021; Levine et al., 2019; Moxley-Wyles et al., 2020; Turner et al., 2020). Not only do they excel at this, but they can be applied to multiple images concurrently, enabling

rapid and high-throughput analysis (Coudray et al., 2018; Echle et al., 2021; Levine et al., 2019; Turner et al., 2020; Zuraw et al., 2020). It is possible that with rigorously validated systems, AI could significantly lessen the amount of straightforward second opinions sought from secondary pathologists; however, complex cases are unlikely to be undertaken by AI soon (Moxley-Wyles et al., 2020).

Advanced applications, as mentioned previously, can include the identification of previously unknown and actionable knowledge which could change how we develop treatments, categorize patients, or study diseases (Ching et al., 2018; Echle et al., 2021; Levine et al., 2019; Sultan et al., 2020). An example of this would be the concept of image-based profiling (morphological profiling) on histologic WSIs such that models are used for segmentation and feature extraction for functionally annotating genes and alleles, identifying the cellular target of small molecules, identifying disease-specific phenotypes suitable for drug screening, and providing predictions of survival and therapy response (Ching et al., 2018; Echle et al., 2021; Levine et al., 2019; Sultan et al., 2020).

Currently, evaluation of the ever-growing biomarkers in human medicine increases the cost and time for decision-making in routine daily oncology practice and often requires additional tumor tissue for assays (e.g., IHC, in situ hybridization, polymerase chain reaction, or next-generation sequencing) in addition to the routine diagnostic material (often, histologic examination); therefore, pivoting to the use of DL for the analysis and identification of these actionable features could significantly streamline the process of risk stratification and treatment decisions while decreasing the overall financial costs and time (Echle et al., 2021; Sultan et al., 2020; Tsou et al., 2019). Notably, the difference between a prognostic and predictive biomarker is that prognostic biomarkers categorize patients according to the risk of disease progression or

death to better customize the intensity of treatment while predictive biomarkers enable a particular targeted treatment to be chosen for a specific patient group (precision medicine) (Echle et al., 2021).

Most work in the field of histopathologic AI image analysis has been in human breast, lung, and prostate cancers, and recent studies show AI models can act as pCAD tools in the histologic classification of breast tissue in humans (Echle et al., 2021; Polonia et al., 2021; Sultan et al., 2020).

Specifically for digital pathology of human prostate cancer, DL has been developed for tumor detection and grading prostatectomies, tissue microarrays, and biopsies, while several studies show pathologist-level performance within the limits of the study design (Bulten et al., 2021). As an example, the human Gleason grading scheme for prostatic cancers has significant inter- and intraobserver variability, and an AI trained on this scheme had higher sensitivity and higher specificity at classifying tumors into different grade groups as well as providing some indication of where in a grade a case was positioned (i.e., providing that a tumor is grade 3.3 versus grade 3.7), which is, again, beyond the ability of a human observer (Bulten et al., 2021; Moxley-Wyles et al., 2020). Another study found that pathologists with assistance from an AI model trained in the Gleason grading scheme had improved agreement with either an expert reference standard or international experts as compared to the same pathologists without AI assistance (Bulten et al., 2021). AI applications to the Gleason grading scheme are still being investigated and have been highly cited to showcase the consistency and benefit AI models can provide to routine diagnostics.

In another exciting study, a deep CNN was trained to automatically classify human lung tumors into adenocarcinoma, squamous cell carcinoma, or normal tissue, which was successful

and comparable to the pathologist's performance (Coudray et al., 2018). This same study trained the CNN to predict the ten most commonly mutated genes in lung adenocarcinoma and found that 6 of these mutations could be accurately predicted from the images (Coudray et al., 2018).

There is a small amount of literature on the application of AI to human thyroid carcinomas. One proof-of-concept study specifically on human thyroid carcinomas found that DL could accurately predict between mutually exclusive *BRAF* and *RAS* mutations in papillary thyroid carcinomas; these mutations are known to correlate to histopathologic patterns (Tsou et al., 2019). A different study found that deep learning could detect histologic features associated with the amount of either *BRAF* or *RAS* gene expression which aids in distinguishing indolent noninvasive follicular thyroid neoplasms with papillary-like nuclear features (NIFTP) from papillary thyroid carcinomas, which can directly alter treatment recommendations (Dolezal et al., 2021). Another study on human thyroid nodules successfully used a deep CNN to differentiate between normal thyroid tissue, adenoma, nodular goiter, papillary thyroid carcinoma, FTC, MTC, and anaplastic thyroid carcinoma (ATC) (Wang et al., 2019b). The size and staining of the nucleus were the primary classification mechanism used here, and the model unsurprisingly had the most difficulty differentiating between normal tissue and adenomas (Wang et al., 2019b).

Because pathology is often considered to be the gold standard for patient diagnosis, some pathologists have been conservative in adopting digital pathology, among other things, although, as of 2017, WSIs are now considered a class II medical device by the FDA (Levine et al., 2019; Tosun et al., 2020; Wang et al., 2019b).

A current critique of the use of AI in a diagnostic setting is that the model functions as a "black box", in that we do not fully understand how the model generates outputs from a given input; this is similar to those FDA-approved drugs with unknown mechanisms of action (Levine

et al., 2019; Tosun et al., 2020). Research is ongoing in attempting to elucidate this process and includes explainable AI (xAI mechanisms) which may further augment CAD processes (Levine et al., 2019; Tosun et al., 2020). Similarly, because of the state-of-the-art nature, there is a lack of consensus on how pathologists should supervise or work with these models, which may also be circumvented by xAI (Tosun et al., 2020). The goal of xAI is to provide clear justifications to the user for the automated recommendations made in the diagnostic workflow to then promote safety, reliability, and accountability when addressing issues concerning bias, transparency, safety, and causality (Tosun et al., 2020). Legal issues must also be considered before clinical implementation, such as who is liable for machine error (Levine et al., 2019).

*Artificial Intelligence in Veterinary Medicine*

Future applications of artificial intelligence in diagnostic veterinary medicine include many of the previously mentioned applications in all medical imaging modalities. For histopathology applications specifically, these may range from pCAD models to standardize and expedite workflows to advanced applications intending to elucidate actionable outcomes. Much of this has not been explored in the context of veterinary histopathology.

Recent publications utilizing AI in veterinary medicine are wide-ranging and not necessarily restricted to image analysis. Some examples in veterinary medicine include detecting left atrial enlargement in canine thoracic radiographs, predicting survivability and need for surgery in horses that present for acute abdomen (colic), modeling milk productivity on a robotic dairy farm, toxicopathologic applications including detecting compound induced changes or detecting, classifying, and scoring cardiomyopathy in rodents, and enhancing active surveillance for avian influenza (Fraiwan and Abutarbush, 2020; Fuentes et al., 2020; Li et al., 2020; Pischon et al., 2021; Tokarz et al., 2021; Walsh et al., 2019). Currently, very few reports use deep

learning in veterinary pathology. One reports an algorithm out-performed veterinary pathologists in detecting the mitotically most active tumor region, while another article reports a completely annotated WSI image dataset of canine breast cancer to aid in human breast cancer research (Aubreville et al., 2020a; Aubreville et all, 2020b).

## Conclusion

Confirmation of the cell of origin (FTC or MTC) and accurate diagnosis of CTCs remains challenging without the use of ancillary IHC stains. To expedite results and reduce costs of this diagnostic process, the following study explored the viability of applying AI to routine and readily available H&E-stained slides, compared the signalment from a retrospective set of cases to what is described in the literature, compared the accuracy of human pathologist interpretation of compact FTCs and MTCs without immunohistochemistries, and compared the results of the human pathologists without IHCs to interpretations generated from this model's output data. This study will continue to advance the use of artificial intelligence in veterinary medicine. Specifically, because AI models can consistently and cost- and time-efficiently diagnose and subtype tumors, this model could be applied to both prospective and previous CTC studies for more consistent classification of CTCs as well as set up a skeleton for advanced studies on possible actionable outcomes, such as correlating subtle histologic features with altered neoplastic cell genotypes.

# CHAPTER 3: UTILIZING DEEP LEARNING TO ACCURATELY DETERMINE CELL OF ORIGIN IN CANINE FOLLICULAR AND MEDULLARY THYROID CARCINOMAS ON ROUTINELY PROCESSED, H&E-STAINED TISSUE SECTIONS


## Abstract

Canine thyroid carcinomas (CTC) are common endocrine malignancies that include neoplasms derived from both follicular cells and medullary cells which require histopathologic examination with costly and time-consuming IHCs for definitive diagnosis. In this study, 137 retrospective cases with at least one accompanying IHC-stain from the University of Illinois at Urbana-Champaign Veterinary Diagnostic Laboratory (UIUC VDL) between January 2015 and June 2021 with a diagnosis of CTC were identified. These cases included both follicular thyroid carcinomas (FTCs) and medullary thyroid carcinomas (MTCs) derived from cervical and ectopic locations. For all cases, a diagnosis of CTC, subtyping (if applicable), criteria of malignancy (mitotic figures, vascular invasion, desmoplasia, etc.), available IHCs, and slide quality for scanning as a whole slide image (WSI) were evaluated. FTCs were subtyped into follicular, compact, and mixed; while MTCs were not subtyped. For inclusion, cases required at least one of the following IHC-stains: thyroglobulin (Tg), calcitonin, synaptophysin (SYP), and/or chromogranin A (CgA). A review of the diagnoses for each case by a single pathologist yielded that 61.6% (85) were follicular origin (FTCs), 25.4% (35) were medullary origin (MTCs), and 13.0% (18) were equivocal. Of the confirmed diagnoses, the most common diagnosis was compact FTCs followed in descending order by MTCs, mixed FTCs, and follicular FTCs.

The training group encompassed 75 diverse images across 57 cases with 24 instances of mixed FTCs, 22 compact FTCs, 6 follicular FTCs, 22 MTCs, and one instance of a bilateral FTC with differing contralateral diagnoses (left was compact and right was mixed). This latter case

was included to bolster the diversity of the training image set. The developed model is a supervised segmentation deep learning model. Each convoluted neural net (CNN) of the model was trained to discriminate between certain features such that the first, and broadest, layer detected high quality tissue, the next layers distinguishing neoplastic tissue from non-neoplastic tissues (including stroma, non-neoplastic thyroid tissue, parathyroid glands, and lymph node architecture), and subsequent layers were trained to distinguish between MTCs and the follicular and compact patterns of FTCs.

Based on the validation data from the present study, most layers (CNN 1: high quality tissue, CNN 2: carcinoma versus remnant) are ready for use in a diagnostic setting in conjunction with interpretation by a pathologist and the continual addition of images with periodic re-training and re-validation of the model. Although the validation results are promising, caution is still recommended with the use of CNN 3 (follicular FTC pattern versus compact FTC pattern versus MTC pattern), especially for differentiating compact FTCs from MTCs. Further development of the model by adding additional training images of confirmed compact FTCs and MTCs from additional cases is required. The need for caution with the use of CNN 3 is highlighted by a comparison of the algorithm's output to a subset of WSIs that represent compact FTCs and MTCs. Regardless of the segmentation maps, a component of human interpretation is still required for the subtyping of FTCs. This is because the mixed subtype is composed of approximately equivalent regions of both follicular and compact FTC patterns.

The diagnostic challenge for veterinary pathologists reliably differentiating between compact FTCs and MTCs without ancillary testing, such as IHCs, is highlighted in this study based on low measures of agreement (Kappa values). The Kappa values used are between the verified diagnoses and IHC-blinded pathologists (range from 0.10 to 0.60 [poor to moderate

agreement], average of 0.30 [poor agreement]) and inter-pathologist agreement (range from -0.05 to 0.41 [poor to weak agreement], average of 0.13 [poor agreement]) for a subset of WSIs.

## Introduction

The primary objective of this project is to develop and validate an artificial intelligence (AI) deep learning algorithm that can accurately determine the cell of origin (follicular or medullary) in hematoxylin and eosin (H&E) stained, whole slide images (WSIs) of canine thyroid carcinomas (CTCs) without the use of ancillary immunohistochemical (IHC) stains. This is because some types of CTCs are challenging to discriminate between without the assistance of costly IHC stains, which may take days to weeks to be performed. Looking forward, routinely achieving an accurate and rapid diagnosis in diagnostic and research settings is useful to ensure correct treatment protocols are implemented, conclusions from studies are not skewed by failing to discriminate between types of CTCs, and could contribute to the development and implementation of highly specific therapeutic protocols. In this study, the primary hypothesis is that an algorithm can accurately determine the cell of origin (FTC or MTC) in CTCs on routine H&E-stained histopathology slides. The secondary objective was to review and compare demographic information and histologic characteristics between non-subtyped FTCs or MTCs and between subtyped FTCs and MTCs. Tertiary objectives were to evaluate the ability of veterinary anatomic pathologists to correctly identify compact FTCs from MTCs on H&E alone and further evaluate the function of the developed AI model in comparison to these IHC-blinded pathologists and the original IHC-based diagnosis.

## Materials and methods

***Case Identification and review.*** One hundred and thirty-seven (137) archival necropsy or surgical biopsy cases from the University of Illinois at Urbana-Champaign Veterinary Diagnostic

Laboratory (UIUC VDL) between January 2015 and June 2021 with an original diagnosis of CTC, including FTCs and MTCs, with at least one accompanying IHC, were identified using the search terms: thyroid carcinoma, thyroid follicular carcinoma, follicular carcinoma, medullary thyroid carcinoma, medullary carcinoma, C-cell carcinoma, and parafollicular carcinoma. Cases were also searched for by the presence of either thyroglobulin (Tg) or calcitonin IHC. Cases were included if archival slides had an accompanying IHC of either thyroglobulin (Tg), calcitonin, synaptophysin (SYP), or chromogranin A (CgA) that were available for manual review.

**Table 1** contains the antibody and dilution information that the UIUC VDL employs. H&E-stained slides with fewer IHCs and special histochemically stained slides from all identified cases were then re-examined by a veterinary anatomic pathology resident (JMA) using brightfield microscopy for quality, verification of the original diagnosis (FTC or MTC), and subtyping of FTCs (if not previously performed). These are hereby referred to as the "verified cell or origin" or "verified diagnosis," respectively. The definitive microscopic diagnosis and confirmation of cell of origin (FTC or MTC) rest with IHC interpretation, as outlined in **Table 2**. The ground truth used for AI model training on the H&E-stained slides was therefore determined by the IHCs. Eighteen cases were unable to be definitively diagnosed by the provided H&E-stained and IHC-stained slides. For these cases, JMA consulted one of two board-certified anatomic veterinary pathologists for assistance in reaching the final diagnosis. The inability of definitive diagnosis in this case subset was mainly due to equivocal IHC staining patterns or lack of a complete IHC panel; of the latter, the main issue was determining between a compact FTC versus MTC, but in a few instances, challenge arose with subtyping FTCs into mixed or follicular patterns. Two of these eighteen cases were considered poorly differentiated, and one

may represent a carcinosarcoma. Equivocal cases were excluded from model development and statistical testing.

Additionally, neoplasms were evaluated for the histologic presence of subendothelial invasion, intravascular neoplastic emboli, osseous metaplasia (with or without mineralization), necrosis, desmoplasia (also known as a scirrhous response), amyloid, and the number of mitotic figures in ten representative high power (400x) fields (Newkirk et al., 2017). The amount of necrosis, if present, was scored as follows:

1 for necrosis that encompasses 1 to 25% of the tumor as a whole,

2 for necrosis that encompasses 26 to 50% of the tumor as a whole,

3 for necrosis that encompasses 51 to 75% of the tumor as a whole, and

4 for necrosis that encompasses 76 to 100% of the tumor as a whole.

In cases with equivocal interpretation, additional board-certified veterinary anatomic pathologists were consulted (MDV and KLB) which generally resulted in the determination that additional IHCs are needed for definitive diagnosis. Cases were excluded if H&E-stained slides of the CTC were absent, or if upon re-examination, the neoplasm was thought to be of parathyroid or other tissue origins. All acceptable slides were then submitted (Charles River Laboratories) for scanning, yielding 1076 unique WSIs consisting of a combination of routine H&E-stained slides containing CTC lesions, IHCs, and a few miscellaneous special or immunohistochemical stains, such as Congo Red, Giemsa, cytokeratin IHC, and CD31 IHC that were ordered as part of the original diagnostic workup.

***Development of the AI model.*** All WSIs were uploaded to the Aiforia Cloud platform (Aiforia Inc., Cambridge, MA, USA) as .svs files without additional processing or accompanying metadata regarding demographics, diagnoses, etc. One hundred and nineteen (119) cases (452

H&E-stained WSI) with unequivocal diagnoses (follicular subtypes or MTCs) were selected for possible use in model development. Of these, 75 diverse WSIs across 57 cases (~17% of the total available H&E-stained WSIs; ~45% of the total available cases) were selected based on diverse CTC tissue architecture, diagnoses, and good quality WSI scanning (minimal scan artifacts such as scan lines or blurry bands that severely compromise image quality). The remaining 377 WSIs from the cases with unequivocal diagnoses were withheld for validation and possible further testing.

Development and training of the models were performed with the Aiforia Create platform (initially under Aiforia version 5.1 and the remaining majority and final training under Aiforia version 5.2). One person (JMA) trained and developed the model with guidance from an Aiforia representative.

**Figure 2** shows a schematic overview of the model structure. The superficial three layers (CNN) of the model were trained by supervised learning for segmentation, while the remaining CNN was intended to be an object detection layer.

The first CNN ("high quality tissue", CNN 1) was trained to detect any tissue, which, despite the name, included tissue folds, superimposed fragments of tissue, and small patches of blurriness. This is more appropriately called a tissue detector layer. CNN 2 is a child layer to CNN 1 and is a binary segmentation model to separate "carcinoma" (also known as neoplastic thyroid tissues) and "remnant" tissues (also known as non-neoplastic tissues, including stroma, non-neoplastic thyroid tissue, parathyroid glands, and lymph node architecture). Within this layer, skeletal muscle was trained out as background rather than stroma due to issues with the model identifying skeletal muscle bundles as thyroid colloid follicles. CNN 3 is a child layer to the "carcinoma" CNN and is a trinary segmentation model to separate between histologic

patterns of tumor growth. Only follicular and compact FTC subtypes are utilized because the diagnosis of mixed FTC is made when there are approximately equal amounts of follicular and compact patterns. During early model development, training a specific layer for a mixed pattern was deemed not feasible due to challenges in model accuracy. All subtype layers then had one additional, interconnected, child CNN ("mitotic figures"), which was intended to function as a mitotic figure counter. The configuration for the mitotic figure layer allows mitotic figures to be detected in all neoplastic tissue regardless of the subtype in question.

Training annotations for segmentation layers were made by encircling representative regions with the desired characteristic(s) (colored circles), while the training regions were indicated by encircling appropriately annotated regions (black circles) (**Figure 2**). Typically, training annotations (colored circles) were drawn larger than training regions (black circles) to ensure all tissues were accounted for; in some instances, only training regions without training annotations were made to teach the model background information (e.g., to help the model distinguish between tissue on the slide versus blank areas of scanned slides). In some instances, especially for CNN 2, interface regions delineating neoplastic and non-neoplastic patterns were desirable to improve the segmentation margins and learning of the model.

Because CNN 4 was intended to be an object detection layer for counting mitotic figures, the annotation style is slightly different. For these annotations, a stamp-like object marker was used to make training annotations which were then surrounded by a training region. The size of this marker can be changed, but 12 um diameter appeared optimal. Mitotic figures were annotated according to Donovan et al. (2021).

Each CNN was manually annotated and trained in sequential order starting first with CNN 1 (high quality tissue). After providing a baseline set of annotations in a small number of

training images, the model was trained at 100 iterations (100i). After each training, the Aiforia program's verification pane provides a concise, rank-ordered review of discrepancies (such as false positives and false negatives) between the annotations provided and the model's prediction. The use of this pane allows for the refinement of existing training annotations. Following higher iteration training, the algorithm may also perform an analysis of regions of the WSI or the whole WSI itself to allow visual assessment of the current overall function of the model. Similar to the verification pane, this allows for further refinement of existing annotations as well as a visual guide of where new annotations should be placed (typically in areas where the model is incorrectly segmenting the tissue). With the early promising function of the model, most of the additional training images were added at one time until the total 75 images were annotated. After annotations were refined and/or added, the model was re-trained, usually with increasing iterations and occasionally with adjustments of advanced parameters for optimization. This process was repeated until the model reached subjective proficiency by examination of the verification pane results and analysis of regions' layer segmentation masks.

At this point, a final high-iteration training (between 10,000i and 15,000i) was performed for each segmentation CNN to allow for further refined learning and confirmation of appropriate function. After this, the individually trained CNNs were internally released and collated into a final model. The final pre- and post-training CNN parameters are summarized in **Table 3**.

***AI Model Validation.*** Twenty-five (25) diverse WSIs from 25 discrete cases that were withheld from the training set were selected for validation. To be included, each case required accompanying thyroglobulin and calcitonin, synaptophysin, or chromogranin IHCs, since these are required for accurate histologic interpretation, as outlined above. Thirteen (13) of these 25 WSIs (52%) were from cases where a sister WSI was used to train the model. Although these

sister WSIs derive from the same patient, their use is acceptable because they contain differences in architectural arrangements, artifacts, amount of neoplastic tissue, and/or location (e.g., lymph node metastasis rather than a primary mass). These 25 WSIs for validation encompassed 5 diagnoses of mixed FTCs, 6 compact FTCs, 3 follicular FTCs, 11 MTCs. The images were randomly assigned a new letter ID (from A to Y) in Microsoft Excel to preserve the blindness of the validators. The model was then applied to these images to generate segmentation layer masks for each CNN.

Three (3) board-certified general veterinary anatomic pathologists with varying years of diagnostic experience assisted in the validation of this model. The pathologists were asked to visually assess and score each layer segmentation mask that the model generated. The scoring system utilized is similar to a system reported in the literature for a histologic image segmentation AI model (Pai et al., 2021) and an additional undisclosed study from a different Aiforia client that is in progress for publication. The scoring key is as follows: 1 (perfect or near-perfect accuracy [95-100%, no significant errors]); 2 (very good accuracy [80-95%, minor errors]); 3 (good accuracy [70-80%, significant errors but still captures the features well]); and 4 (insufficient accuracy [<70%, significant errors compromising feature recognition]). Validators were provided minimal additional instructions or guidance and were asked to classify each image based on their interpretation of the cell of origin (follicular or medullary; also known as FTC or MTC) as well as what their ultimate diagnosis would be (follicular, compact, or mixed FTC or MTC).

***IHC-blinded Pathologist success in differentiating compact FTC and MTC (Tertiary Objective).*** Twenty (20) WSIs including 10 compact FTC and 10 MTC each were given to 3 general veterinary anatomic pathologists with varying years of diagnostic experience that were

not associated with the training or validation of the AI model. The pathologists were tasked to provide whether they interpreted a particular tumor as a compact FTC or MTC without ancillary IHCs ("IHC-blinded pathologists"). Ten (10) WSIs were from cases where a sister WSI was used in the training set, 2 WSIs were from cases with another image included in the validation set, and the remaining 8 WSIs were included in the reserved set. Both compact FTCs and MTCs are represented in all groups. These images were also randomly assigned a new letter ID (from A to T) in Microsoft Excel.

Following the validation process, the algorithm was applied to this image set. The interpretation of model results was only by JMA via both visual assessment or by taking the highest model-generated percent segmentation area for CNN 3 (follicular FTC pattern, compact FTC pattern, or MTC pattern). This percentage is out of the total area that the model identified as "carcinoma" tissue. For visual assessment, the neoplasm was classified by the predominant colored layer mask with compact FTC as dark blue, MTC as light blue, and follicular FTC as red. The use of the category "FTC" rather than the more specific diagnosis of "compact FTC" was preferred to highlight the main intended function of the model (differentiation of FTCs from MTCs).

***Statistical analyses.*** R/RStudio [RStudio 2021.09.1+372 "Ghost Orchid" Release (8b9ced188245155642d024aa3630363df611088a, 2021-11-08) for Windows Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) QtWebEngine/5.12.8 Chrome/69.0.3497.128 Safari/537.36) was used for most statistical testing with an alpha of 0.05. Microsoft Excel (Version 2002 [Build 12527.20252 Click-to-Run]) was used for various calculations including means, variance, standard deviations, and Cohen's

Kappa, and was also used to build confusion matrices and organize data for import into R/RStudio.

For various components of the study binary or quaternary confusion matrices were constructed to compare many things. Categories of binary confusion matrices constructed for this study with their specific comparisons include:

- Agreement of validator pathologists
    - Pairwise agreement of interpretations of validator pathologists and the verified cell of origin (FTC or MTC)
    - Pairwise agreement between each validator pathologist (interobserver agreement) for the cell of origin (FTC or MTC)
- Agreement of IHC-blinded pathologists
    - Pairwise agreement of diagnosis of IHC-blinded pathologists to the verified diagnosis (compact FTC or MTC)
    - Pairwise IHC-blinded interpathologist agreement of the cell of origin (FTC or MTC)
- Agreement with the model's predictions
    - Agreement of the verified cell of origin (FTC or MTC) versus results generated by the model (as visually interpreted by JMA)
    - Agreement of the verified cell of origin (FTC or MTC) versus results generated by the model (by taking the highest percent segmentation area)
    - Pairwise agreement of the cell of origin (FTC or MTC) between each IHC-blinded pathologist and the model (as visually interpreted by JMA)

- Pairwise agreement of the cell of origin (FTC or MTC) between each IHC-blinded pathologist and the model (by taking the highest percent segmentation area)

For each comparison, Cohen's Kappa, the accuracy, and their respective averages (if applicable) were calculated. Accuracy (the proportion of agreeing predictions) was also calculated by summing the true positives and true negatives and dividing by the total number of cases (e.g., 25 for the first category).

Quaternary confusion matrices were constructed to compare the:

- Pairwise agreement of validation scores from all pathologists for all layers
- Pairwise agreement of each validator pathologist's ultimate diagnosis (follicular FTC, compact FTC, mixed FTC, or MTC).

For the second category of confusion matrices, Fleiss' Kappas were calculated and included evaluation of the:

- Pairwise agreement between each validator pathologist's ultimate diagnosis (subtyped FTCs or MTC) and the verified diagnosis
- Total agreement between all validator pathologists' ultimate diagnoses (FTC subtypes and MTC) (total interpathologist agreement)

For the first set of pairwise comparisons, the average Fleiss' Kappa was calculated and interpreted as outlined above.

As an alternative evaluation of the data, the average score, variance, and standard deviation for each segmentation layer mask for each validator pathologist were calculated. Then, the average score, variance, and standard deviation across all images and all validator pathologists for each segmentation layer mask were calculated. This was calculated first

including all images and then again excluding images where the total consensus of the cell of origin (FTC or MTC) was not achieved between all validator pathologists.

For evaluating demographic data, cases with missing data were excluded from the relevant tests as well as the single herding dog with bilateral FTC and differing contralateral subtypes. Age groups were made to attempt to equally distribute the age ranges. Given the wide spectrum of dog breeds and "mixed breed dogs", dogs were instead grouped by the American Kennel Club (AKC) groups (2020). Furthermore, breeds considered to have increased risk were kept separate, while mixed breed dogs with at least one dog breed provided (e.g., "boxer-mix") were included with the breed listed. In instances where multiple breeds were listed, the animal was classified into the first given breed (e.g., classified as a boxer for a "boxer-lab mix"). This applies to the beagle, boxer, golden retriever, and Labrador retriever strata.

Demographic features (age groups, stratified sex, non-stratified sex, location, and breed) as well as histologic features (subendothelial invasion, intravascular invasion, presence of osseous metaplasia, presence of necrosis, and presence of desmoplasia) were each compared in a pairwise fashion against categories of either all FTCs and MTCs (2 categories) or subtyped FTCs (follicular, mixed, or compact) and MTCs (4 categories). These pairs were initially evaluated for independence using a chi-square test. A Pearson's chi-squared test was then evaluated for a statistically significant relationship between independent pairs. A likelihood ratio was used to evaluate the contribution of the remaining categories to the development of all FTCs or MTCs; or of subtyped FTC or MTC.

Quantitative demographic values (mean age and standard deviation) and histologic features (mean number of mitotic figures and standard deviation; mean necrosis score and standard deviation) were calculated for all CTCs, as well as for the different diagnoses. One-way

analysis of variances (ANOVAs) were performed to evaluate whether there were statistically significant differences in mean age on all FTCs and MTCs or subtyped FTCs and MTCs. Before the ANOVA, Levene's test was used to first confirm that population variances were equal for each.

## Results

The training WSI set encompassed 24 diagnoses of mixed FTCs, 22 compact FTCs, 6 follicular FTCs, 22 MTCs, and one instance of a bilateral FTC with differing contralateral diagnoses on the same WSI (left gland was diagnosed as a compact FTC while the right was diagnosed as a mixed FTC). Within this image set, there were between 1 to 3 distinct WSIs from the same animal, deriving from different areas of the neoplasm.

Understanding how to interpret AI models to make testable hypotheses about the system under study remains an open challenge, as of 2018 (Ching et al.). A common approach to validating deep learning algorithms in oncologic histopathology is to assess the performance of algorithms as compared to an expert pathologist (Sultan et al., 2020). Because a consensus on how to evaluate this data does not yet exist, attempts were made to follow what previous studies have done.

To evaluate the model's function, Kappa values, confusion matrices, and a few other methods were utilized. Cohen's Kappa is used to compare 2 raters using a categorical variable, while Fleiss' Kappa may be used for comparing 2 or more raters using 2 or more categorical variables. Both are used throughout this study. Interpretation of all Kappa statistics are as follows: <0.40 (poor agreement); 0.40-0.54 (weak agreement); 0.55-0.69 (moderate agreement); 0.70-0.84 (good agreement); and 0.85-1.00 (excellent agreement) (Schober et al., 2021).

Confusion matrices may be used to visually compare the agreement between two assays or observers and are commonly used in evaluating artificial intelligence models. Binary (2 x 2) or quaternary (4 x 4) confusion matrices were built to compare a variety of things, as outlined below. For confusion matrices, instances of agreement occur along the diagonal of intersection spanning from the upper left to lower right cells, while instances of disagreement are represented by anything outside of this diagonal. All confusion matrices are provided in **Supplementary Figure 3**.

Cohen's Kappas for the pairwise agreement of the interpretations of the validator pathologists and verified cell of origin (FTC or MTC) are 0.84 (good agreement; pathologist A), 0.92 (excellent agreement; pathologist B), and 0.92 (pathologist C) with a mean of 0.89 (excellent agreement). The accuracies for these comparisons are 0.92, 0.96, and 0.96 with a mean of 0.95.

Measures of agreement between validator pathologists (interobserver agreement) for the verified cell of origin are provided in **Table 4**. Cohen's Kappas for the pairwise agreement between each validator pathologist (interobserver agreement) for the cell of origin (FTC or MTC) are 0.84 (good agreement), 0.92 (excellent agreement), and 0.92 (excellent agreement) with a mean of 0.89 (excellent agreement). The accuracies for these comparisons are 0.92, 0.96, and 0.96 with a mean of 0.95.

Fleiss' Kappa for the pairwise agreement between each validator pathologists' ultimate diagnosis (subtyped FTCs or MTCs) and the verified diagnosis are 0.822 (good agreement; pathologist A), 0.593 (moderate agreement; pathologist B), and 0.712 (good agreement; pathologist C) with an average of 0.709 (good agreement).

Fleiss' Kappa for the total agreement between all validator pathologists' ultimate diagnosis (FTC subtypes and MTC; total interpathologist agreement) was 0.626 (moderate agreement).

Measures of agreement between IHC-blinded pathologists and verified IHC-based diagnosis (compact FTC or MTC) are provided in **Table 5**. Cohen's Kappas for the pairwise agreement of IHC-blinded pathologists to the verified diagnosis are 0.60 (moderate agreement), 0.20 (poor agreement), and 0.10 (poor agreement) with a mean of 0.30 (poor agreement). Accuracy for each pair is 0.80, 0.60, and 0.55, with a mean of 0.65.

Measures of agreement between IHC-blinded pathologists (interobserver agreement) for cell of origin (FTC or MTC) are provided in **Table 6**. Cohen's Kappas for the pairwise IHC-blinded interpathologist agreement of the cell of origin are 0.41 (weak agreement), -0.05 (poor agreement), and 0.03 (poor agreement), with a mean of 0.13 (poor agreement). Accuracy for each pair is 0.70, 0.55, and 0.65, with a mean of 0.63.

Cohen's Kappa for comparing the agreement of the verified cell of origin and visual assessment of the model's results is 0.60 (moderate agreement) with an accuracy of 0.80.

Cohen's Kappa for comparing the agreement of the verified cell of origin and taking the highest percent segmentation area is 0.50 (weak agreement) with an accuracy of 0.75.

Cohen's Kappas for the pairwise agreement of cell of origin between IHC-blinded pathologists and visual interpretation of the model are 0.42 (weak agreement), 0.07 (poor agreement), and 0.04 (poor agreement) with an average of 0.18 (poor agreement). Accuracy for each pair is 0.70, 0.50, and 0.45 with a mean of 0.55.

Cohen's Kappas for the pairwise agreement of cell of origin between IHC-blinded pathologists and taking the highest percent segmentation area provided by the model are 0.31

(poor agreement), -0.06 (poor agreement), and -0.12 (poor agreement) with an average of 0.04 (poor agreement). Accuracy for each pair is 0.65, 0.45, and 0.40 with a mean of 0.50.

Considering validator interpathologist agreement on cell of origin (FTC or MTC) for the 25 validation WSI subset, consensus was achieved for all images except for 2. In these instances, only one of the three (1/3) validator pathologists disagreed. For one image, the majority diagnosed an MTC while one pathologist diagnosed FTC; the reverse is true for the other image. Therefore, if the determination of the cell of origin (FTC or MTC) for these images is by the majority, the cell of origin ends up the same as the verified diagnosis, which is supported by good to excellent Cohen's Kappa values between the verified cell of origin and each validator pathologist.

Macroscopic images of the model's CNN 3 WSI segmentation masks for the comparison of the model's predictions to IHC-blinded pathologists in determining compact FTCs and MTCs are provided in **Figure 3**, while **Table 7** compares interpretation of the model's predictions with each IHC-blinded pathologist. Comparison of the interpretation of the model's predictions to the verified cell of origin found misclassification by visual assessment in 4 of 20 images (images F, P, Q, and R) and misclassification by segmentation area percentage in 5 of 20 images (images B, F, P, Q, and R). In one image (image K), the follicular FTC pattern visually dominated and had the highest segmentation area percentage (43.7%). This represents a misclassification, as only WSI originally diagnosed as compact FTC or MTC were selected for this study component. For this image, the compact FTC segmentation area percentage was 15.27%, while the medullary segmentation area percentage was 14.57%. Looking at the specific segmentation patterns, interpretation of the algorithm's results yielded three FTCs misclassified as MTCs (images F, P, and Q), while only one misclassification of an MTC as an FTC (image R). Looking at the highest

segmentation percent areas, three FTCs were misclassified as MTCs (images F, P, and Q), while two MTCs are misclassified as FTCs (images B and R).

The highest percentage of segmentation patterns for each image ranged from 43.7% (image K, follicular FTC pattern) to 99.45% (image M, compact FTC pattern), while the absolute difference between the compact FTC pattern and MTC pattern for each image ranged from 0.7% to 99.352%.

Of the incorrectly classified images by both visual assessment and percentage, only 2 images had a sister slide that was involved in the training set. Therefore, the remaining 8 images with a sister slide involved in training classified the image correctly.

Given that a scoring system from non-randomized rater pathologists with the intent to evaluate agreement was used for rating the model, a two-way mixed effect, absolute agreement, single raters intraclass correlation coefficient (ICC) was selected and performed for each layer. The ICC scores were first calculated using scores from all validation images and then re-calculated excluding images where the total consensus of the cell of origin (FTC or MTC) was not achieved between all validator pathologists. ICC scores are interpreted like the Kappa statistics, as outlined above, and the ICC scores for all images ranged from 0.00 to 0.59 (poor to moderate agreement). The ICC scores for all images with validator pathologist consensus on the cell of origin ranged from 0.00 to 0.57 (poor to moderate agreement).

The averages, variances, and standard deviations for each segmentation layer for each image from each validator pathologist are listed in **Table 8**. Below, is the overall average score for each segmentation layer for all images.

1. CNN 1, High Quality Tissue: 1.11 (SD 0.27)
2. CNN 2, Carcinoma: 1.48 (SD 0.26)

3. CNN 2, Remnant: 1.52 (SD 0.24)
4. CNN 3, Follicular: 1.83 (SD 0.55)
5. CNN 3, Compact: 2.16 (SD 0.75)
6. CNN 3, Medullary: 2.21 (SD 0.82)

Unfortunately, the mitotic figure counter did not successfully make it through the pre-validation phase due to excessively high false-positive rates, relatively rare high-quality examples among the 75 WSIs used for training, and project timeline constraints. The most recent model has 986 high quality manual annotations, but the algorithm identified a total of 3,185 mitotic figures yielding a total object error of 323.02%, a false positive percentage of 321.50%, and a false negative percentage of 1.52%. This model also reports a precision of 23.45%, a sensitivity of 98.48%, and an F1 score of 37.88%. This model was trained at 5000i with a training loss of 0.1162. Altered advanced parameters include extra complex complexity and 80% maximum object overlap for layer features, and a mini-batch size of 160 and 500i iterations without progress for training procedures. Complexity refers to how difficult it is to recognize features from the rest of the image, mini-batch size essentially splits the data set into smaller batches to avoid running out of available memory, and iterations without progress refer to a mechanism where the training will stop if no progress is made after a certain number of iterations.

Clinical information provided by the client generally ranged from reporting a left, right, or bilateral thyroid mass, neoplasm, or carcinoma; to a thyroid mass without a specified side or location; or a ventral cervical or laryngeal mass. Few cases were from ectopic locations, such as a cranioventral cervical mass, heart base mass, subcutaneous mass at the right thoracic inlet, or a pericardial mass associated with the brachiocephalic trunk. In general, the accompanying clinical information was highly variable and was mostly restricted to indicating a thyroid mass with or

without providing a location. Rarely, additional information, such as results from a variety of clinical assessments, clinical signs, or concerns about metastasis, was provided. Comparison of presenting clinical signs or comorbidities was not pursued due to the incomplete and inconsistent reporting of this information in the study sample.

**Table 9** summarizes the distribution of all FTCs and MTCs or subtyped FTCs and MTCs between various demographic categories. Animals were classified by stratified sex (sex with spay or neuter status) and non-stratified sex (male or female), by numeric age, by age groups, by breed, and by neoplasm location. There is one case of a 15-year-old, AKC herding group dog that had bilateral FTC but was subtyped differently on each side (left was compact and right was mixed); this animal was excluded from the calculations of mean age, mitotic figures, and scoring of necrosis as well as subsequent statistical analyses but yielded a total of 138 diagnoses from 137 animals. The remaining animals with bilateral neoplasms were diagnosed with the same type of CTC on each side and were subsequently considered as one diagnosis for statistical testing.

There were 85 FTCs (61.6%), 35 MTCs (25.4%), and 18 cases with equivocal diagnoses (13.0%). Of the FTCs, the compact subtype was most common with 51 cases (37.0% of the total; 60% of FTCs), the mixed subtype was second most common with 27 cases (19.6% of the total; 31.8% of FTCs), and the follicular subtype was least common with 7 cases (5.1% of the total; 8.2% of FTCs). When considering only cases with unequivocal diagnoses (n = 120), FTCs compose 70.8% while MTCs compose 29.2%. The FTC subtypes considering only unequivocal diagnoses are as follows: compact was 42.5% of total unequivocal diagnoses, mixed was 22.5% of total unequivocal diagnoses, and follicular was 5.8% of total unequivocal diagnoses.

Age groups, stratified sex, non-stratified sex, location, and breed were compared against categories of either the cell of origin (FTC or MTC) or specific diagnosis (FTC subtypes or

MTC). The only independent pair was the age groups and categorizing tumors into either FTC or MTC. Subsequent Pearson's chi-squared test confirmed there is a statistically significant relationship between age groups and developing CTC. (p-value = 0.01477). A likelihood ratio was used for the remaining pairs, and none were significant. A statistically significant relationship using the Pearson's chi-squared test (P-value = 0.01477) was found between age groups (3 to 6 years, 7 to 10 years, and 11-15 years) and developing either all FTCs or MTCs. Based on the age groups, the 7 to 10 age range contained the most cases.

The overall mean age between all diagnoses is 9.2±2.3 years, while all FTCs had a mean age of 9.6±2.3 years and MTCs had a mean of 8.4±2.3 years. Separating FTC subtypes, follicular FTCs had a mean age of 9.6±1.6 years, mixed FTCs had a mean of 9.0±2.5 years, and compact FTCs had a mean of 9.8±2.2 years.

ANOVA was used to determine if there were statistically significant differences in mean age on all FTCs or MTCs (2 categories) and subtyped FTCs (follicular, mixed, and compact) or MTCs (4 categories). Levene's test was done to first confirm population variances were equal for both tests (first test F = 0.0128, p-value: 0.9101; second test F = 0.3344, p-value = 0.8005). The first ANOVA revealed that were was a statistically significant difference in mean age between FTCs (9.6 years [1.6 SD] years) and MTCs (8.4 years [2.3 SD]) (F = 5.648, p-value = 0.0192). The second ANOVA revealed that there was not a statistically significant difference in mean age between at least two of these groups (F = 2.613, p-value = 0.0549).

Examined histologic features include the presence of subendothelial invasion, intravascular invasion, osseous metaplasia, necrosis, and desmoplasia. Additionally, mitotic counts from ten high-power (400x) fields in the area of the tumor with the most mitoses were

performed. In cases with necrosis, the amount of necrosis present within a neoplasm was scored, as previously described.

The presence of subendothelial invasion, intravascular invasion, osseous metaplasia, necrosis, and desmoplasia were compared against categories of either all FTCs and MTCs or subtyped FTCs and MTCs. A chi-square test confirmed independence in four pairs which include 1) intravascular invasion and FTC or MTC, 2) desmoplasia and FTC or MTC, 3) desmoplasia and FTC subtype or MTC, and 4) presence of necrosis and FTC or MTC. A Pearson's chi-squared test was performed for each pair, and the p-values for each pair are as follows: 1) 1.0, 2) 0.004885, 3) 0.012, and 4) 0.2912. Therefore, statistically significant relationships were only identified between desmoplasia and both CTC cell of origin and subtypes. There were 34 cases with desmoplasia. Of these, 13 were FTCs, 14 were MTCs, and 7 were equivocal. Of the FTCs, 9 were compact, and 2 each were mixed or follicular. The relationships between intravascular invasion and cell of origin and the presence of necrosis and cell of origin are statistically insignificant. A likelihood ratio was used for the remaining pairs, and a statistically significant relationship was confirmed between the presence of osseous metaplasia and subtypes (p-value 0.00041158). Twelve (12) cases had the presence of osseous metaplasia, and these included 5 mixed FTCs, 3 follicular FTCs, 0 compact FTCs, 2 MTCs, and 2 neoplasms with equivocal diagnoses

The overall mean number of mitotic figures between all diagnoses is 5.5±4.3, while FTCs had a mean of 5.3±4.4 mitotic figures and MTCs had a mean of 6.1±4.5 mitotic figures. Separating FTC subtypes, follicular FTCs had a mean of 3.9±2.6 mitotic figures, mixed FTCs had a mean of 5.6±4.8 mitotic figures, and compact FTCs had a mean of 5.3±4.5 mitotic figures. ANOVA was performed to evaluate for statistical significance of mean mitotic figures on all

FTCs or MTCs, and on subtyped FTC and MTCs. Levene's test first confirmed that population variances were equal for both tests (first test F = 0.6374, p-value: 0.4263; second test F = 0.7727, p-value = 0.5116). Both ANOVAs revealed that there were no statistically significant differences in mean mitotic figures between at least two groups (first test F = 0.824, p-value = 0.366, second test F = 0.56, p-value = 0.642).

For cases with necrosis, the overall mean score for necrosis was 1.6±0.8, while all FTCs had a mean of 1.6±0.7 and MTCs had a mean of 1.6±0.9. Separating FTC subtypes, follicular FTCs had a mean score of 2.2±0.8, mixed FTCs had a mean of 1.7±0.8 mitotic figures, and compact FTCs had a mean of 1.4±0.6. ANOVA was performed to evaluate for statistical significance of mean scores on all FTCs or MTCs and subtyped FTC and MTCs. Levene's test first confirmed that population variances were equal for both tests (first test F = 0.0263, p-value: 0.8715; second test F = 0.7994, p-value = 0.4977). Both ANOVAs revealed that there were no statistically significant differences in mean scores between at least two groups (first test F = 0.026, p-value: 0.872; second test F = 2.089, p-value = 0.108).

Amyloid was not detected in any neoplasm, so no statistical testing was pursued.

## Discussion

After performing the literature review, the need for a consistent, cost-effective, and efficient method of differentiating FTCs and MTCs was obvious. This is because even current (2020 and 2021) clinical literature regarding treatment modalities frequently fails to distinguish these entities yielding conclusions about CTCs as a whole, which may not accurately reflect what is happening on a genomic or biochemical level (Giannasi et al., 2021; Hassan et al., 2020; Jegatheeson et al., 2021; Nadeau and Kitchell, 2011). This systemic failure of distinguishing MTCs from FTCs coupled with the paucity of comparative studies between these entities may

perpetuate skewed data with lower MTC prevalence and/or yield unreliable results and conclusions from CTC studies. By combining these neoplasms under the umbrella of CTCs, masking of clinically useful information, like differences in prognoses or therapeutic response, may be occurring (Barber, 2007). In future studies of CTCs, it is critical to begin routinely distinguishing these neoplasms, as preliminary evidence supports genetic differences between canine FTCs and MTCs which may eventually lead to differences in treatment, as in human medicine (Al Rasheed and Xu, 2019; Bai et al., 2020; Cabanillas et al., 2019; Cabanillas et al., 2018; Campos et al., 2014c; Ceolin et al., 2019; Haddad et al., 2018; Hassan et al., 2020; Valerio et al., 2017; Varricchi et al., 2019).

Overall, most convoluted neural nets (CNNs) of the model are successful and are ready for use in a diagnostic setting in conjunction with interpretation by a pathologist and the continual addition of images with periodic re-training and re-validation of the model. This is supported by relatively low (good) average validator scores for CNN 1 (high quality tissue), CNN 2 (carcinoma and remnant), and only the follicular pattern of CNN 3. Furthermore, these scores trend similar to what is found for the total area error percentages during the final training (**Table 8**), with CNN 1 having the least area errors overall when considering CNNs 2 and 3 without subdividing them into their segmentation layers. Total area error is essentially the total error per training area, including both false positive and false negative areas. CNN 3 (follicular versus compact versus medullary) shows promising validation results, although the results from this CNN are less reliable and should be used with caution. A mixed layer for the third CNN was not included; instead, the model generates a tissue area percentage of each segmentation mask, so a pathologist could diagnose a mixed FTC after confirming an algorithm's output gives approximately equal tissue area percentages for the follicular and compact segmentation regions

or visually assessing the segmentation masks. This further supports the requirement that a pathologist interprets the model-generated results. Because there are no clearly defined guidelines in diagnosing a mixed FTC beyond approximately equal portions of follicular and compact patterns, their diagnosis remains subjective. The proportions could range from a pure 50-50 to a more skewed 60-40, etc., depending on the pathologist's interpretation and adherence to this definition as well as the area of mass examined (Kiupel et al., 2008; Rosol and Meuten, 2017; Rosol and Frone, 2016). While this is considered a minor point in this study, it is still notable and should be considered when formulating future classification schemes of CTCs, especially if AI models for the classification of tumors continue to be developed based on the amount of pattern present.

Looking at the overall average scores and standard deviation for each segmentation layer, it can be expected that CNNs 1 and 2 will function well when applied to future images, but CNN 3 will perform less reliably with most concern given to the compact and medullary layers. Validation scores for CNN 3 have an estimated accuracy of 80-95% for the model's predictions, while the scores for CNN 1 and 2 ranged from slightly higher to 100%. Therefore, CNN 3 still requires additional development, especially with providing additional compact FTCs and MTCs from distinctly different cases as compared to what was used so far.

Retrospectively, a more effective model structure could be that CNN 1 is for high quality tissue (as is here), CNN 2 is for neoplastic tissue versus non-neoplastic tissue (as is here), and CNN 3 is for all FTCs (non-subtyped) versus MTCs, while subsequent CNNs could assist in subtyping FTCs, counting mitotic figures, or even correlating patient outcomes or genotypes with histologic patterns (Castellino, 2005; Chan et al., 2020; Ching et al., 2018; Echle et al., 2021; Laury et al., 2021; Levine et al., 2019; Sultan et al., 2020; Tosun et al., 2020; Zuraw, 2020). The latter

design is specifically inspired by what Laury et al. (2021) describe, but in this case, after their preliminary segmentation layers were trained, the whole tissue was then encircled by a training annotation relating to patient outcome. As more AI models for diagnostic use become validated and reported in the literature, determining the appropriate model structure for a particular diagnostic challenge will become easier.

The ground truth for this current model could be improved by incorporating only those H&E-stained WSI whose diagnoses have been reviewed and agreed upon with consensus by a group of board-certified veterinary anatomic pathologists in conjunction with a robust IHC panel. This robust IHC panel should ideally include at least two FTC markers, 2 MTC markers, one or more NE markers, and parathyroid hormone which would allow for increased confidence in the diagnoses as well as the possible incorporation of more poorly differentiated CTCs (Campos et al., 2014b; Carver et al., 1995; Hassan et al., 2020; Soares et al., 2020; Moore et al., 1984).

Overall, the interpretation of the model's predictions is similar to the agreement between the verified IHC-based diagnosis and the results of IHC-blinded pathologist. However, agreement is poor when comparing the model to the IHC-blinded pathologists as a whole. The reasons for discordance between the model's function as compared to just IHC-blinded pathologist A versus the whole group of IHC-blinded pathologists are likely due to a combination of factors, including over-reliance on ancillary IHCs during routine diagnostic use, the unfamiliarity of subtle features that can be used to differentiate compact FTCs and MTCs, a decreased inclination to diagnose MTCs due to much of the literature suggesting a low prevalence, and/or differing years of experience. IHC-blinded pathologists tended to misclassify MTCs as compact FTCs more frequently than they would misclassify compact FTCs as MTCs, based on the associated confusion matrices comparing the blinded pathologists to the verified

diagnoses by IHC. In contrast, the model tended to misclassify FTCs as MTCs more than it would misclassify MTCs as FTCs. When considering FTCs as positive and MTCs as negative, this model tends to classify things as false negative more frequently than as false positive (lower sensitivity). For some models, specificity or sensitivity may be prioritized, depending on the desired outcome; however, in this case, since there are not yet obvious differences in prognoses, it is unclear whether sensitivity or specificity should be prioritized.

Overall, the agreement between visual assessment or segmentation area percentage with the verified cell of origin (FTC or MTC) was moderate or weak, respectively. While this may seem concerningly low, there is an inherent bias involved in the WSI used here, as these WSI were selected based on their original diagnosis of compact FTCs or MTCs. The low value can therefore be correlated with the validation results, in that the model has the most trouble differentiating between compact FTCs and MTCs. This lower agreement was therefore expected and is not reflective of the overall function of the model.

Both visually and according to the highest segmentation area percentage, only an overlapping subset of WSI were misclassified. In images with near equivalent segmentation area percentages, it can be inferred that the model likely struggled in differentiating between compact FTC and MTC patterns. The errors in model function for CNN 3, especially the medullary pattern, are likely associated with the high error rates found in CNN 3 during the final training.

A potential extrapolation by a human observer would be to consider the range in differences between the segmentation area percentages, as this may correlate with how well the model is functioning. Theoretically, MTCs should not contain areas of follicular or compact FTC growth. Conversely, FTCs should not contain areas of MTC growth. Canine neoplasms deriving from both follicular thyrocytes and medullary cells have not yet been reported and are considered

unlikely. In light of this, FTCs should have a high segmentation percentage for compact and/or follicular patterns with low to zero MTC percentages, and vice versa. WSIs with small differences in percentages separating the FTC subtype from MTC may suggest that the model struggled with correctly assigning segmentation masks, while those with larger percentage differences could suggest more reliable model-generated output. However, based on the percentages in this image subset, there is one MTC that was incorrectly classified as an FTC with a large difference in percent segmentation areas for compact FTC (85.34%) and MTC (14.0%), which highlights the need for further development. Because only 2 of the misclassified images (by visual and percentages) had a sister slide involved in the training sets, it appears that the model is more successful in classifying WSIs that are similar to the training set, which is an expected finding.

From the subset of WSI used in this component of the study, the segmentation masks generated for CNN 3 suggest that image K is a heterogeneous neoplasm and would be more appropriately interpreted as a follicular FTC, although this is incorrect based on the original verified diagnosis as well as a re-examination of the original slide and IHCs. On H&E, this neoplasm is relatively homogenous with lobules and sheets of polygonal cells that are occasionally vacuolated or pulled away from the basement membranes. As an example, correlating regions of the H&E-stained slide with the segmentation masks shows that the model is erroneously segmenting areas with vacuolated cells as the follicular pattern and areas where cells are pulling away from the basement membranes as MTC (**Figure 4**). However, attributing these features to how the algorithm is segmenting regions is speculative at best, since this model and most currently used AI models function within a black box (Levine et al., 2019; Tosun et al., 2020). There are also large multifocal regions with no overlying CNN 3 segmentation mask

(26.46%) which further supports that the model struggled to assign a segment those regions, although they were generally detected by CNN 2 as carcinoma tissue. This discordance represents a failure of the model to achieve segmentation, which can be rectified by further training. At this time, this is a minor error, as both compact and follicular patterns are still categorized under FTCs as a whole and are therefore not being misclassified as an MTC; however, this error should ideally be trained out in future models. This image did not have a sister slide within the training set.

Comparisons between the IHC-blinded pathologists and model generated predictions on a subset of compact FTCs and MTCs was pursued for two reasons. The first, tangible reason was to directly compare the function of the AI model to the IHC-blinded pathologists. For this, the model was found to generally be more successful than the group of IHC-blinded pathologists. Secondly, compact FTCs and MTCs were targeted to specifically highlight the flaws of this model in differentiating between compact FTCs and MTCs, as these were the two most unreliable layers. This unreliability was further showcased by several WSIs that were incorrectly interpreted by the AI-generated results which included both completely new images and fewer images that had sister slides involved in the training WSI set. While this data is skewed towards the comparison of compact FTCs and MTCs, it still appears to enable frequent correct interpretations. Continued development of CNN 3 is required to minimize incorrect classifications, as well as to ensure a robust capability of distinguishing all diagnostic classes considered here (follicular, mixed, or compact FTCs, and MTCs). Improvement for this particular comparison include having the same IHC-blinded pathologists interpret the model's output (make a diagnosis) and compare that to the verified diagnosis. Alternatively, a more desirable comparison would be having a new panel of pathologists interpret results and formulate

a diagnosis from this model's CNN 3 on WSIs from all diagnosis categories (follicular, mixed, or compact FTCs and MTCs) to compare to verified diagnoses and/or cell of origin.

Overall, expanding the training data set is highly recommended by incorporating H&E-stained WSIs of CTCs from outside institutions, from the UIUC VDL archives, and/or from new cases that presented to the UIUC VDL after the study enrollment period. Furthermore, should this model ever be deployed for routine use in a diagnostic lab, training images should be continually added with periodic re-training and re-validation of the model to maintain quality assurance and quality control and ensure the appropriate function of the model. An expanded example library for the model to learn from which should alleviate some of the previously described issues and errors.

Image matching could also improve the discriminatory ability of this model. This would entail pairing the H&E-stained WSI with the IHC-stained WSI and training a model to incorporate information from both; however, given the large variation in immunoreactivity for both thyroglobulin and calcitonin, this could be challenging to work out (Ma et al., 2021). The incorporation of multiple IHCs for each FTCs and MTCs, as outlined above, could be more helpful.

The inaccuracy of pathologists' diagnosis between compact FTC and MTCs solely by H&E-stained images was confirmed by the low agreement between pathologists blinded to IHCs as evaluated by Cohen's Kappa. This reaffirms the continued need for ancillary modalities for differentiation, whether it be ordering IHCs, utilizing a fully validated AI model, or other diagnostic modalities. This is the first study to explicitly evaluate and characterize this discordance, although this concept is already well-accepted in CTC literature (Carver et al., 1995; Pineyro et al., 2014; Rosol and Meuten, 2017; Rosol and Frone, 2016).

There are several instances of discordance in this project, including between validator pathologists' determination of cell of origin (FTC or MTC) and between validator pathologists' ultimate diagnosis (follicular, compact, or mixed FTC or MTCs). Discordance for these comparisons may derive from a combination of factors. Firstly, the validator pathologists could have been at a disadvantage in interpretation without access to all slides from a case. Secondly, as mentioned previously there is subjectivity in subtyping FTCs as well as differences in validator pathologist interpretations of IHC staining. Furthermore, interpretations of one neoplasm between several pathologists may differ slightly depending on the area of the neoplasm examined; some of these ideas can be extrapolated to explain the discordance in scoring as well. Discordance between FTC subtypes is considered acceptable for this project, given this inherent subjectivity and that the main goal was to distinguish between FTCs and MTCs.

Validator interpathologist agreement and agreement with the verified diagnosis are both supported by moderate to good agreement. Less agreement here could be due to comparing three observers (three pathologists) rather than two (one pathologist and JMA), which inherently introduces more variability. The differences in Fleiss' Kappa as compared to the Cohen's Kappa may be in part due to differences in subtyping FTCs as well as differences in the statistical tests used.

Regarding the validation scoring, improvements in the validation process could be made with more explicit instructions on how to score segmentation errors as well as decreasing the possible scores within the scoring system. Interpathologist differences in scoring could have arisen with differences in the approach of images with areas of the converse segmentation pattern. As an example, say an MTC was primarily segmented out with the medullary pattern, but it also contains regions of either follicular FTC or compact FTC patterns. For some, a small

amount of erroneous segmentation areas could have been considered within acceptable limits and given a better, lower score (a 1 or 2), while for others the presence of any erroneous segmentation areas could have been considered unacceptable, yielding a higher score (3 or 4). In this case, minimal guidance was given to the validator pathologists for this issue which surely contributed, at least in part, to differences in scoring. Aeffner et al. (2017) recommend minimizing the number of categories when scoring or grading samples to minimize the effects of the human tendency to avoid extremes. Therefore, a two- (acceptable or not acceptable) or three-tiered system (excellent, acceptable, or not acceptable) could more accurately represent the function of the model.

The ICC scores calculated for each layer were surprisingly low, even though the data for some layers appear relatively concordant, especially considering CNN 1. Although ICCs may be used to evaluate the agreement of quantitative data, like ratings, it is well known that ICC values may be spuriously low with low between-subject variance (Girard J, 2016). I suspect this is what is happening in this case and that another statistical test may be more appropriate. Enlisting the assistance of a statistician with an interest in artificial intelligence could aid in determining and/or confirming appropriate statistical tests.

An F1 score calculation was also attempted for each confusion matrix comparing validator pathologists' scoring for each segmentation layer. The F1 score is essentially a measure of accuracy, is commonly used in evaluating predictive performance in ML, and is defined as the harmonic mean of precision (of all positives, how many are true positives; measures the extent of the error caused by false positives) and recall (of all true positives, how many are predicted positive; measures the extent of the error caused by false negatives) (Mohajon, 2020; Zeya, 2021). A high F1 score indicates a well-performing model (Mohajon, 2020; Zeya, 2021). With

this data, calculation of an F1 score was not possible for each of these confusion matrices, as the mathematical formula would require dividing by zero, in some instances.

CTC literature generally maintains that while follicular, compact, and mixed FTC subtypes are all relatively common, the mixed or compact subtypes may be most common, depending on the study referenced (Campos et al., 2014c; Kiupel et al., 2008; Pessina et al., 2014; Rosol and Meuten, 2017; Rosol and Frone, 2016). Although the literature is inconsistent, this study suggests that compact and mixed FTCs may be more common than follicular FTCs (Campos et al., 2014c, Pessina et al., 2014). Another interesting finding is that the rate of MTCs considering unequivocal cases is 29.2%, which is similar to what is reported in several recent papers on CTCs and in direct contrast to several older, as well as more recent clinical papers (Campos et al., 2014a; Campos et al., 2014b; Campos et al., 2014c; Carver et al., 1995; Hassan et al., 2020; Kiupel et al., 2008; Liptak, 2007; Pessina et al., 2014; Pineyro et al., 2014; Ramos-Vara, 2002; Rosol and Meuten, 2017; Rosol and Frone, 2016; Soares et al., 2020). This difference in MTC prevalence (near 30% or less than 5%) could be a manifestation of the systemic lack of differentiating between FTCs and MTCs for each study and the perpetuation of data from older studies.

The numeric mean age correlates with the age group most represented here (7-10 years). These findings correlate with literature that maintains that advancing age is associated with an increased risk of developing CTCs and the dogs with CTCs tend to be between 9 and 10 years of age (Barber, 2007; Campos et al., 2014a; Hassan et al., 2020; Hayes and Fraumeni, 1975; Liptak, 2007; Rosol and Meuten, 2017; Rosol and Frone, 2016; Soares et al., 2020). The peak in age found here could be due to a combination of factors including an inherent peak age range for developing CTCs, that many dogs do not live to the 11- to 15-year age range due to natural

causes or humane euthanasia, or that owners of dogs in this latter age range may not seek advanced medical care for dogs nearing the end of their natural expected lifespan. A statistically significant relationship was found between age groups and developing all FTC subtypes or MTCs, but this relationship was not substantiated when FTCs were divided into subtypes. Accordingly, an ANOVA revealed a statistically significant difference in mean age between all FTCs or MTCs, while an ANOVA separating the FTC subtypes and MTCs did not reveal a statistically significant difference in mean age between at least two of these groups. Current literature does not provide mean ages for the different subtypes, as they are generally grouped together as FTCs. The lack of statistically significant relationships when subtyping FTCs in this study could support this grouping and the notion that FTC subtypes have similar biologic behaviors, as the canine literature currently suggests (Campos et al., 2014b). This latter point, however, is in contrast with what is found in human medicine and may be refuted if less-differentiated FTCs (compact or mixed) can be correlated with worse biologic behavior or outcome (Bai et al., 2020; Carver et al., 1995; Castillo et al., 2016). Examination of a study set from various institutions could help elucidate and/or confirm these findings. Discordance of statistical modalities evaluating all FTCs or MTCs versus subtyped FTCs or MTCs could be a function of the overall small sample size per category in the likelihood ratio test. Larger sample sizes will assist in further elucidating any significant relationship here.

Few articles specifically provide either a mean (9.6 years) or median age (9 years) for MTCs with a range of either 4 to 12, 13, or 16 years as compared to the 8.4±2.3 years found here; most articles group MTCs with FTCs yielding a mean age for CTCs as a whole (Carver et al., 1995; Patnaik and Lieberman, 1991). These results suggest that MTCs may affect slightly younger dogs. Additional studies are needed to confirm or refute these findings.

In this study, there does not appear to be any statistically significant associations between the sex, location, and breed compared to either all FTCs and MTCs or subtyped FTCs and MTCs. This is similar to what is currently described and accepted in canine CTC literature, as no sex or side predisposition (left versus right thyroid gland) are reported (Barber, 2007; Campos et al., 2014a; Hassan et al., 2020; Hayes and Fraumeni, 1975; Leav et al., 1976; Liptak, 2007; Patnaik and Lieberman, 1991; Pessina et al., 2014; Rosol and Meuten, 2017; Rosol and Frone, 2016; Soares et al., 2020). In contrast, there is one report of canine MTCs being more common in males, and in humans, it is well accepted that older human women tend to have a higher risk of thyroid cancer, while men tend to experience more aggressive cancer (Hassan et al., 2020; Hayes and Fraumeni, 1975; Patnaik and Lieberman, 1991). Human medicine also does not report a side predisposition (Hassan et al., 2020; Hayes and Fraumeni, 1975). Ectopic thyroid carcinomas at locations other than the neck appear to be a minority of cases both here and in the literature (Liptak, 2007; Rosol and Meuten, 2017).

There is a lack of consensus in the literature on predisposed breeds, but suggested breeds include boxers, beagles, Siberian huskies, golden retrievers, and mixed breed dogs (Hassan et al., 2020; Hayes and Fraumeni, 1975; Liptak, 2007; Rosol and Meuten, 2017; Rosol and Frone, 2016). This study's population contains all of these breeds, but Labrador retrievers and herding dogs (as categorized by the AKC) were the most common with 15 cases each (11% of the study population). Hassan et al. (2020) suggest that some breeds within the AKC herding group may have increased risk, including Shetland collies (shelties, Shetland sheepdog), old English sheepdogs, and Cairn terriers. The AKC herding group also includes breeds like the Australian cattle dog, Australian shepherd, bearded collie, border collie, and Welsh corgis (List of Breeds by Group, 2020). Because herding-type dogs were common here and another study has proposed

increased risk, there may be a true increased risk for these latter breeds, although examination of this relationship would require additional studies with a much larger sample population.

Desmoplasia is often associated with malignant carcinomas and is relatively nonspecific to the neoplastic cell of origin outside of carcinomas (Newkirk et al., 2017). Therefore, desmoplasia is a useful feature to look for when diagnosing CTCs but may not be useful in differentiating between all FTCs and MTCs or even between FTC subtypes. In this study, a statistically significant relationship with the presence of desmoplasia and all FTCs or MTCs and subtyped FTCs or MTCs was identified. However, the presence and amount of desmoplasia in CTCs have not been previously evaluated in the context of prognosis or grading and could be useful in the development of future prognostic or grading criteria.

Osseous metaplasia has been described in various literature sources with conflicting connotations and diagnostic terms. Some sources would name these neoplasms as malignant mixed thyroid tumors ("carcinosarcoma", "undifferentiated thyroid carcinoma of spindle cell type with osseous or cartilaginous metaplasia", by the WHO scheme) due to concurrent malignant thyroid follicular cells and mesenchymal elements (Kiupel et al., 2008; Ramos-Vara et al., 2002; Rosol and Meuten, 2017; Rosol and Frone, 2016). The WHO scheme specifies that the clonality of the mesenchymal cells has not been investigated, while Rosol and Meuten describe the neoplastic spindle cells present as resembling those found in chondrosarcoma and osteosarcoma, which could suggest neoplastic qualities (Kiupel et al., 2008; Rosol and Meuten, 2017). Another article reports the presence of woven bone with an MTC (Pineyro et al., 2014). In the current study, a statistically significant relationship was confirmed between the presence of osseous metaplasia and subtyped FTCs or MTCs, suggesting this influences the ultimate diagnosis. In this study, cases with osseous metaplasia generally did not contain a robust

spindloid neoplastic population, and the formed spaces tended to resemble normal bone marrow; this change was also identified in some MTCs. In most cases a diagnosis of malignant mixed thyroid tumor or carcinosarcoma did not seem appropriate, except for one case with an unequivocal diagnosis. This case contained a robust and highly infiltrative spindloid population as well as a distinct infiltrative epithelial population. The heterogeneity for naming neoplasms with malignant mesenchymal populations versus the presence of benign-appearing osseous metaplasia needs to be clearly defined and, ideally, with distinctive nomenclature to avoid confusion with the term "mixed FTCs". Larger sample sizes would assist in further elucidating any significant relationship here. At this time, the remaining categorical histologic features evaluated in this study do not appear to hold diagnostic significance.

ANOVAs of the mitotic figures between all FTCs and MTCs or between subtyped FTCs and MTCs showed no statistically significant differences, and do not appear to be useful for differentiation. Additionally, mitotic figures have not been successfully correlated to outcome values or therapeutic responsiveness (Campos et al., 2014b). Both Campos et al. (2014b) and Soares et al. (2020) evaluated the proliferative marker Ki-67 and found no significant differences between differentiated FTCs and MTC; Campos also found that at the time of diagnosis, Ki-67 was positively associated with local invasiveness and negatively associated with time to metastasis but was not considered an independent predictor. In humans, Ki-67 is associated with clinical stage and survival in both differentiated FTC and MTC and cutoffs have been suggested to correlate with the neoplasm's biologic behavior (Campos et al., 2014b). However, until there are more studies evaluating the role of mitotic figures in CTCs, mitotic figures should continue to be reported. This could include performing standardized mitotic counts, such as providing counts from a tissue area of 2.37 $mm^2$ as Donovan et al. propose (2021). It is possible that

correlating mitotic counts with metastatic rates, therapeutic responsiveness, or prognosis could provide clinically useful information.

Necrosis is another relatively nonspecific change that is often found with malignant neoplasms with discordant growth rates between the neoplastic cells and their vascular supply (Newkirk et al., 2017). ANOVAs of the mean score of necrosis between all FTCs and MTCs or between subtyped FTCs and MTCs showed no statistically significant differences groups.

Campos et al. (2014b) found that macroscopic (identification of tumor thrombi in cervical blood vessels) and histologic (tumor growth into blood vessels) vascular invasion were independent negative predictors for disease-free survival and corroborates an earlier study suggesting that vascular invasion is one of the most important histologic criteria for the overall grade of malignancy. The identification of intravascular invasion could be improved by serial sectioning of tissue blocks, as suggested by Soares et al. (2020). However, this is impractical for routine diagnostic use, and, based on the results here, does not suggest a significant association in differentiating these tumors. Therefore, continued vigilance for the identification of intravascular invasion and tumor emboli is recommended as it may be of clinical importance.

Most of the clinical findings (when provided) from this study appear consistent with what is found in the literature (e.g., dyspnea, vascular invasion, possible hypothyroidism, etc) (Lee et al., 2020; Liptak, 2007; Rosol and Meuten, 2017). Surveys or questionnaires given to referring veterinarians to standardize the clinical information received could assist in further determining the significance of those clinical findings that are not explicitly described in the literature (e.g., chylothorax) or appear superficially unrelated to CTCs (e.g., a concurrently excised soft tissue sarcoma).

A large limitation of the overall study is that a single resident (JMA) confirmed the initial diagnosis with only occasional assistance from a board-certified pathologist. This may have resulted in errors and/or unintended bias in the training WSI set, as well as for the secondary and tertiary objectives. Ideally, evaluation by consensus of the entire H&E-stained slides by a group (at least 3) of board-certified pathologists with or without residents could reduce these unintended biases or possible errors in diagnostic interpretation and improve the overall reliability and significance of these results.

Other sources of bias include that these cases all came from one, tertiary institution or from the inclusion of those cases that already had accompanying IHC stains. For the former, the study set could be skewed to represent patients with more severe or unusual diseases or that have owners more motivated to pursue medical treatment. For the latter, there may be some missed cases where IHCs were not pursued, based on the tissue patterns present on routine H&E-stained slides (e.g., a pure follicular FTC). To the author's knowledge, since one to two IHC stains are generally included in the cost of necropsy or surgical biopsy examination at the University of Illinois Urbana-Champaign Veterinary Diagnostic Laboratory (UIUC VDL), pathologists here tend to request at least either thyroglobulin (Tg), calcitonin, synaptophysin (SYP), or chromogranin A (CgA) as the standard of care for solid neoplasms that could represent either a compact FTC or MTC. Countering these biases would entail recruiting cases of CTCs from other institutions (from primary general practices to other tertiary referral hospitals or diagnostic labs) with tissue slide processing performed at both UIUC and other facilities. AI models tend to function better with highly variable training images; in this case, the most diverse images were selected to compensate for this inherent weakness, but this will be bolstered with the inclusion of slides from external institutions.

An additional obstacle encountered during model development was the identification of thyroid follicles and discrimination between remnant, non-neoplastic thyroid follicles versus those in the follicular FTC pattern. For the former, the CNN 2 of the model would misidentify cross-sections of skeletal myocytes as thyroid follicles which was resolved by training skeletal muscle as background rather than as non-neoplastic tissue. For the latter, neoplastic and non-neoplastic follicles can be remarkable similar, which translates to being a feature that is challenging for the model to accurately predict. Furthermore, because entrapped remnant thyroid follicles within MTCs are known to occur, any potentially entrapped follicles or medullary follicles were not specifically trained out. This could be a future avenue of investigation concurrent with image-based (morphological) profiling or image matching with IHC-stained WSI for this model.

A limitation for comparing the IHC-blinded pathologists to interpretations of the model's predictions is that interpretations included both a subjective visual assessment and an objective classification based on a model-generated percentage of segmentation tissue area. Specific issues with comparison includes that both interpretations were performed by one person (JMA), this person had prior access to the diagnosis of these slides, and humans are inherently prone to visual bias which can skew our interpretations (Aeffner et al., 2017). Therefore, the use of an objective, data-driven outcome is preferable, as it is quantifiable, repeatable, and less likely to suffer these biases. However, the use of only the percentage segmentation area may also present diagnostic issues, as outlined below.

A potential source of bias for the demographic and histologic analyses could be from excluding the single dog with bilateral FTC with differing contralateral subtypes; the right was mixed while the left was compact. Without the capability to thoroughly review the gross tissues

and how these neoplastic tissues were architecturally and anatomically related to one another, a diagnosis of bilateral FTC with differing contralateral subtypes was used. However, this may represent either a large mixed or compact thyroid tumor, with sectioning through an area that was predominately the opposite subtype. As a result, this animal was excluded from the statistical analyses. The effect of excluding this animal is likely small, given the number of cases available.

Two potential future studies would be applying this specific model to the 18 withheld cases with equivocal diagnoses as well as evaluation of this model when it is applied to WSIs of CTCs from outside of the UIUC VDL system. For both, ideally, at least three pathologists should be enlisted with one person responsible for one stage of model development, including the design and training of the model, quality control, and verification of the model's output, as outlined by Zuraw et al. (2020).

Because of the challenge in accurately annotating mitotic figures and the model identifying an excessively high rate of false positives during training (mainly due to artifact, necrosis, or inadequate fixation), a potential future workaround could be re-staining the same H&E slides which contain mitotic figures with phosphohistone H3 (PHH3) and correlating the WSI of the H&E-stained and PHH3-stained slides using image matching (Ma et al., 2021; Tellez et al., 2018). PHH3 is a relatively new IHC that can identify cells that are undergoing mitosis and can therefore yield easily identifiable, high-contrast mitotic figures (**Supplementary Figure 4**) (Donovan et al., 2021; Tellez et al., 2018). Other proliferative markers, like Ki-67, AgNOR, or thymidine kinase 1, could be used but do not appear to highlight the actual mitotic figures like PHH3 appears to do (Ramos-Vara and Borst, 2017; Wang et al., 2017). Alternatively, transfer

learning could be utilized, which entails incorporating a pre-existing working mitotic figure detector CNN from another model and fine-tuning it into this model.

The diagnostic challenge of veterinary pathologists reliably differentiating between compact FTCs and MTCs without ancillary testing, such as IHCs, is highlighted in this study based on low measures of agreement which supports their continued need. Another significant finding is that current CTC studies often fail to discriminate between FTCs and MTCs. This study suggests that a supervised segmentation deep learning model could be a novel and potentially more cost-effective way to rapidly distinguish between canine FTCs and MTCs, although more development is needed. Based on the validation data from the present study, most layers (CNN 1: high quality tissue, CNN 2: carcinoma versus remnant) are ready for use in a diagnostic setting in conjunction with interpretation by a pathologist and the continual addition of images with periodic re-training and re-validation of the model. Caution is still recommended with the use of CNN 3 (follicular FTC pattern versus compact FTC pattern versus MTC pattern), especially for differentiating compact FTCs from MTCs. Further baseline development of the model could include adding diverse training images from a multitude of sources and/or incorporating additional data (e.g., image matching with IHCs and/or correlating with other biomarkers or clinical data), while ancillary features could include mitotic figure detection or predictions about response to treatment modalities or outcome.

**Figures and Tables**

**Figure 1.** Representative photomicrographs of canine medullary thyroid carcinoma (A, left) and compact follicular thyroid carcinoma (B, right) illustrate the similar histomorphology contributing to challenges in determining cell of origin in canine thyroid carcinoma (H&E, 20x).

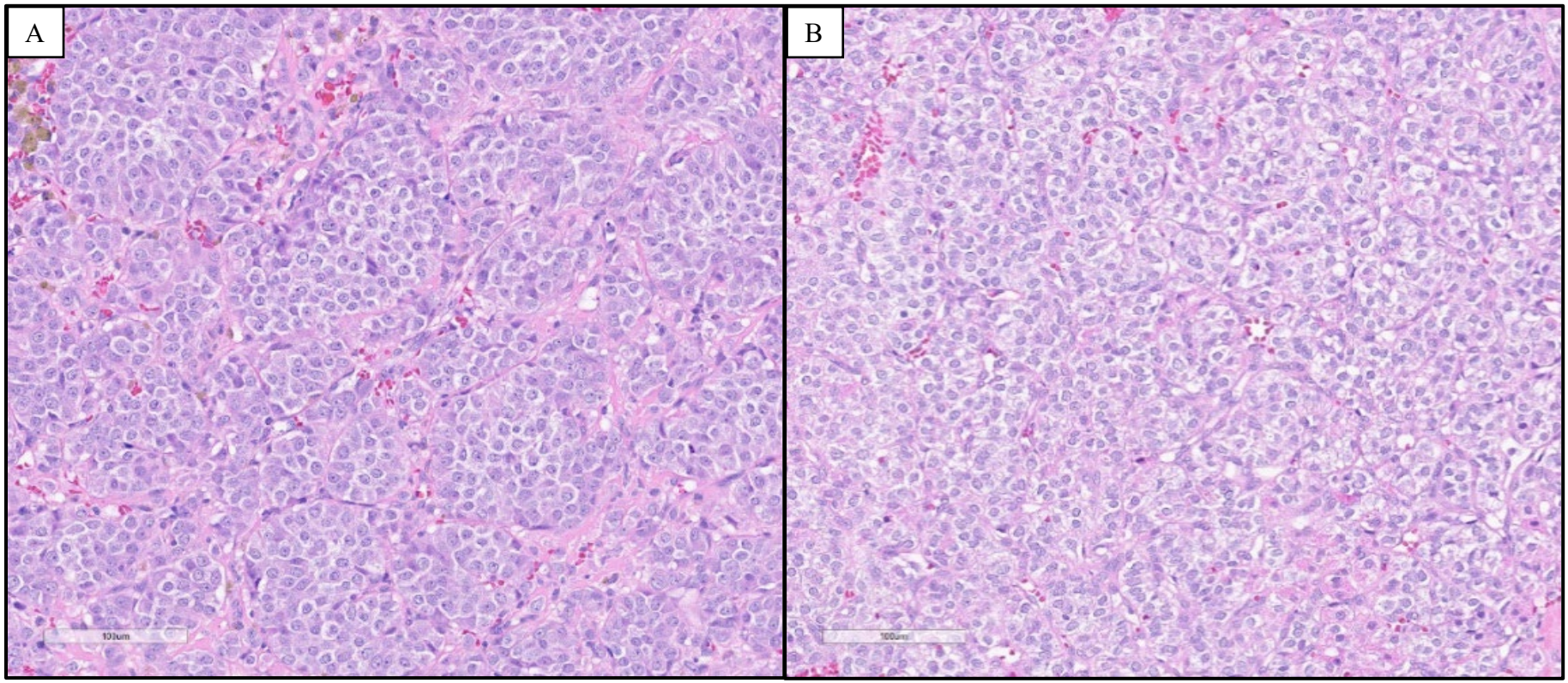


**Figure 2**. Artificial intelligence model structure and training for determining cell of origin (FTCs or MTCs) in CTCs. A) The overall structure of the AI model in schematic form. Of note, the "mitotic figures" layers are connected between the three "subtype" layers ("medullary", "compact", and "follicular") so that identification of mitotic figures is universal and not restricted by any neoplastic pattern. Hematoxylin and eosin (H&E). Scale bar = 1 mm. B) Representative annotations for the "high quality tissue" layer, which included any tissue present on the slide, despite the misnomer. The areas circled in green (annotations) indicate these areas are tissue regions (training annotations). The regions encircled in black are what is what the model "sees" to train on (training regions). The black circle without an outer green circle (asterisk) indicates this region should be interpreted as background and thus not included in subsequent analyses. H&E. Scale bar = 1 mm.

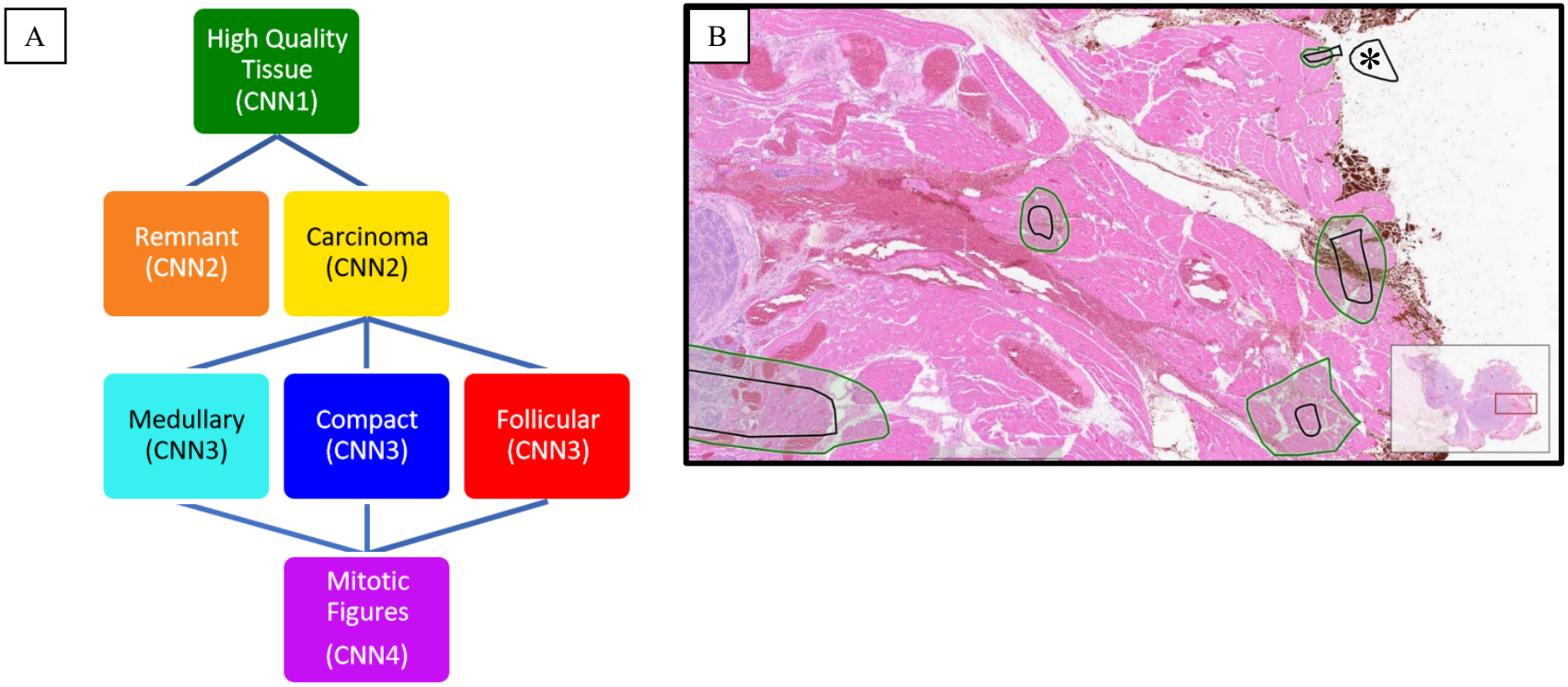


**Figure 2 (continued).** C) Representative annotations for the "remnant" and "carcinoma" layers. Neoplastic tissue is encircled in yellow, non-neoplastic tissue is encircled in orange, and skeletal muscle is only encircled by black (rendering it to be interpreted as background at this layer). H&E. Scale bar = 1 mm. D-F) Representative annotations for the "medullary" (D), "compact" (E), and "follicular" (F) layers; IHCs were used to determine whether to annotate as an MTC or FTC. H&E. Scale bar = 100 um.

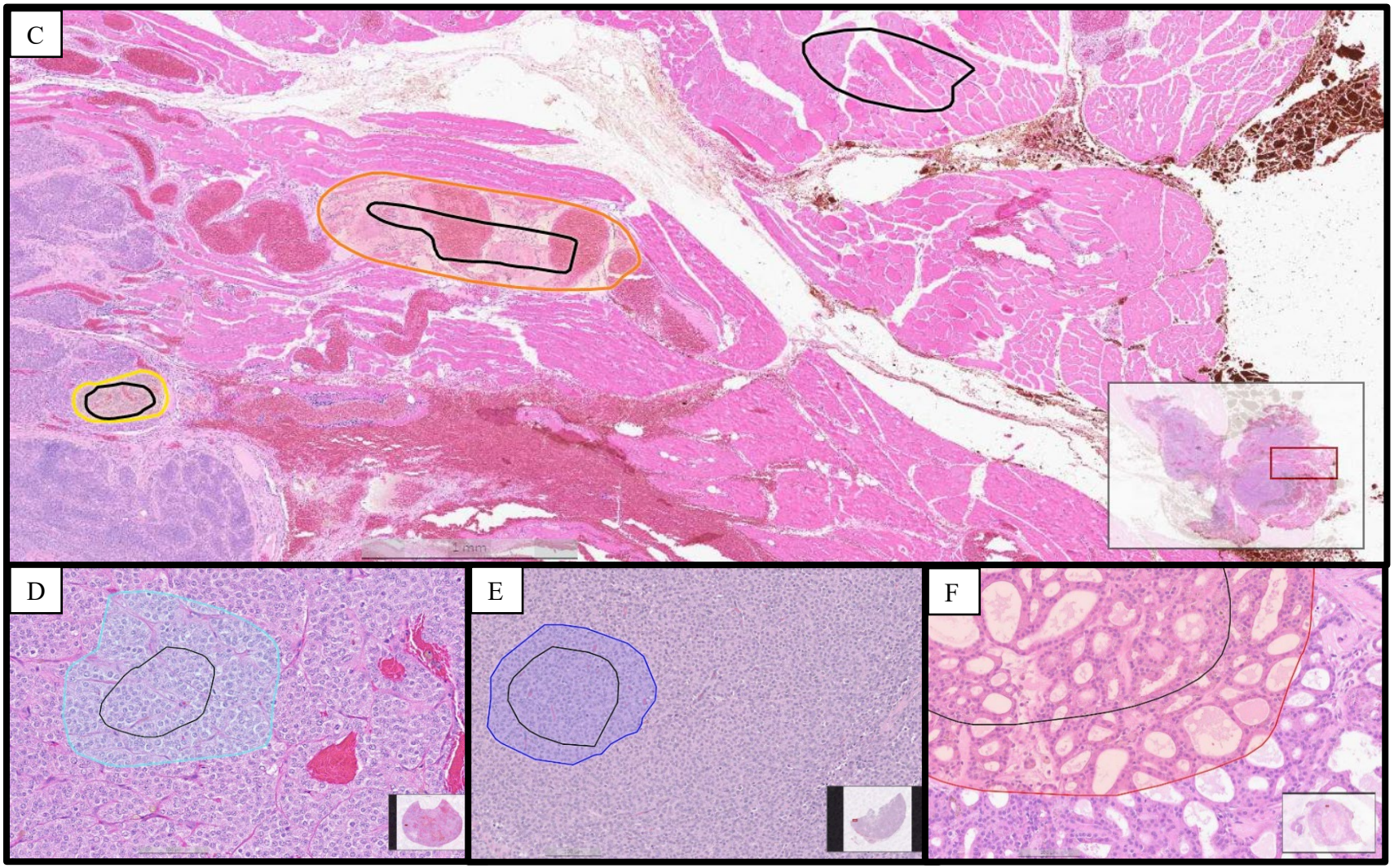


**Figure 2 (continued).** G) Representative image segmentation masks from the development period illustrating output for all three segmentation CNNs. This image is from the single case with a bilateral FTC differing contralateral diagnoses; in this case, the tissue piece on the left should be interpreted as a mixed FTC with some error, while the tissue piece on the right should be interpreted as a compact FTC.

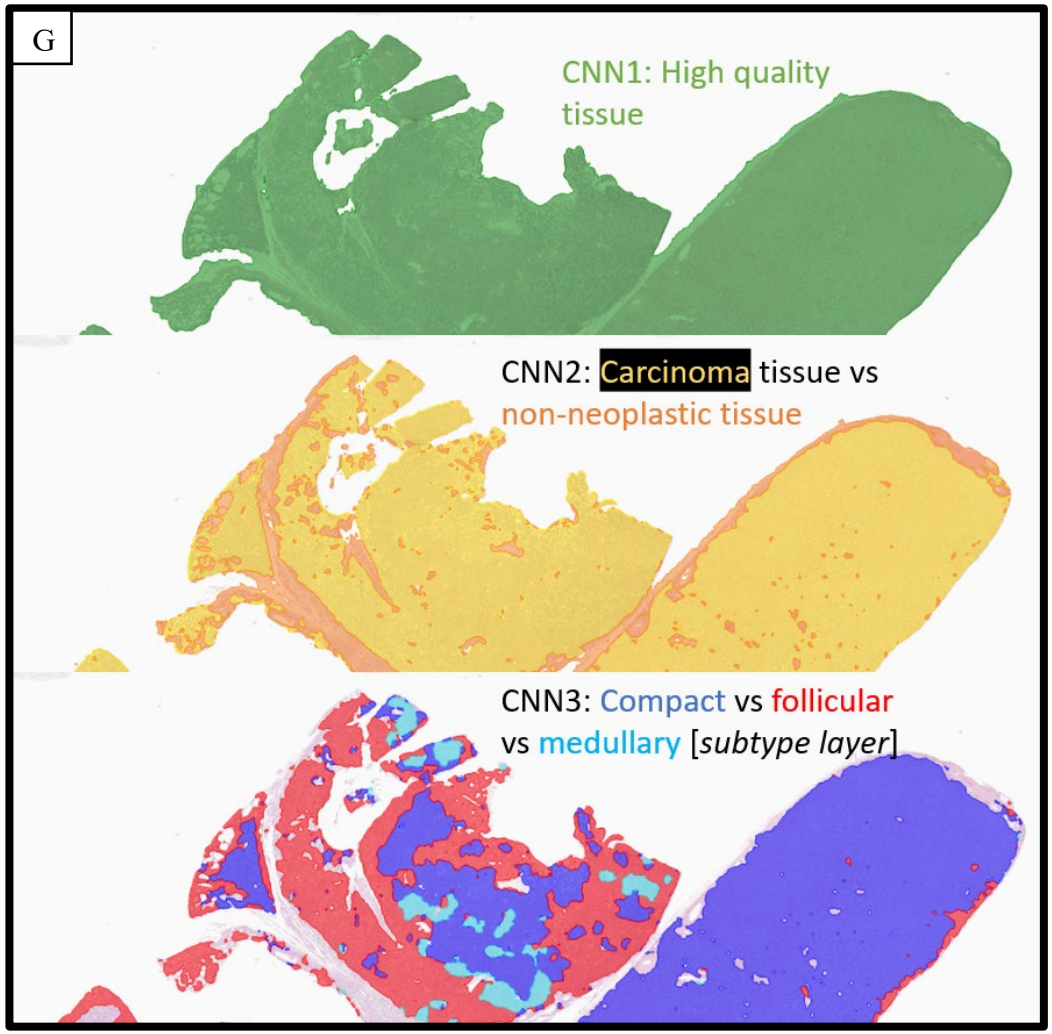


**Figure 2 (continued).** H-J) Representative annotations made for CNN 4, the “mitotic figures” layer. H&E. Scale bar = 20 um.

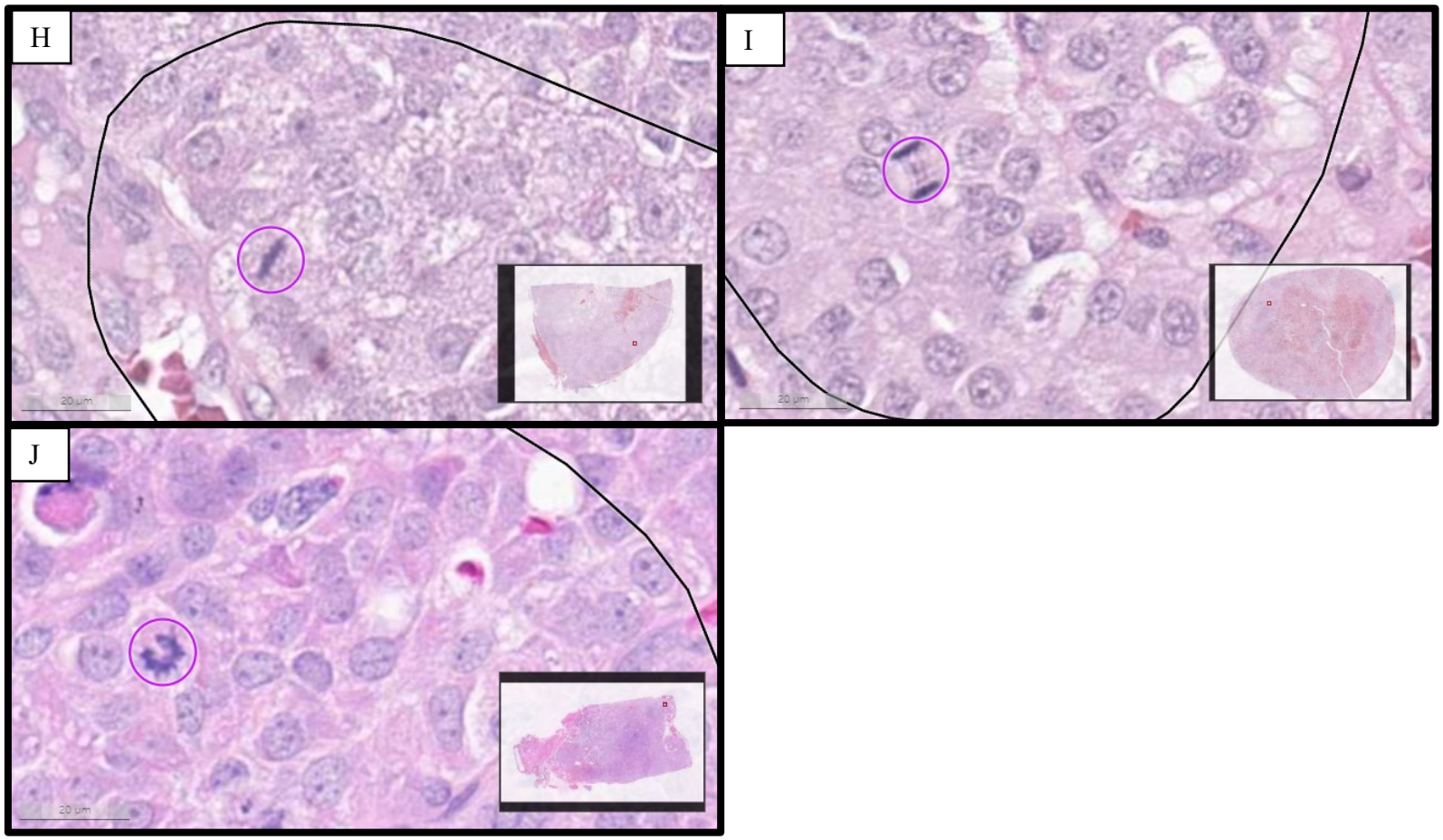


**Figure 3,** Macroscopic view of the CNN 3 segmentation masks from the image set used in the third objective. Notably, image K appears to be a follicular FTC with 43.7% segmented as follicular FTC pattern, 15.27% as compact FTC, and 14.57 for MTC. The image was originally verified and re-verified as a compact FTC.

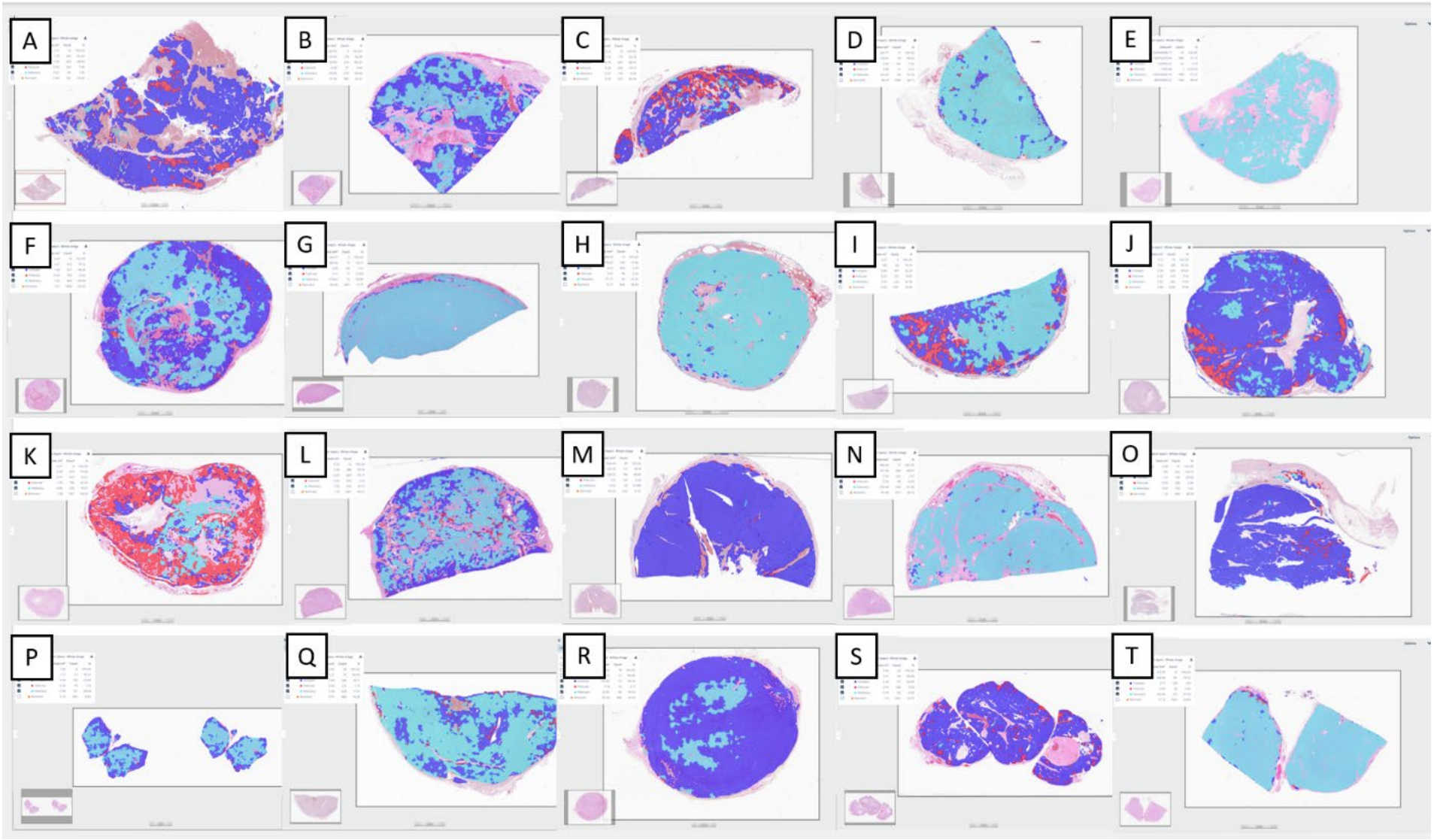
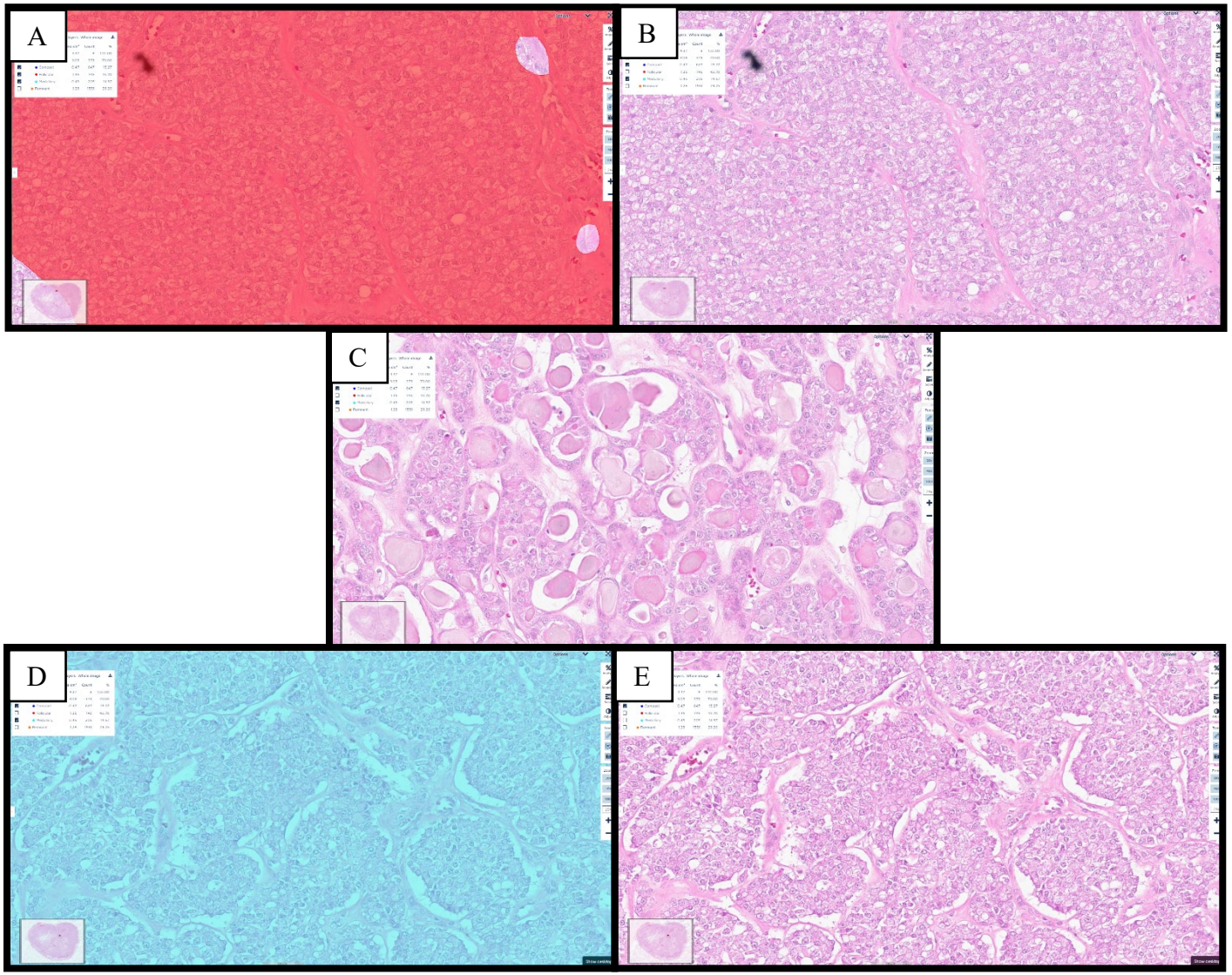


**Figure 4,** Representative regions of Image K from the subset of images used in the third objective. A-B) An area incorrectly segmented as follicular FTC. C) An area correctly segmented a follicular FTC with numerous colloid follicles. D-E) An area incorrectly segmented as MTC. All images except for C more appropriately resemble a compact FTC.



| **Table 1. Primary antibodies used for immune characterization of canine thyroid carcinomas.** | | | | |
|---|---|---|---|---|
| Antibody | Species/Type | Company | Dilution | Positive Control |
| Thyroglobulin | RP | Dako | 1:2000 | Canine Thyroid |
| Calcitonin | RP | Biocare Medical | 1:300 | Canine Thyroid |
| Chromogranin A | RP | ImmunoStar | 1:2000 | Canine Adrenal Gland, Pancreas, +/- Thyroid |
| Synaptophysin | MM | Biocare Medical | 1:100 | Canine Adrenal Gland, Pancreas, +/- Thyroid |
| *Abbreviations: MM, Mouse monoclonal antibody; RP, rabbit polyclonal antibody.* | | | | |

| **Table 2. Ideal discriminatory diagnostic features for well-differentiated follicular canine thyroid neoplasms. Adapted from the WHO classification scheme for canine thyroid neoplasms.** | | | | | |
|---|---|---|---|---|---|
| | *H&E* | *Immunohistochemistry* | | | |
| *Type* | *Features* | *Thyroglobulin* | *Calcitonin* | *Chromogranin A* | *Synaptophysin* |
| Follicular | The majority of tumor forms variably sized follicles with colloid that may be unremarkable, clumped, or mineralized | Positive | Negative | Negative | Negative |
| Compact | The majority forms solid sheets of aggregated cells | Positive | Negative | Negative | Negative |
| Mixed | Approximately equal proportions of follicular and compact growth<br>Colloid follicles may be smaller and contain less colloid | Positive | Negative | Negative | Negative |
| Medullary | Solid neoplasm growth with typical neuroendocrine packeting<br>+/- amyloid (considered rare in dogs) | Negative | Positive | Positive | Positive |
| Positivity to thyroglobulin: moderate to strong, 10% to 100% immunoreactivity of cytoplasm, apical membranes, intracytoplasmic droplets (if those cells form colloid follicles), and/or colloid<br>Positivity to calcitonin: faint to moderate, scant to diffuse granular cytoplasmic immunoreactivity, possibly of individualized, clustered, or the majority of cells<br>Positivity to chromogranin A and synaptophysin: faint to strong granular cytoplasmic positivity in the majority of cells<br>(Kiupel et al., 2008; Moore et al., 1984; Pineyro et al., 2014) | | | | | |

| **Table 3. Summarized convoluted neural network training with pre-training and post-training advanced parameters and verification error rates** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| CNN | Field of View | Complexity | Misc. Adjusted Advanced Parameters | Total Iterations | Training Loss | Total Area Detected as Error | Total Area Annotated | Post-training differences |
| CNN 1: High Quality Tissue | 225 um | Complex | Image Analysis: Region Merging Starting Level of 100<br>Heatmap starting level: 16<br>Region merging starting level: 16 | 10,000 | 0.0002 | 1.316 $mm^2$ | 493.15 $mm^2$ | Heatmap starting level: 200<br>Region merging starting level: 100 |
| CNN 2: Carcinoma versus Remnant tissue | 150 um | Extra Complex | Training Procedure: Mini-batch size of 20<br>Image Augmentation: Scale -15 to 15, Aspect Ratio 15, Maximum Shear 15, Luminance -15 to 15, Contrast -15 to 15, Maximum White Balance Change 3, Noise 2 levels<br>Image Analysis: Region Merging Starting Level of 50 | 15,000i | 0.0027 | 5.27 $mm^2$ | 167.534 $mm^2$ | No differences |
| CNN 3: Subtypes (follicular, compact, or medullary) | 400 um | Extra Complex | Training Procedure: Mini-batch size of 20 | 15,000i | 0.0031 | 12.814 $mm^2$ | 289.166 $mm^2$ | No differences |

**Table 3. Summarized convoluted neural network training with pre-training and post-training advanced parameters and verification error rates, continued**

| CNN | Total Area Error | Area Error | False Positive | False Negative | Precision | Sensitivity | F1 score | Error % (FP/FN) |
|---|---|---|---|---|---|---|---|---|
| CNN 1: High Quality Tissue | 0.27% | 0.27% | 0.21% | 0.06% | 99.63% | 99.90% | 99.77% | 0.47% (0.37/0.10) |
| CNN 2: Carcinoma versus Remnant tissue (combined) | 3.15% | 1.57% | 0.67% | 0.91% | 98.49% | 97.95% | 98.22% | |
| CNN 3: Subtypes (combined) | 4.43% | 1.48% | 0.12% | 1.36% | 99.60% | 95.69% | 97.61% | |

*Selected Definitions*

-Total Area Error: the sum of all false positive and false negative areas / sum of total regions of interest areas; this is essentially the total error per training area.

-Area Error: the difference between analysis result and annotations, calculated of training region area

-Precision: the percentage of analysis result area found within the annotation area

-Sensitivity: the percentage of annotation area that was found by the analysis

-F1 score: the harmonic mean of precision and sensitivity

-Error % (FP/FN) is the sum of all false positive and or false negative areas for the respective class / sum of all region of interest areas; in other words, this is the sum of all errors in all training regions containing the class

-FP is the false positive as a percentage of the whole area of all annotations per class or per whole layer

-FN is the false negative as a percentage of the whole area of all annotations per class or per whole layer

| **Table 3. Summarized convoluted neural network training with pre-training and post-training advanced parameters and verification error rates, continued** | | | | | |
|---|---|---|---|---|---|
| CNN | Area Error | Precision | Sensitivity | F1 score | Error (FP/FN) % |
| CNN 2: Carcinoma Only | 1.32% | 97.29% | 98.86% | 98.07% | 3.89% (2.75/1.14) |
| CNN 2: Remnant Only | 1.83% | 99.26% | 97.39% | 98.31% | 3.34% (0.73/2.61) |
| CNN 3: Follicular Only | 0.73% | 98.42% | 95.30% | 96.84% | 6.23% (1.53/4.701) |
| CNN 3: Compact Only | 0.75% | 99.37% | 97.82% | 98.59% | 2.80% (0.62/2.18) |
| CNN 3: Medullary Only | 2.97% | 99.97% | 94.76% | 97.29% | 5.27% (0.03/5.24) |

| **Table 4. Interobserver agreement between validator pathologists for determining the cell of origin (FTC or MTC)** | | | |
|---|---|---|---|
| Pathologist Pairs | A x B | A x C | B x C |
| Accuracy (proportion of agreeing predictions) | 0.92 | 0.96 | 0.96 |
| Mean Accuracy | 0.95 | | |
| Cohen's Kappa: | 0.84 | 0.92 | 0.92 |
| Cohen's Kappa Interpretation: | Good Agreement | Excellent Agreement | Excellent Agreement |
| Mean Cohen's Kappa: | 0.89 | | |
| Mean Cohen's Kappa Interpretation: | Excellent Agreement | | |

| **Table 5. Agreement between IHC-blinded pathologists and IHC-based diagnosis** | | | |
|---|---|---|---|
| IHC-Blinded Pathologist | A | B | C |
| Accuracy (proportion of correct predictions) | 0.8 | 0.6 | 0.55 |
| Mean Accuracy | 0.65 | | |
| Cohen's Kappa: | 0.60 | 0.20 | 0.10 |
| Cohen's Kappa Interpretation: | Moderate Agreement | Poor Agreement | Poor Agreement |
| Mean Cohen's Kappa | 0.30 | | |
| Mean Cohen's Kappa Interpretation: | Poor Agreement | | |

| **Table 6. IHC-blinded interpathologist agreement** | | | |
|---|---|---|---|
| Pathologist Pairs | A x B | A x C | B x C |
| Accuracy (proportion of agreeing predictions) | 0.70 | 0.55 | 0.65 |
| Mean Accuracy | 0.63 | | |
| Cohen's Kappa: | 0.41 | -0.05 | 0.03 |
| Cohen's Kappa Interpretation: | Weak Agreement | Poor Agreement | Poor Agreement |
| Mean Cohen's Kappa: | 0.13 | | |
| Mean Cohen's Kappa Interpretation: | Poor Agreement | | |

**Table 7, Comparison of the AI model's function (as interpreted by JMA) to IHC-blinded pathologists.** Cells filled with red indicate a difference with the verified diagnosis. The one cell surrounded by an orange box indicates the one FTC from this image set where the AI results suggest a diagnosis of follicular FTC.

Comparison of blinded pathologist diagnosis to interpretation of the model results by JMA

| | Path A | Path B | Path C | By Visual | By % | Highest Percentage | Original |
|---|---|---|---|---|---|---|---|
| A | FTC | MTC | FTC | FTC | FTC | Compact - 89.65% | FTC |
| B | MTC | MTC | MTC | MTC | FTC | Compact - 49.2% | MTC |
| C | FTC | FTC | FTC | FTC | FTC | Compact - 66.01% | FTC |
| D | FTC | FTC | FTC | MTC | MTC | Medullary - 91.93% | MTC |
| E | MTC | FTC | FTC | MTC | MTC | Medullary - 97.37% | MTC |
| F | FTC | MTC | FTC | MTC | MTC | Medullary - 49.94% | FTC |
| G | FTC | FTC | FTC | MTC | MTC | Medullary - 98.84% | MTC |
| H | MTC | MTC | FTC | MTC | MTC | Medullary - 97.22% | MTC |
| I | FTC | FTC | MTC | MTC | MTC | Medullary - 52.34% | MTC |
| J | FTC | FTC | FTC | FTC | FTC | Compact - 80.67% | FTC |
| K | FTC | FTC | FTC | FTC | FTC | Follicular - 43.7% | FTC |
| L | MTC | MTC | FTC | MTC | MTC | Medullary - 61.12% | MTC |
| M | FTC | FTC | MTC | FTC | FTC | Compact - 99.45% | FTC |
| N | MTC | FTC | FTC | MTC | MTC | Medullary - 97.59% | MTC |
| O | FTC | FTC | FTC | FTC | FTC | Compact - 95.85% | FTC |
| P | MTC | FTC | FTC | MTC | MTC | Medullary - 58.09% | FTC |
| Q | FTC | FTC | FTC | MTC | MTC | Medullary - 71.31% | FTC |
| R | MTC | MTC | FTC | FTC | FTC | Compact - 85.34% | MTC |
| S | FTC | FTC | FTC | FTC | FTC | Compact - 93.54% | FTC |
| T | MTC | FTC | FTC | MTC | MTC | Medullary - 97.56% | MTC |
| Total Correct | 16 | 12 | 11 | 16 | 15 | | |
| Total Correct as % | 80.00% | 60.00% | 55.00% | 80.00% | 75.00% | | |

**Table 8. Validator Pathologist Scores and Averaged Validator Pathologist Scores, with Averages, Variance, and Standard Deviation per Segmentation Layer.** Columns for images B and R are outlined in red, as they are the images with discordance among validator pathologists in the interpretation of the cell of origin (FTC or MTC).

Validator Pathologist A

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Ave. | Var. | Std. Dev. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HQT | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.12 | 0.11 | 0.33 |
| Carcinoma | 2 | 3 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 3 | 3 | 2 | 2 | 4 | 3 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2.32 | 0.31 | 0.56 |
| Remnant | 3 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 2 | 2.24 | 0.19 | 0.44 |
| Follicular | 3 | 4 | 3 | 3 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 2 | 2 | 2 | 2 | 3 | 1 | 2 | 3 | 4 | 2.64 | 0.57 | 0.76 |
| Compact | 3 | 4 | 4 | 2 | 4 | 3 | 2 | 2 | 2 | 2 | 4 | 4 | 2 | 4 | 2 | 2 | 2 | 4 | 2 | 2 | 4 | 3 | 2 | 3 | 4 | 2.88 | 0.86 | 0.93 |
| Medullary | 2 | 4 | 4 | 2 | 4 | 2 | 2 | 2 | 3 | 2 | 4 | 4 | 2 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 3 | 2 | 3 | 2.72 | 0.71 | 0.84 |

Validator Pathologist B

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Ave. | Var. | Std. Dev. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HQT | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.00 | 0.00 | 0.00 |
| Carcinoma | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1.08 | 0.08 | 0.28 |
| Remnant | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1.12 | 0.19 | 0.44 |
| Follicular | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 2 | 2 | 1.24 | 0.27 | 0.52 |
| Compact | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1.40 | 0.42 | 0.65 |
| Medullary | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 3 | 3 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 1.64 | 0.41 | 0.64 |

Validator Pathologist C

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Ave. | Var. | Std. Dev. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HQT | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 1.20 | 0.42 | 0.65 |
| Carcinoma | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.04 | 0.04 | 0.20 |
| Remnant | 1 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.20 | 0.17 | 0.41 |
| Follicular | 1 | 1 | 4 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 4 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 4 | 1 | 1 | 1 | 1 | 1 | 1.60 | 1.00 | 1.00 |
| Compact | 4 | 2 | 4 | 1 | 1 | 4 | 1 | 2 | 2 | 2 | 3 | 2 | 1 | 1 | 1 | 1 | 2 | 4 | 1 | 4 | 4 | 1 | 2 | 1 | 4 | 2.20 | 1.58 | 1.26 |
| Medullary | 2 | 4 | 4 | 1 | 4 | 1 | 1 | 2 | 4 | 1 | 3 | 3 | 1 | 1 | 1 | 2 | 4 | 2 | 4 | 4 | 1 | 1 | 2 | 1 | 3 | 2.28 | 1.63 | 1.28 |

Averaged Validator Pathologist Scores

| | | | | | | | | | | | | | | | | | | | | | | | | | | All Validation Images | | | Excuding Discordant Images | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Ave. | Var. | Std. Dev. | Ave. | Var. | Std. Dev. |
| HQT | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.67 | 1.00 | 1.00 | 1.00 | 1.00 | 1.67 | 1.00 | 1.33 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 2.00 | 1.00 | 1.00 | 1.11 | 0.07 | 0.27 | 1.12 | 0.08 | 0.28 |
| Carcinoma | 1.33 | 1.67 | 1.33 | 1.33 | 1.33 | 1.33 | 1.67 | 1.33 | 1.33 | 1.67 | 2.33 | 1.33 | 1.33 | 2.00 | 1.67 | 1.33 | 1.33 | 1.33 | 1.33 | 1.67 | 1.67 | 1.33 | 1.33 | 1.33 | 1.33 | 1.48 | 0.07 | 0.26 | 1.48 | 0.07 | 0.26 |
| Remnant | 1.67 | 1.33 | 1.67 | 2.00 | 1.33 | 1.67 | 1.33 | 1.67 | 1.33 | 1.67 | 2.00 | 1.33 | 1.33 | 1.67 | 1.33 | 1.33 | 2.00 | 1.33 | 1.33 | 1.33 | 1.67 | 1.33 | 1.33 | 1.67 | 1.33 | 1.52 | 0.06 | 0.24 | 1.51 | 0.05 | 0.22 |
| Follicular | 1.67 | 2.33 | 2.67 | 2.00 | 1.33 | 1.33 | 1.33 | 1.67 | 2.33 | 2.00 | 2.00 | 3.00 | 1.33 | 1.33 | 1.67 | 2.00 | 1.67 | 1.33 | 1.33 | 3.00 | 1.67 | 1.00 | 1.33 | 2.00 | 2.33 | 1.83 | 0.30 | 0.55 | 1.81 | 0.31 | 0.56 |
| Compact | 3.33 | 2.33 | 3.00 | 1.33 | 2.00 | 3.00 | 1.33 | 2.00 | 2.00 | 1.67 | 3.00 | 2.33 | 1.33 | 2.00 | 1.33 | 1.33 | 1.67 | 3.67 | 1.33 | 2.67 | 3.33 | 1.67 | 1.67 | 1.67 | 3.00 | 2.16 | 0.57 | 0.75 | 2.17 | 0.60 | 0.78 |
| Medullary | 2.00 | 3.33 | 3.33 | 1.33 | 3.33 | 1.33 | 1.33 | 1.67 | 3.00 | 1.33 | 3.00 | 3.00 | 1.33 | 1.00 | 2.00 | 2.00 | 3.33 | 2.67 | 3.00 | 3.00 | 1.67 | 1.33 | 2.33 | 1.33 | 2.33 | 2.21 | 0.67 | 0.82 | 2.12 | 0.60 | 0.78 |

| Table 9. Signalment and microscopic diagnosis of dogs with thyroid carcinoma. | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Animals (n=137)** | | | **Diagnoses (n=138)** | | | | |
| | | | **FTC (n=85)[1]** | | | **MTC[1]** | **Unsure[1]** |
| **Sex** | **#** | **%** | **Fol (n=7)** | **Com (=51)** | **Mix (n=27)** | **Med (n=35)** | **Uns (n=18)** |
| Male | 70 | (51) | 3 | 28 | 11 | 19 | 9 |
| MI | 4 | (3) | — | 2 | 1 | 1 | 1 |
| MC | 65 | (47) | 3 | 26 | 10 | 18 | 8 |
| M-NS | 1 | (1) | — | 1 | — | — | — |
| Female | 63 | (46) | 4 | 21 | 15 | 15 | 9 |
| FI | 2 | (1) | — | 1 | 1 | — | — |
| FS | 61 | (45) | 4 | 20 | 14 | 15 | 9 |
| F-NS | 0 | (0) | — | — | — | — | — |
| Not specified | 4 | (3) | — | 2 | 1 | 1 | — |
| **Age, years (mean [SD])** | 9.2 | [2.3] | 9.6 [1.6] | 9.8 [2.2] | 9.0 [2.5] | 8.4 [2.3] | 9.3 [2.3] |
| **Age, years** | | | | | | | |
| 3-6 | 18 | (13) | — | 3 | 3 | 9 | 3 |
| 7-10 | 67 | (49) | 4 | 23 | 16 | 16 | 8 |
| 11-15 | 47 | (34) | 3 | 23 | 7 | 8 | 7 |
| NS | 5 | (4) | — | 2 | 1 | 2 | — |

[1] Diagnoses were made by a single pathologist reviewing all tumors with accompanying immunohistochemical stains. An additional pathologist was consulted for a few equivocal cases.
[2]Cases with signalment such as "boxer-mix" are included with the emphasized breed. If multiple breeds were provided, cases were classified using the first listed breed. This applies to the beagle, boxer, golden retriever, and Labrador retriever strata. The remaining strata are based on American Kennel Club breed classifications (2020).
[3]Includes masses from the cranial mediastinum, heart base, pericardium, and subcutaneous mass near the thoracic inlet (ectopic neoplasms).

One dog with bilateral FTC had differing contralateral subtypes. This dog was a 15-year-old, female spayed, herding dog. This yields 138 total diagnoses with only 137 animals. This animal was excluded from mean age and standard deviation calculations. The remaining dogs with bilateral disease were found to have the same subtypes between the left and right sides and were therefore considered one case each.

Abbreviations: FTC, Follicular Thyroid Carcinoma; MTC, Medullary Thyroid Carcinoma; Fol, Follicular; Com, Compact; Mix, Mixed; Med, Medullary; Uns, Unsure; NS, not specified; MI, intact male; MC, castrated male; FI, intact female; FS, spayed female.

| Table 9, continued. Signalment and microscopic diagnosis of dogs with thyroid carcinoma. | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Animals (n=137)** | | | **Diagnoses (n=138)** | | | | |
| | | | **FTC (n=85)[1]** | | | **MTC[1]** | **Unsure[1]** |
| **Breed[2]** | **#** | **%** | **Fol (n=7)** | **Com (=51)** | **Mix (n=27)** | **Med (n=35)** | **Uns (n=18)** |
| Beagles | 10 | (7) | — | 4 | 4 | 1 | 1 |
| Boxers | 9 | (7) | 1 | 1 | 3 | 1 | 3 |
| Golden Retrievers | 8 | (6) | 1 | 3 | 1 | 2 | 1 |
| Siberian Huskies | 2 | (1) | 1 | — | — | 1 | — |
| Labrador Retrievers | 15 | (11) | 2 | 5 | 4 | 1 | 3 |
| Pit Bull | 11 | (8) | — | 3 | 1 | 4 | 3 |
| Mixed Breed Dogs | 13 | (9) | — | 4 | 2 | 7 | — |
| Herding | 15 | (11) | — | 7 | 3 | 3 | 3 |
| Hound | 7 | (5) | — | 4 | 3 | — | — |
| Toy | 14 | (10) | — | 7 | 2 | 4 | 1 |
| Non-Sporting | 11 | (8) | — | 2 | 3 | 6 | — |
| Sporting | 5 | (4) | — | 2 | — | 1 | 2 |
| Terrier | 8 | (6) | 2 | 3 | 1 | 2 | — |
| Working | 6 | (4) | — | 4 | — | 1 | 1 |
| NS | 3 | (2) | — | 2 | — | 1 | — |
| **Location** | | | | | | | |
| Right Neck | 52 | (38) | 1 | 19 | 7 | 14 | 11 |
| Left Neck | 37 | (27) | 1 | 10 | 10 | 13 | 3 |
| Cervical (side NS) | 35 | (27) | 3 | 16 | 7 | 6 | 3 |
| Bilateral | 5 | (3) | — | 2 | 3 | — | 1 |
| Elsewhere[3] | 4 | (3) | — | 3 | — | 1 | — |
| NS | 4 | (2) | 2 | 1 | — | 1 | — |

## CHAPTER 4: CONCLUSIONS AND FUTURE DIRECTIONS

The primary goal of this project was to train an artificial intelligence (AI) model to detect and categorize histologic whole slide images (WSIs) of canine thyroid carcinomas (CTCs) as either follicular origin (FTCs) or medullary origin (MTCs). Attempts were made to incorporate subtyping of well-differentiated FTCs according to the World Health Organization's Histologic Classification of the Tumors of the Endocrine System of Domestic Animals and for the development of a mitotic figure counter. Overall, most convoluted neural nets (CNNs) of the model are successful and are ready for use in a diagnostic setting in conjunction with interpretation by a pathologist and the continual addition of images with periodic re-training and re-validation of the model. While CNN 3 (follicular FTC versus compact FTC versus MTC) shows promising validation results, these results are less reliable and should be used with caution.

Future work for CNN 3 includes adding more training images of all subtypes from other institutions, from archival, or from new H&E-stained slides from within the UIUC VDL. Other areas of improvement include the determination of ground truth by consensus from a group of board-certified pathologists; possible utilization of image matching with accompanying IHCs; and enlisting a panel of experts, (ideally, at least three with one person per stage of model development, which includes the design and training, quality control, and the verification of the algorithm's output). Directions for the elaboration of this model include the complete development of a mitotic figure counter possibly with correlations of patient outcomes or treatment responses or image matching with an IHC specific for mitotic figures (e.g., phosphohistone H3 [PHH3]); comparing pathologist diagnoses with and without the use of AI and/or IHCs; incorporating more rare CTC subtypes (such as carcinosarcoma); evaluation of how

a fully validated model incorporates into and, likely improves, workflow efficiency; developing and comparing a similar model specifically focusing on nuclear features (similar to Wang et al., 2019b); correlating FTC subtypes and MTCs to response to radioactive iodine therapy; and/or correlating histologic image to genotypic derangements and/or patient outcomes (similar to what Coudray et al., 2018 and Laury et al 2021 describe). This latter point could allow for the development and use of highly customized precision medicine in veterinary medicine as the standard of care for relatively little time or cost, which is what the current trend in human medicine is. Additionally, by having specific tissue patterns highlighted by highly contrasting segmentation layer makes, previously overlooked histologic patterns may be observed which could then be presented to another unsupervised AI model to classify images independent from human input which could shed light on new relationships between histologic patterns and clinical features.

Future avenues of investigation that are not necessarily related to the use of an AI model include applying some of the reported human grading schemes to canine tumors to evaluate for prognostic ability and as well as the refinement and/or the provision of additional image examples of the current canine classification schemes to reflect those cases that are composed of more atypical or less-differentiated neoplastic thyroid cells (e.g., oxyphil cells, clear cells, giant cells, small cells, etc).

Overall, the signalments from this study resemble what is described in the literature. The compact FTC subtype was most common (supported by a minority of the literature sources) and MTCs may occur in slightly younger dogs, as is seen in the few studies that distinguish FTCs from MTCs. Notably, current CTC literature somewhat routinely fails to discriminate between FTCs and MTCs. Continual grouping of these tumors may obscure useful clinical correlations

but also promotes the potentiation of antiquated data, such as MTCs accounting for approximately 5% of all CTCs, although several more recent papers report rates similar to what is found here.

The diagnostic challenge of differentiating compact FTCs and MTCs without IHCs was highlighted in this study and is aligned with what is accepted in CTC literature.

Possible sources of bias include that this was a uni-institutional study with relatively homogenous hematoxylin and eosin (H&E) staining and slide quality (all slides were produced in the same laboratory over several consecutive years); cases were selected for use if they already had attendant IHCs; and the exclusion of one dog with a bilateral FTC that was found to have differing FTC subtypes on each side. The effects of the latter example are likely very small, given the overall number of cases identified. For the tertiary objective, interpretation of AI model output was not blinded and was only performed by one person, which is likely a source of bias.

This thesis provides a baseline for future applications of AI in CTC and can provide an outline for additional AI applications in veterinary medicine to allow for more consistent diagnoses or the investigation of subtle histologic changes in other veterinary disease processes, including, but not limited to, neoplasia (e.g., melanoma, mast cell tumors, etc). Inconsistencies of immunoreactivity with available IHCs, differences in veterinary pathologist interpretation of both IHC-stained and H&E-stained slides, and, historically, the lack of available calcitonin IHC could all have been confounding factors in previous CTC studies. The AI model developed here could assist in resolving many of these issues and allow for more accurate, repeatable, and cost-effective results. After further refinement, this model could then be applied to WSIs of future CTC studies, previous CTC studies, and possibly be implemented into the workflow of veterinary diagnostic laboratories in the future.

## BIBLIOGRAPHY

Abas FS, Shana'ah A, Christian B, Hasserjian R, Louissaint A Jr, Pennell M, Sahiner B, Chen W, Niazi MKK, Lozanski G, Gurcan M. (2017). Computer-assisted Quantification of CD3+ T Cells in Follicular Lymphoma. Cytometry Part A. 91A, 609-621.

Aeffner F, Wilson K, Martin NT, Black JC, Luengo Hendriks CL, Bolon B, Rudmann DG, Gianani R, Koegler SR, Krueger J, Young GD. (2017). The Gold Standard Paradox in Digital Image Analysis: Manual Versus Automated Scoring as Ground Truth. Archives of Pathology and Laboratory Medicine. 141, 1267-1275.

Al-antari MA, Al-masni MA, Choi MT, Han SM, Kim TA. (2018). A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. International Journal of Medical Informatics. 117, 44-54.

Al Rasheed MRH and Xu B. (2019). Molecular Alterations in Thyroid Carcinoma. Surgical Pathology Clinics. 12(4), 921-930.

Antonelli A, Ferrari SM, Fallahi P. (2018). Current and future immunotherapies for thyroid cancer. Expert Review of Anticancer Therapy. 18(2), 149-159.

Aubreville M, Bertram CA, Donovan TA, Marzahl C, Maier A, Klopfleisch R. (2020a). A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. Scientific Data. 4 (417), 1-10.

Aubreville M, Bertram CA, Marzahl C, Gurtner C, Dettwiler M, Schmidt A, Bartenschlager F, Merz S, Fragoso M, Kershaw O, Klopfleisch R, Maier A. (2020b). Scientific Reports. 10 (16447), 1-11.

Bai Y, Kakudo K, Jung CK. (2020). Updates in the Pathologic Classification of Thyroid Neoplasms: A review of the World Health Organization Classification. Endocrinology and Metabolism. 35(4), 696-715.

Barber LG. (2007). Thyroid Tumors in Dogs and Cats. Veterinary Clinics Small Animal Practice. 37, 755-773.

Benchoufi M, Matzner-Lober E, Molinari N, Jannot AS, Soyer P. (2020). Interobserver agreement issues in radiology. Diagnostic and Interventional Imaging. 101(10), 639-641.

Brearley MJ, Hayes AM, Murphy S. (1999). Hypofractionated radiation therapy for invasive thyroid carcinoma in dogs: a retrospective analysis of survival. Journal of Small Animal Practice. 40, 206-210.

Bulten W, Balkenhol M, Belinga JJA, Brilhante A, Cakir A, Egevad L, Eklund M, Farre X, Geronatsiou K, Molinie V, Pereira G, Roy P, Saile G, Salles P, Schaafsma E, Tschui J, Vos AM, ISUP Pathology Imagebase Expert Panel, van Boven H, Vink H, van der Laak J, Hulsbergen-van der Kaa C, Litjens G. (2021). Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. Modern Pathology. 34, 660-671.

Cabanillas ME, Ferrarotto R, Garden AS, Ahmed S, Busaidy NL, Dadu R, Williams MD, Skinner H, Gunn B, Grosu H, Iyer P, Hofmann MC, Zafereo M. (2018). Neoadjuvant BRAF- and Immune-Directed Therapy for Anaplastic Thyroid Carcinoma. Thyroid. 28(7), 945-951.

Cabanillas ME, Ryder M, Jimenez C. (2019). Targeted Therapy for Advanced Thyroid Cancer: Kinase Inhibitors and Beyond. Endocrine Reviews. 40(6), 1573-1604.

Castellino RA. (2005). Computer aided detection (CAD): an overview. Cancer Imaging. 5, 17-19.

Castillo V, Pessina P, Hall P, Cabrera Blatter MF, Miceli D, Soler Arias E, Vidal P. (2016). Post-surgical treatment of Thyroid Carcinoma in Dogs with Retinoic acid 9 cis Improves Patient Outcome. Open Veterinary Journal. 6(1), 6-14.

Cameselle-Teijeiro JM, Eloy C, Sobrinho-Simoes M. (2020). Pitfalls in Challenging Thyroid Tumors: Emphasis on Differential Diagnosis and Ancillary Biomarkers. Endocrine Pathology. 31, 197-217.

Campos M, Ducatelle R, Kooistra HS, Rutteman G, Duchateau L, Polis I, Daminet S. (2014a). Immunohistochemical Expression of Potential Therapeutic Targets in Canine thyroid carcinoma. Journal of Veterinary Internal Medicine. 28, 564-570.

Campos M, Ducatelle R, Rutteman G, Kooistra HS, Duchateau L, de Rooster H, Peremans K, Daminet S. (2014b). Clinical, Pathologic, and Immunohistochemical prognostic factors in dogs with thyroid carcinoma. Journal of Veterinary Internal Medicine. 28, 1805-1813.

Campos M, Kool MMJ, Daminet S, Ducatell R, Rutteman G, Kooistra HS, Galac S, Mol JA. (2014c). Upregulation of the PI3K/Akt Pathway in the Tumorigenesis of Canine Thyroid carcinoma. Journal of Veterinary Internal Medicine. 28, 1814-1823.

Carver JR, Kapatkin A, Patnaik AK. (1995). A Comparison of Medullary Thyroid Carcinoma and Thyroid Adenocarcinoma in Dogs: A Retrospective Study of 38 Cases. Veterinary Surgery. 24, 315-319.

Ceolin L, Amaro da Silveira Duval M, Benini AF, Ferreira CV, Maia AL. (2019). Medullary thyroid carcinoma beyond surgery: advances, challenges, and perspectives. Endocrine-Related Cancer. 26(9), R499-R518.

Chan HP, Samala RK, Hadjiiski LM, Zhou C. (2020). Deep Learning in Medical Image Analysis. Deep Learning in Medical Image Analysis: Challenges and Applications, Volume 1213. Ed. Lee G and Hujita H. Springer, Cham, Switzerland. 3-21.

Chapman, J. (2015). ATLAS-CITLStatisticsPublicWiki. Illinois Wiki. https://wiki.illinois.edu//wiki/display/ATLASCITLStatisticsPublicWiki

Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM, Xie W, Rosen GL, Lengerich BJ, Israeli J, Lanchantin J, Woloszynek S, Carpenter AE, Shrukumar A, Xu J, Cofer EM, Lavender CA, Turaga SC, Alexandar AM, Lu Z, Harris DJ, DeCaprio D, Qi Y, Kundaje A, Peng Y, Wiley LK, Segler MHS, Boca SM, Swamidass J, Huang A, Gitter A, Greene CS. (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of the Royal Society Interface. 15(141), 20170387.

Chmielik E, Rusinek D, Oczko-Wojciechowska M, Jarzab M, Krajewska J, Czarniecka A, Jarzab B. (2018). Heterogeneity of Thyroid Cancer. Pathobiology. 85,117-129.

Cohen S. (2021). The basics of machine learning: strategies and techniques. Chapter 2: Artificial Intelligence and Deep Learning in Pathology. Ed. S Cohen. Elsevier, Amsterdam, Netherlands. 13-39.

Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyo D, Moreira AL, Razavian N, Tsirigos A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nature Medicine. 24, 1559-1567.

Dolezal JM, Trzcinska A, Liao CY, Kochanny S, Blair E, Agrawal N, Keutgen XM, Angelos P, Cipriani NA, Pearson AT. (2021). Deep learning prediction of BRAF-RAS gene

expression signature identifies noninvasive follicular thyroid neoplasms with papillary-like nuclear features. Modern Pathology. 34, 862-874.

Donovan TA, Moore FM, Bertram CA, Luong R, Bolfa P, Klopfleisch R, Tvedten H, Salas EN, Whitley DB, Aubreville M, Meuten DJ. (2021). Mitotic Figures—Normal, Atypical, and Imposters: A Guide to Identification. Veterinary Pathology. 58(2), 243-257.

Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. (2021). Deep learning in cancer pathology: a new generation of clinical biomarkers. British Journal of Cancer. 124, 686-696.

Erdogan MF, Gursoy A, Erdogan G, Kamel N. (2006). Radioactive iodine treatment in medullary thyroid carcinoma. Nuclear Medicine Communications. 27, 359-362.

Fraiwan MA, Abutarbush SM. (2020). Using Artificial Intelligence to Predict Survivability Likelihood and Need for Surgery in Horses Presented With Acute Abdomen (Colic). Journal of Equine Veterinary Science. 90, 102973.

Fuchs TL, Nassour AJ, Glover A, Sywak MS, Sidhu SB, Delbridge LW, Clifton-Bligh RJ, Gild ML, Tsang V, Robinson BG, Clarkson A, Sheen A, Sioson L, Chou A, Gill AJ. (2020). A Proposed Grading Scheme for Medullary Thyroid Carcinoma Based on Proliferative Activity (Ki-67 and Mitotic Count) and Coagulative Necrosis. American Journal of Surgical Pathology. 44(10), 1419-1428.

Fuentes S, Gonzalez Viejo C, Cullen B, Tongson E, Chauhan SS, Dunshea FR. (2020). Artificial Intelligence Applied to a Robotic Dairy Farm to Model Milk Productivity and Quality based on Cow Data and Daily Environmental Parameters. Sensors (Basel). 20(10):2975.

Gambardella C, Offi C, Patrone R, Clarizia G, Mauriello C, Tartaglia E, Di Capua F, Di Martino S, Romano RM, Fiore L, Conzo A, Conzo G, Docimo G. (2019). Calcitonin negative

Medullary Thyroid Carcinoma: a challenging diagnosis or a medical dilemma?. BMC Endocrine Disorders. 19(Suppl 1), 42.

Gamer M. (2019). Cran.R-Project.org. Package 'irr'; Various Coefficients of Interrater Reliability and Agreement. https://cran.r-project.org/web/packages/irr/irr.pdf

Giannasi C, Rushton S, Rook A, Van Den Steen N, Venier, F, Ward P, Bell R, Trevail T, Lamb V, Eiras A, Ellis J, Roberts E. (2021). Canine thyroid carcinoma prognosis following the utilization of computed tomography assisted staging. Veterinary Record. E55, 1-9.

Girard J [Jeffrey Girard]. (2016, September 23) *First a note on Kodiologist's answer, in order to avoid confusion: It can be appropriate to use ICCs to estimate* [Comment on the online forum post *What to do with negative ICC values? Adjust the test or interpret it differently?]*. Stack Exchange. https://stats.stackexchange.com/questions/214124/what-to-do-with-negative-icc-values-adjust-the-test-or-interpret-it-differently.

Haddad RI, Nasr C, Bischoff L, Busaidy NL, Byrd D, Callender G, Dickson P, Duh QY, Ehya H, Goldner W, Haymart M, Hoh C, Hunt JP, Iagaru A, Kandeel F, Kopp P, Lamonica DM, McIver B, Raeburn CD, Ridge JA, Ringel MD, Scheri RP, Shah JP, Sippel R, Smallridge RC, Sturgeon C, Wang TN, Wirth LJ, Wong RJ, Jognson-Chilla A, Hoffman KG, Gurski LA. (2018). Thyroid Carcinoma, Version 2.2018, Featured Updates to the NCCN Guidelines. NCCN Guidelines Insights, Thyroid Carcinoma. 16(12), 1429-1440.

Hassan BB, Altstadt LA, Dirksen WP, Elshafae SM, Rosol TJ. (2020). Canine Thyroid Cancer: Molecular Characterization and Cell Line Growth in Nude Mice. Veterinary Pathology. 57(2), 227-240.

Hayes HM, Fraumeni JF. (1975). Canine Thyroid Neoplasms: Epidemiologic Features. Journal of the National Cancer Institute. 55(4), 931-934.

Hondelink LM, Huyuk M, Postmus PE, Smit VTHBM, Blom S, von der Thusen JN, Cohen D. (2021). Development and validation of a supervised deep learning algorithm for automated whole-slide programmed death-ligand 1 tumour proportion score assessment in non-small cell lung cancer. Histopathology. 80(4), 635-647.

Jegatheeson S, Zuber M, Woodward AP, Cannon CM. (2021). Response of canine thyroid carcinomas to radioiodine. Veterinary and Comparative Oncology. 1-11.

Kassambara A. (2018). Inter-rater reliability measures in R/Intraclass Correlation Coefficient in R. Data Novia. https://www.datanovia.com/en/lessons/intraclass-correlation-coefficient-in-r/#computing-icc-in-r

Kiupel M, Capen C, Miller M, Smedley R. (2008). World Health Organization International Histological Classification of Tumors of Domestic Animals: Histological Classification of Tumors of the Endocrine System of Domestic Animals, Second Series, Volume XII with Armed Forces Institute of Pathology, Washington, D.C. Ed. Schulman FY. Charles Louis Davis DVM Foundation, Gurnee, IL. 25-38 and 109-130.

Koo TK, Li MY. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine. 15, 155-163.

Laury AR, Blom S, Ropponen T, Virtanen A, Carpen OM. (2021). Artificial intelligence-based image analysis can predict outcome in high-grade serous carcinoma via histology alone. Scientific Reports. 11, 19165.

Leav I, Schiller AL, Rijnberk A, Legg MA, der Kinderen PJ. (1976). Adenomas and Carcinomas of the Canine and Feline Thyroid. American Journal of Pathology. 83(1), 61-94.

Lee BI, LaRue SM, Seguin B, Griffin L, Prebble A, Martin T, Leary D, Boss MK. (2020). Safety and efficacy of stereotactic body radiation therapy (SBRT) for the treatment of canine thyroid carcinoma. Veterinary and Comparative Oncology. 18, 843-853.

Lee JJ, Larsson C, Lui WO, Hoog A, Von Euler H. (2006). A dog pedigree with familial medullary thyroid cancer. International Journal of Oncology. 29(5), 1173-82.

Levine AB, Schlosser C, Grewal J, Coope R, Jones SJM, Yip S. (2019). Rise of the Machines: Advances in Deep Learning for Cancer Diagnosis. Trends in Cancer. 5(3), 157-169.

Li S, Wang Z, Visser LC, Wisner ER, Cheng H. (2020). Pilot study: Application of artificial intelligence for detecting left atrial enlargement on canine thoracic radiographs. Veterinary Radiology and Ultrasound. Nov;61(6), 611-618.

Liu YJ, Qiang W, Shi J, Lv SQ, Ji MJ, Shi BY. (2013). Expression and significance of IGF-1 and IGF-1R in thyroid nodules. Endocrine. 44(1), 158-164.

Liptak JM. (2007). Canine Thyroid Carcinoma. Clinical Techniques in Small Animal Practice. 22, 75-81.

List of Breeds by Group. (2020). American Kennel Club. https://www.akc.org/public-education/resources/general-tips-information/dog-breeds-sorted-groups/

Ma J, Jiang X, Fan A, Jiang J, Yan J. (2021). Image Matching from Handcrafted to Deep Features: A Survey. International Journal of Computer Vision. 129, 23-79.

Maekawa N, Konnai S, Nishimura M, Kagawa Y, Takagi S. Hosoya Km Ohta H, Kim S, Okagawa R, Izumi Y, Deguchi T, Kato Y, Yamamoto S, Yamamoto K, Toda M, Nakajima C, Suzuki Y, Murata S, Ohashi K. (2021). PD-L1 immunohistochemistry for canine cancers and clinical benefit of anti-PD-L1 antibody in dogs with pulmonary metastatic oral malignant melanoma. Nature Partner Journals Precision Oncology. 5(10).

Martucciello G, Lerone M, Bricco L, Tonini GP, Lombardi L, Del Rossi CG, Bernasconi S. (2012). Multiple Endocrine Neoplasias Type 2B and RET proto-oncogene. Italian Journal of Pediatrics. 38, 9.

Meijer JAA, Bakker LEH, Valk GD, de Herder WW, de Wilt JHW, Netea-Maier RT, Schaper N, Fliers E, Lips P, Plukker JT, Links TP, Smit JA. (2013). Radioactive iodine in the treatment of medullary thyroid carcinoma: a controlled multicenter study. European Journal of Endocrinology. 168, 779-786.

Mitchell BR. (2021). The basics of machine learning: strategies and techniques. Chapter 3: Overview of advanced neural network architectures. Ed. S Cohen. Elsevier, Amsterdam, Netherlands. 41-56.

Mohajon J. (2020). Confusion Matrix for Your Multi-Class Machine Learning Model. Towards Data Science. https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826.

Moore, FM, Kledzik GS, Wolfe HJ, DeLellis RA. (2020). Thyroglobulin and Calcitonin Immunoreactivity in Canine Thyroid Carcinomas. Veterinary Pathology. 21(2), 168-173.

Moxley-Wyles B, Colling R, and Verrill C. (2020). Artificial intelligence in pathology: an overview. Diagnostic Histopathology. 26(11), 513-520.

Nadeau ME, Kitchell BE. (2011). Evaluation of the use of chemotherapy and other prognostic variables for surgically excised canine thyroid carcinoma with and without metastasis. Canadian Veterinary Journal. 52, 994-998.

Nelson D. (2020). What is a Confusion Matrix. Unite. AI. https://www.unite.ai/what-is-a-confusion-matrix/.

Newkirk KM, Brannick EM, Kusewitt DF. (2017). Pathologic Bases of Veterinary Disease, 6th ed. Chapter 6: Neoplasia and Tumor Biology. Ed. JF Zachary. Elsevier, St Louis, Missouri. 286-321.

Nitulescu GM, Margina D, Juzenas P, Peng Q, Olaru OT, Saloustros E, Fenga C, Spandidos DA, Libra M, Tsatsakis AM. (2015). Akt inhibitors in cancer treatment: The long journey from drug discovery to clinical use (Review). International Journal of Oncology. 48, 869-885.

Pai RK, Hartman D, Schaeffer DF, Rosty C, Shivji S, Kirsch R, Pai RK. (2021). Development and initial validation of a deep learning algorithm to quantify histologic features in colorectal carcinoma including tumor budding/poorly differentiated clusters. Histopathology. 79, 391-405.

Patnaik AK, Lieberman PH. (1991). Gross, Histologic, Cytochemical, and Immunocytochemical Study of Medullary Thyroid Carcinoma in Sixteen Dogs. Veterinary Pathology. 28, 223-233.

Patnaik AK, Lieberman PH, Erlandson RA, Acevedo WM, and Lio S-K. (1978). Canine Medullary Carcinoma of the Thyroid. Veterinary Pathology. 15, 590-599.

Pessina P, Castillo VA, Cesar D, Sartore I, Meikle A. (2016). Proliferation, angiogenesis, and differentiation related marked in compact and follicular-compact thyroid carcinomas in dogs. Open Veterinary Journal. 6(3), 247-254.

Pessina P, Castillo V, Sartore, I, Borrego J, Meikle A. (2014). Semiquantitative immunohistochemical marker staining and localization in canine thyroid carcinoma and normal thyroid gland. Veterinary and Comparative Oncology. 14(3) e102-e112.

Pineyro P, Vieson MD, Ramos-Vara JA, Moon-Larson M, Saunders G. (2014). Histopathological and immunohistochemical findings of primary and metastatic medullary thyroid carcinoma in a young dog. Journal of Veterinary Science. 15(3), 449-453.

Pischon H, Mason D, Lawrenz B, Blanck O, Frisk AL, Schorsch F, Bertani V. (2021). Artificial Intelligence in Toxicologic Pathology: Quantitative Evaluation of Compound-Induced Hepatocellular Hypertrophy in Rats. Toxicologic Pathology. 49(4), 928-937.

Polonia A, Campelos S, Ribeiro A, Aymore I, Pinto D, Biskup-Fruzynska M, Veiga RS, Canas-Marques R, Aresta G, Araujo T, Campilho A, Kwok S, Aguiar P, Eloy C. (2021). Artificial Intelligence Improves the Accuracy in Histologic Classification of Breast Lesions. American Journal of Clinical Pathology. 155, 527-536.

R II: Inferential Statistics in R Workbook. Illinois Center for Innovation in Teaching & Learning. 1-12.

Ramos-Vara JA, Borst LB. (2017). Tumors of Domestic Animals. Chapter 3: Immunohistochemistry Fundamentals and Applications in Oncology. Ed. DJ Meuten. Wiley Blackwell, Ames, IA. 44-87.

Ramos-Vara JA, Frank CB, DuSold D, Miller MA. (2016). Immunohistochemical Detection of Pax8 and Napsin A in Canine Thyroid Tumors: Comparison with Thyroglobulin, Calcitonin and Thyroid Transcription Factor 1. Journal of Comparative Pathology. 155, 286-298.

Ramos-Vara JA, Miller MA, Johnson GC, and Pace LW. (2002). Immunohistochemical Detection of Thyroid Transcription Factor-1, Thyroglobulin, and Calcitonin in Canine Normal, Hyperplastic, and Neoplastic Thyroid Gland. Veterinary Pathology. 39, 480-487.

Revelle, W. (2021). Cran.R-Project.org. Package 'psych'; Procedures for Psychological, Psychometric, and Personality. https://cran.r-project.org/web/packages/psych/psych.pdf

Rosol TJ and Meuten DJ. (2017). Tumors of Domestic Animals. Chapter 18: Tumors of the Endocrine Glands. Ed. DJ Meuten. Wiley Blackwell, Ames, IA. 766-833.

Rosol TJ and Frone A. (2016). Jubb, Kennedy, and Palmer's Pathology of Domestic Animals, Volume 3, 6th ed. Chapter 3, Endocrine Glands. Ed. Maxie G. Elsevier, St. Louis, MO. 326-336.

Rudmann D and O'Shea D. (2020). Using Artificial Intelligence-Powered Image Analysis to Drive Discovery and Development, 7 May. 2020, https://www.criver.com/webinar-series-sa-are-you-ready-digital-pathology-revolution.

Scharf VF, Oblak ML, Hoffman K, Skinner OT, Neal KM, Cocca CJ, Duffy DJ, Wallace ML. (2020). Clinical features and outcomes of functional thyroid tumours in 70 dogs. Journal of Small Animal Practice. 61, 504-511.

Schober P, Mascha EJ, Vetter TR. (2021). Statistics From A (Agreement) to Z (z Score): A Guide to Interpreting Common Measures of Association, Agreement, Diagnostic Accuracy, Effect Size, Heterogeneity, and Reliability in Medical Research. Anesthesia and Analgesia. 133(6), 1633-1641.

Sheppard-Olivares S, Bello NM, Wood E, Szivek A, Biller B, Hocker S, Wouda RM. (2020). Toceranib phosphate in the treatment of canine thyroid carcinoma: 42 cases (2009-2018). Veterinary and Comparative Oncology. 18, 519-527.

Soares LMC, Pereira AHB, de Campos, CG, Rocha LS, dos Santos TA, Souza MA, Jark PC, Pescador CA. (2020). Histopathological and Immunohistochemical characteristics of Thyroid carcinoma in the Dog. Journal of Comparative Pathology. 177, 34-41.

Soler Arias EA, Castillo VA, Caneda Aristarain ME. (2016). Calcitonin-negative primary neuroendocrine tumor of the thyroid (nonmedullary) in a dog. Open Veterinary Journal. 6(3), 223-227.

SPSS 2: Inferential Statistics with SPSS. Illinois Center for Innovation in Teaching & Learning. 6-9.

Staup M. (2020). Controlling Quality of Automated Image Analysis. Charles River Laboratories, 29 Apr. 2020, https://www.criver.com/webinar-series-sa-are-you-ready-digital-pathology-revolution.

Stenman S, Siironen P, Mustonen H, Lundin J, Haglund C, Arola J. (2018). The prognostic significance of tall cells in papillary thyroid carcinoma: A case-control study. Tumor Biology. 1-7.

Sultan AS, Elgharib MA, Tavares T, Jessri M, Basile JR. (2020). The use of artificial intelligence, machine learning and deep learning in oncologic histopathology. Journal of Oral Pathology and Medicine. 49, 849-856.

Tellez D, Balkenhol M, Otte-Holler I, van de Loo R, Vogels R, Bult P, Wauters C, Vreuls W, Mol S, Karssemeijer N, Litjens G, van der Laak J, Ciompi F. (2018). Whole-Slide Mitosis Detection in H&E Breast Histology Using PHH3 as a Reference to Train Distilled Stain-Invariant Convolutional Networks. IEEE Transactions on Medical Imaging. 37, 2126-2136)

Tokarz DA, Steinbach TJ, Lokhande A, Srivastava G, Ugalmugle R, Co CA, Shockley KR, Singletary E, Cesta MF, Thomas HC, Chen VS, Hobbie K, Crabbs TA. (2021). Using Artificial Intelligence to Detect, Classify, and Objectively Score Severity of Rodent Cardiomyopathy. Toxicologic Pathology. 2021 49(4), 888-896.

Tosun AB, Pullara F, Becich MJ, Taylor DL, Fine JL, Chennubhotla SC. (2020). Explainable AI (xAI) for Anatomic Pathology. Advances in Anatomic Pathology. 27(4), 241-250.

Tsou P and Wu CJ. (2019). Mapping Driver Mutations to Histopathological Subtypes in Papillary Thyroid Carcinoma: Applying a Deep Convolutional Neural Network. Journal of Clinical Medicine. 8, 1675-1685.

Turner OC, Knight B, Zuraw A, Litjens G, Rudmann DG. (2021). Mini Review: The Last Mile – Opportunities and Challenges for Machine Learning in Digital Toxicologic Pathology. Toxicologic Pathology. 49(4), 714-719.

Turner OC, Aeffner F, Bangari DS, High W, Knight B, Forest T, Cossic B, Himmel LE, Rudmann DG, Bawa B, Muthuswamy A, Aina OH, Edmondson EF, Saravanan C, Brown DL, Sing T, Sebastion MM. (2020). Society of Toxicologic Pathology Digital Pathology and Image Analysis Special Interest Group Article*: Opinion on the Application of Artificial Intelligence and Machine Learning to Digital Toxicologic Pathology. Toxicologic Pathology. 48(s): 277-294.

Turrel JM, McEntee MC, Burke BP, Page RL. (2006). Sodium iodide I 131 treatment of dogs with nonresectable thyroid tumors: 39 cases (1990-2003). Journal of the American Veterinary Medical Association. 229(4), 542-548.

Valerio L, Pieruzzi, Giani C, Agate L, Vottici V, Lorusso L, Cappagli V, Puleo L, Matrone A, Viola D, Romei C, Ciampi R, Molinaro E, Elisei R. (2017). Targeted Therapy in Thyroid Cancer: State of the Art. Clinical Oncology. 29(5), 316-324.

Varricchi G, Loffredo S, Marone G, Modestino L, Fallahi P, Ferrari SM, de Paulis A, Antonelli A, Galdiero MR. (2019). The Immune Landscape of Thyroid Cancer in the Context of Immune Checkpoint Inhibition. International Journal of Molecular Sciences. 20, 3934.

Viera AJ, Garrett JM. (2005). Understanding Interobserver Agreement: The Kappa Statistic. Family Medicine. 37(5), 360-363.

Walsh DP, Ma TF, Ip HS, Zhu J. (2019). Artificial intelligence and avian influenza: Using machine learning to enhance active surveillance for avian influenza viruses. Transboundary and Emerging Diseases. 66(6), 2537-2545.

Wang J, Liu Q, Zhou X, He Y, Guo Q, Shi Q, Eriksson S, Zhou J, He E, Skog S. (2017). Thymidine kinase I expression in ovarian serous adenocarcinoma is superior to Ki-67: A new prognostic biomarker. Tumor Biology. 39(6), 1-8.

Wang S, Yang DM, Rong R, Zhan X, Xiao G. (2019a). Pathology Image Analysis Using Segmentation Deep Learning Algorithms. The American Journal of Pathology. 189(9), 1686-1698.

Wang Y, Guan Q, Lao I, Wang L, Wu Y, Li D, Ji Q, Wang Y, Zhu Y, Lu H, Xiang J. (2019b). Using deep convolutional neural networks for multi-classification of thyroid tumor by histopathology: a large-scale pilot study. Annals of Translational Medicine. 7(18), 468-481.

Wen S, Qu N, Ma B, Wang X, Luo Y, Xu W, Jiang H, Zhang Y, Wang Y, Ji Q. (2021). Cancer-Associated Fibroblasts Positively Correlate with Dedifferentiation and Aggressiveness of Thyroid Cancer. OncoTargets and Therapy. 14, 1205-1217.

Wetstein SC, Stathonikos N, Pluim JP, Heng YJ, ter Hoeve ND, Vreuls CPH, van Diest PJ, Veta M. (2021). Deep learning-based grading of ductal carcinoma in situ breast histopathology images. Laboratory Investigation. 101, 525-533.

Williams ED. (1966). Histogenesis of medullary carcinoma of the thyroid. Journal of Clinical Pathology. 19, 114-118.

Yoshimoto S, Kato D, Kamoto S, Yamamoto K, Tsuboi M, Shinada M, Ideka N, Tanaka Y, Yoshitake R, Eto S, Saeki K, Chambers J, Kinoshita R, Uchida K, Nishimura R, Nakagawa R. (2004). Immunohistochemical evaluation of HER2 expression in canine thyroid carcinoma. Heliyon. 5(7), e02004.

Yu Y, Bovenhuis H, Wu Z, Laport K, Groenen MAM, Crooijmans RPMA. (2021). Deleterious Mutations in the TPO Gene Associated with Familial Thyroid Follicular Cell Carcinoma in Dutch German Longhaired Pointers. Genes. 12(7), 997.

Zaiontz C. (2020). Real Statistics Using Excel. www.real-statistics.com.

Zapf A, Castell S, Morawietz L, Karch A. (2016). Measuring inter-rater variability for nominal data – which coefficients and confidence intervals are appropriate. BMC Medical Research Methodology. 63, 93.

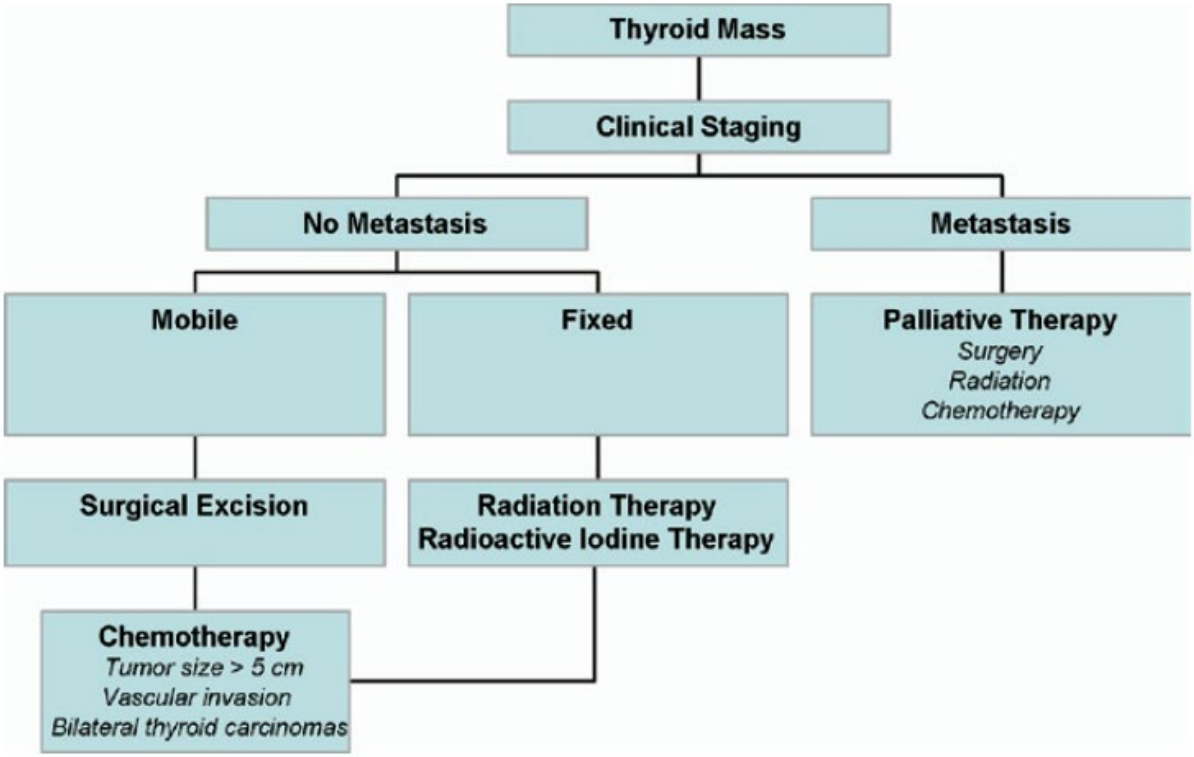
Zeya. (2021). TowardsDataScience.Com. Essential Things You Need To Know About F1-Score. https://towardsdatascience.com/essential-things-you-need-to-know-about-f1-score-dbd973bf1a3

Zuraw A. (2020). Demystifying Digital Pathology. Charles River Laboratories, 22 Apr. 2020, https://www.criver.com/webinar-series-sa-are-you-ready-digital-pathology-revolution.

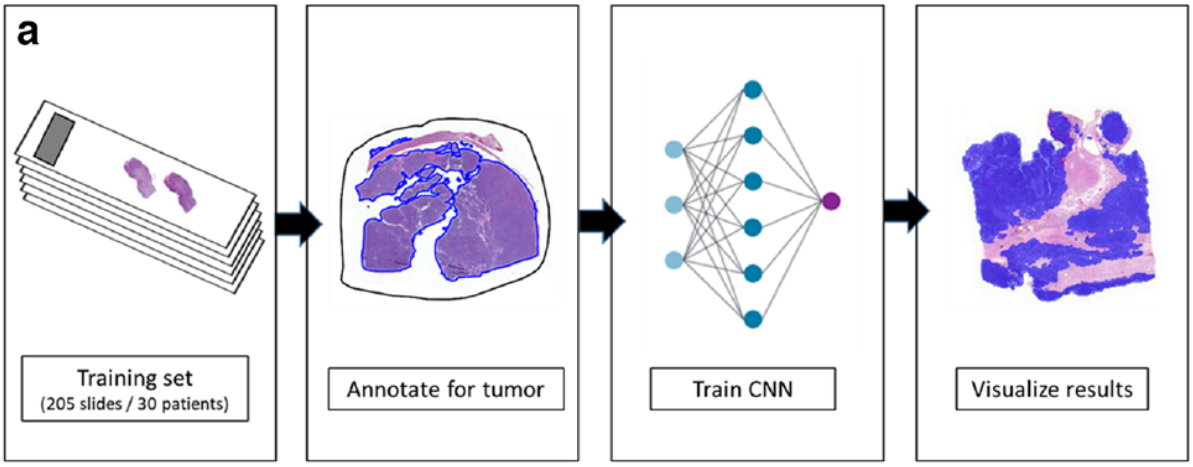
Zuraw A, Staup M, Klopfleisch R, Aeffner F, Brown D, Westerling-Bui T, Rudmann DG. (2020). Developing a Qualification and Verification Strategy for Digital Tissue Image Analysis in Toxicological Pathology. Toxicologic Pathology 49(4), 773-783.

## APPENDIX A: SUPPLEMENTARY FIGURES

**Supplementary Figure 1.** Algorithm for CTC treatment. This figure is from Liptak (2007) and outlines the approach to treating CTCs.



**Supplementary Figure 2.** Training a segmentation AI model. This partial figure is from Laury et al. (2021) and shows the general process of supervised learning for tumor segmentation. The training set WSIs are uploaded and manual segmentation annotations are applied. The model is trained and then outputs segmentation layer masks.



**Supplementary Figure 3.** All built confusion matrices. These are all of the confusion matrices made for this study. The diagonal from the upper left to the lower right shows the true positives, while the surrounding boxes show where errors occurred.

Interobserver Agreement of Validator Pathologists in Determining FTC or MTCs

| | | Path B | | |
|---|---|---|---|---|
| | | FTC | MTC | totals |
| Path A | FTC | 13 | 0 | 13 |
| | MTC | 2 | 10 | 12 |
| | totals | 15 | 10 | 25 |
| | Cohen's Kappa: | 0.84 | | |

| | | Path C | | |
|---|---|---|---|---|
| | | FTC | MTC | totals |
| Path A | FTC | 13 | 0 | 13 |
| | MTC | 1 | 11 | 12 |
| | totals | 14 | 11 | 25 |
| | Cohen's Kappa: | 0.92 | | |

| | | Path C | | |
|---|---|---|---|---|
| | | FTC | MTC | totals |
| Path B | FTC | 14 | 0 | 14 |
| | MTC | 1 | 10 | 11 |
| | totals | 15 | 10 | 25 |
| | Cohen's Kappa: | 0.92 | | |

Average Cohen's Kappa
0.89

Validator Pathologists' Ultimate Diagnosis

Pathologist A vs B

| | | B | | | | |
|---|---|---|---|---|---|---|
| | dx | FF | FC | FM | M | totals |
| | FF | 1 | 0 | 0 | 0 | 1 |
| A | FC | 0 | 3 | 3 | 1 | 7 |
| | FM | 0 | 0 | 5 | 2 | 7 |
| | M | 0 | 0 | 0 | 10 | 10 |
| | totals | 1 | 3 | 8 | 13 | 25 |

Pathologist A vs C

| | | C | | | | |
|---|---|---|---|---|---|---|
| | dx | FF | FC | FM | M | totals |
| | FF | 2 | 0 | 0 | 0 | 2 |
| A | FC | 0 | 4 | 2 | 0 | 6 |
| | FM | 1 | 1 | 3 | 0 | 5 |
| | M | 0 | 1 | 0 | 11 | 12 |
| | totals | 3 | 6 | 5 | 11 | 25 |

Pathologist B vs C

| | | C | | | | |
|---|---|---|---|---|---|---|
| | dx | FF | FC | FM | M | totals |
| | FF | 1 | 0 | 0 | 0 | 1 |
| B | FC | 0 | 3 | 1 | 0 | 4 |
| | FM | 2 | 3 | 4 | 1 | 10 |
| | M | 0 | 0 | 0 | 10 | 10 |
| | totals | 3 | 6 | 5 | 11 | 25 |

Validator Pathologists' Ultimate Diagnosis versus Verified Diagnosis

Pathologist A

| | | Original | | | | |
|---|---|---|---|---|---|---|
| | dx | FF | FC | FM | M | totals |
| | FF | 2 | 0 | 0 | 0 | 2 |
| A | FC | 1 | 5 | 0 | 0 | 6 |
| | FM | 0 | 0 | 5 | 0 | 5 |
| | M | 0 | 1 | 0 | 11 | 12 |
| | totals | 3 | 6 | 5 | 11 | 25 |

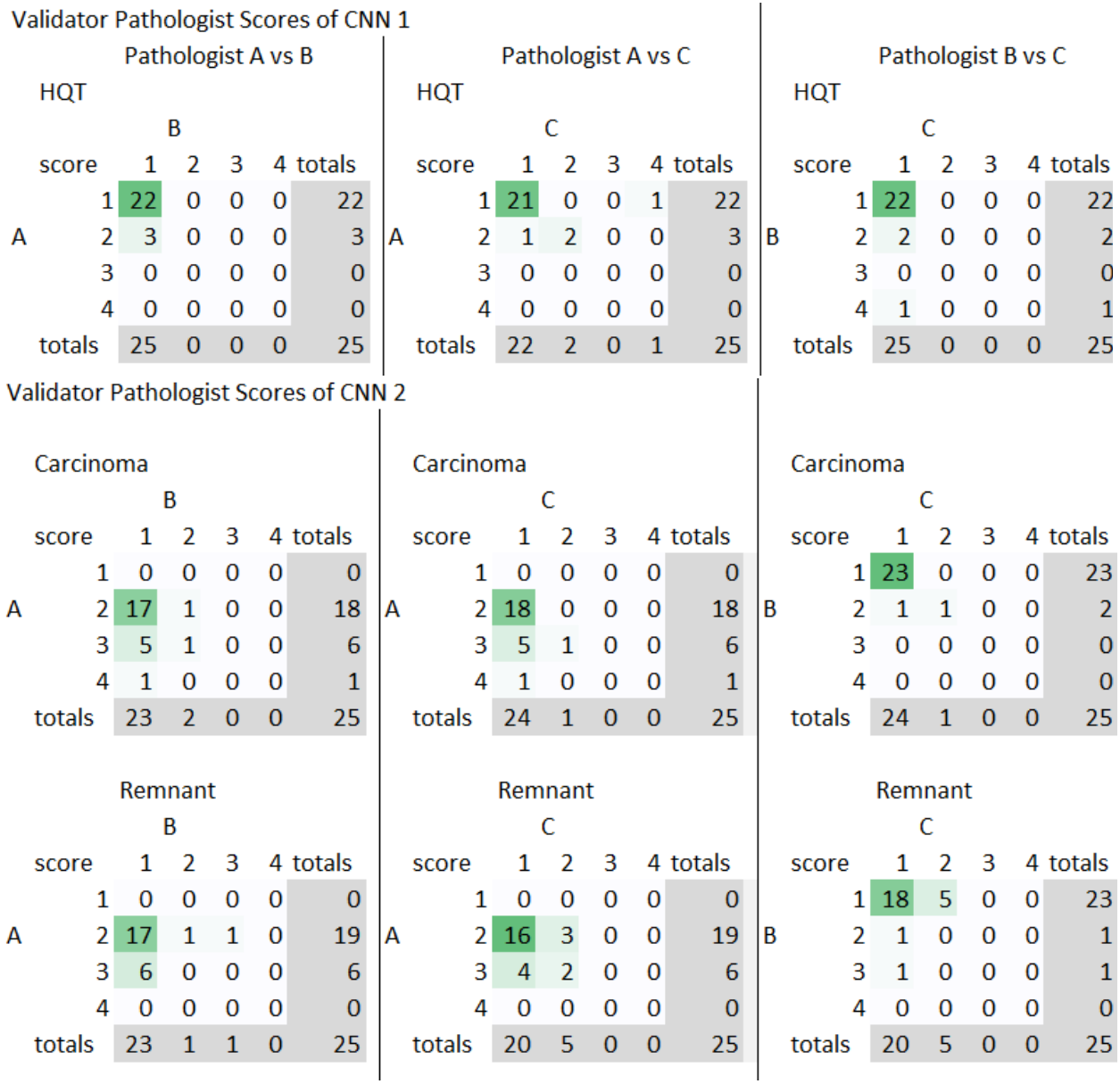
Pathologist B

| | | Original | | | | |
|---|---|---|---|---|---|---|
| | dx | FF | FC | FM | M | totals |
| | FF | 1 | 0 | 0 | 0 | 1 |
| B | FC | 0 | 3 | 1 | 0 | 4 |
| | FM | 2 | 3 | 4 | 1 | 10 |
| | M | 0 | 0 | 0 | 10 | 10 |
| | totals | 3 | 6 | 5 | 11 | 25 |

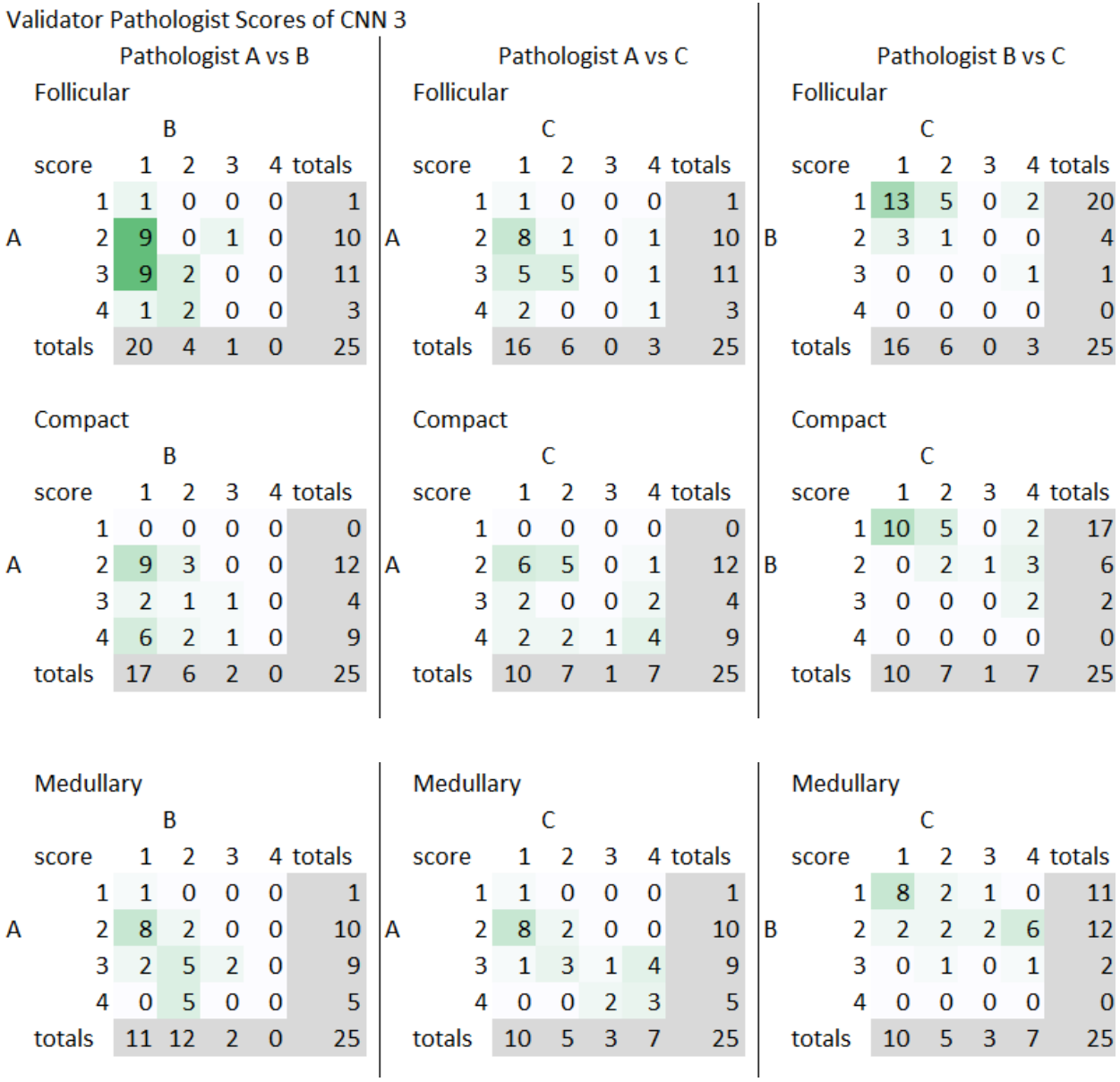
Pathologist C

| | | Original | | | | |
|---|---|---|---|---|---|---|
| | dx | FF | FC | FM | M | totals |
| | FF | 2 | 0 | 1 | 0 | 3 |
| C | FC | 1 | 4 | 1 | 0 | 6 |
| | FM | 0 | 2 | 3 | 0 | 5 |
| | M | 0 | 0 | 0 | 11 | 11 |
| | totals | 3 | 6 | 5 | 11 | 25 |

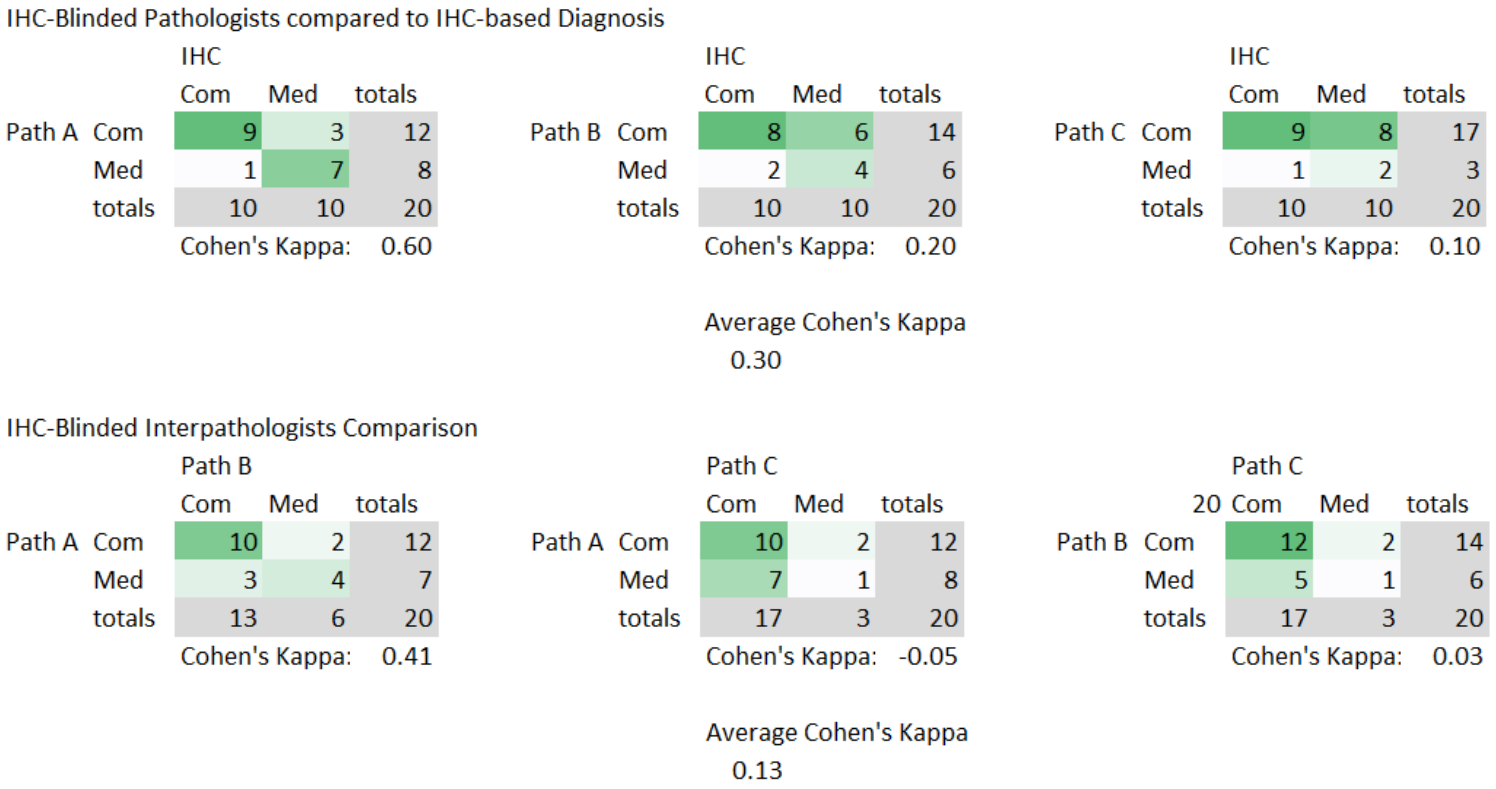
**Supplementary Figure 3 (continued).** All built confusion matrices.



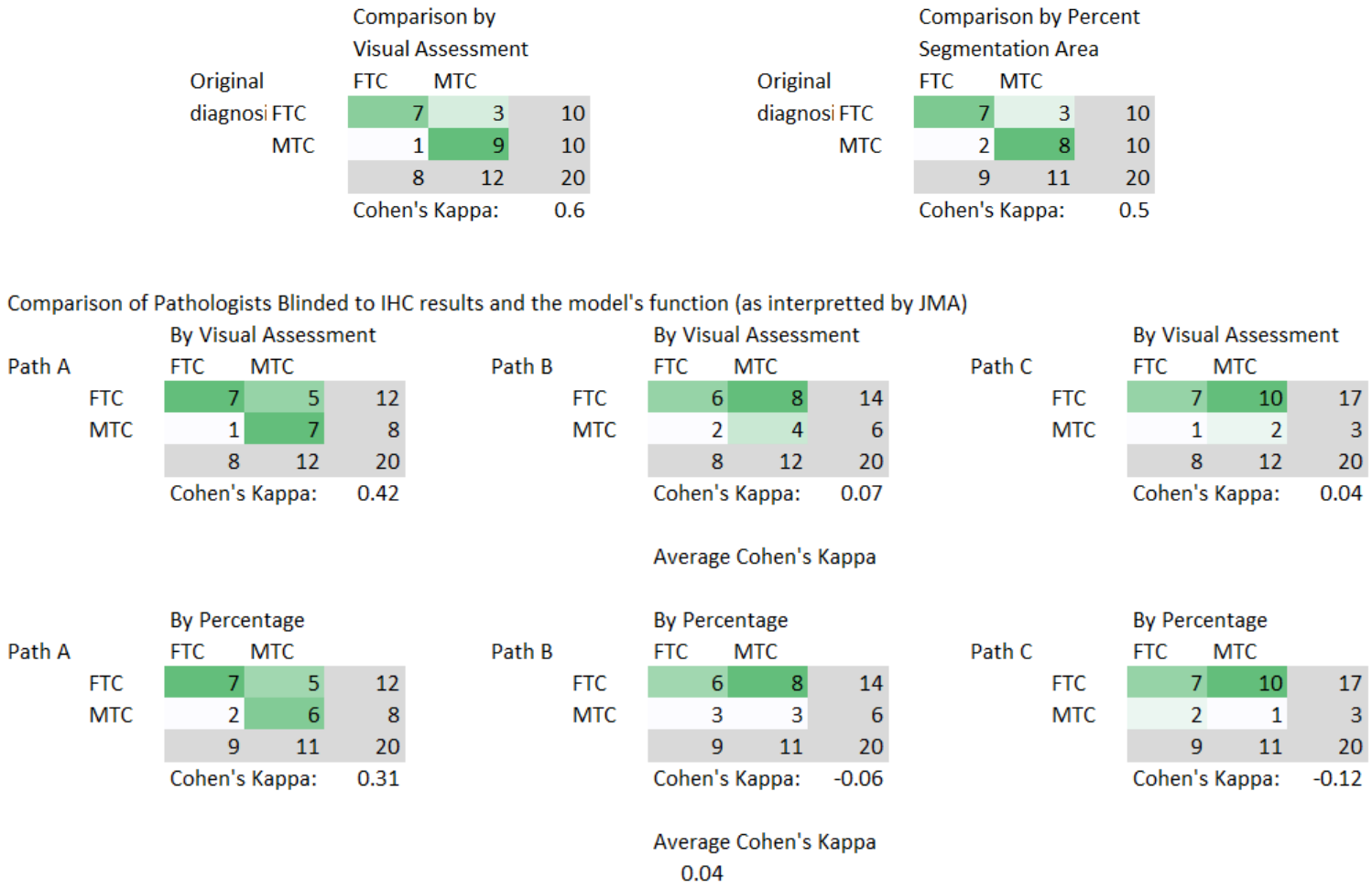
**Supplementary Figure 3 (continued).** All built confusion matrices.



**Supplementary Figure 3 (continued).** All built confusion matrices.



**Supplementary Figure 3 (continued).** All built confusion matrices.



**Supplementary Figure 4,** Example of phosphohistone H3 (PHH3) immunohistochemistry. From Tellez et al. (2018).

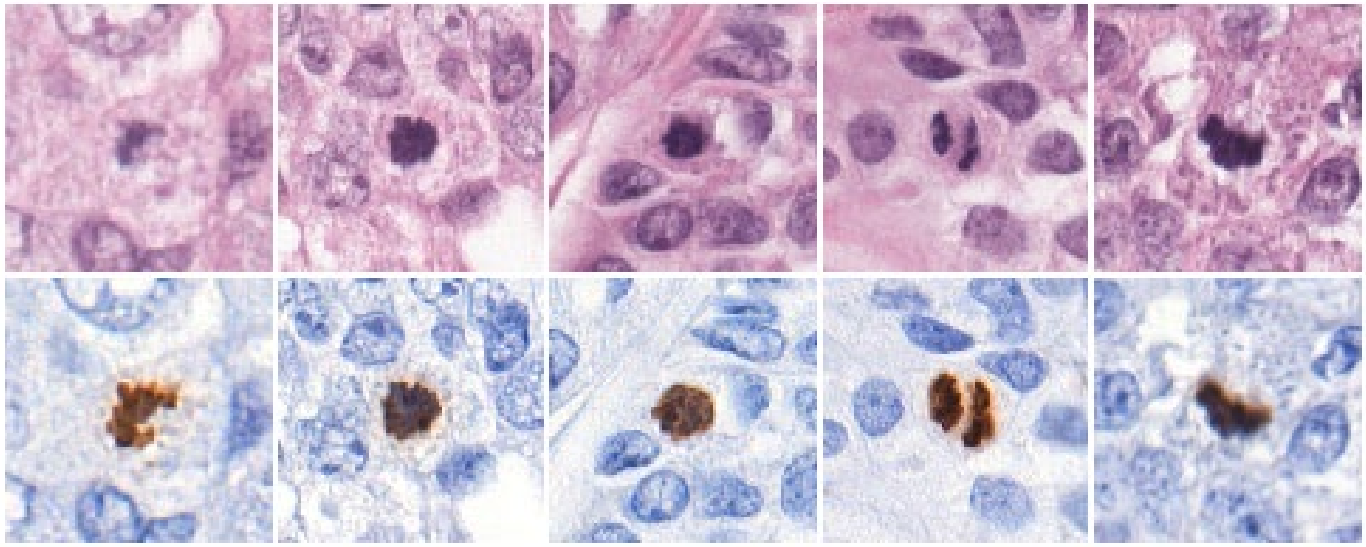


Fig. 1: Examples of image patches containing mitotic figures, shown at the center of each patch. In H&E (top), mitotic figures are visible as dark spots. In PHH3 (bottom), they are visible as brown spots. Mitotic figures in PHH3 stain are easier to identify than in H&E stain.