AUTOMATED DETECTION OF MATERNAL VASCULAR MALPERFUSION LESIONS IN HUMAN PLACENTAS DIAGNOSED WITH PREECLAMPSIA AND FETAL GROWTH RESTRICTION USING MACHINE LEARNING

PURVASHA PATNAIK

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Master of Science Interdisciplinary Health Sciences

Department of Interdisciplinary Health Sciences
Faculty of Health Sciences
University of Ottawa

Preface

Abstract

Introduction: Preeclampsia (PE) and fetal growth restriction (FGR) are common obstetrical complications, often with pathological features of maternal vascular malperfusion (MVM) in the placenta. Current placental clinical pathology methods involve a manual visual examination of histology sections, a practice that can be resource-intensive and demonstrate moderate-to-poor inter-pathologist agreement on diagnostic outcomes, dependant on the degree of pathologist sub-specialty training.

Methods: This thesis aims to apply different machine learning (ML) feature extraction methods to classify digital images of placental histopathology specimens, collected from PE, FGR, PE + FGR, and healthy pregnancies, according to the presence or absence of MVM lesions. 166 digital images were captured from histological placental specimens, manually scored for MVM lesions (MVM- or MVM+) and used to develop various support vector machine (SVM) classifier models, differing in feature extraction methods. Classification performance of each model was assessed through accuracy, precision, and recall using confusion matrices.

Results: SVM models demonstrated accuracies between 47-73% in MVM classification, with poorest performance observed on images with borderline MVM presence, as determined through manual observation. Data augmentation provided little to no improvement to the accuracies.

Conclusion: The results are promising for the integration of ML methods into the placental histopathological examination process. Using this study as a proof-of-concept foundation will lead our group and others to carry ML models further in placental histopathology.

Acknowledgements

Foremost, I would like to express my genuine gratitude to my thesis supervisor, Dr. Shannon Bainbridge. She has taught me the fundamental processes of research and how to effectively communicate it. Her motivation and passion are infectious when it comes to placental research, and I extend my heartfelt appreciation for the late-night meetings, last minute revisions, and her words of encouragement that kept me going. I am extremely grateful to have had the opportunity to conduct such valuable research under her guidance.

I would like to thank Dr. Adrian D. C. Chan for his advice throughout the methodology and coding process. I would also like to thank Dr. David Grynspan for his dedication to the project, and his passionate conversation and guidance. Thank you to Sina Salsabili, for the time and efforts he has dedicated to this project with his valuable research. Additionally, I would like to thank Anika Mukherjee for introducing me to the project and for her willingness to guide me.

A huge thank you to Dr. Goutham Vasam for his endless support for this thesis. Without his help, this project would not be possible. He gave me the best advice at crucial times and helped me to complete this project through a pandemic. His positivity and encouragement drove this project forward, and it was an immense pleasure to work with him on this project.

I would like to thank my grandparents and parents for their unconditional love and support throughout this process. Thank you, Mom, for instilling your passion of teaching and learning in me. Thank you, Dad, for being such an inspiration as a scientist to me. Aja – I feel your blessings every step of the way. Thank you, Aai, for lighting all those lamps for me. Without them, I would not be where I am today. Thank you to my sisters for everything they

have done for me; their humour, their company, and for being my ultimate role models. I am thankful to Sahil Kapal, who kept me grounded and pulled me out of the darkest tunnels. Thank you to Aarti Singla and Aanchal Agarwal for being the sincerest friends in the world.

I would like to thank God. With Your blessings, I have come this far, and I am merely an instrument of Your Hands.

Finally, I dedicate this thesis to all the strong mothers out there, and specifically those who have made this research possible. Your selfless donation of placental samples will help so many other mothers seeking solutions to their health.

"A mother is she who can take the place of all others, but whose place no one else can take." – Cardinal Mermillod

Authorship Contribution

The following description outlines the contribution of our authors for the manuscript prepared for this thesis. Dr. Shannon Bainbridge-Whiteside was the principal investigator for this study. Dr. David Grynspan examined and provided the cumulative MVM scores of the histopathological slides. Dr. Adrian D. C. Chan supervised the machine learning methodology and provided sponsorship for the access to Compute Canada High Performance Computing resources to run the programs. Patch extraction and image pre-processing codes were written by Purvasha Patnaik. Machine learning model codes were written by Dr. Goutham Vasam, Purvasha Patnaik, and Anika Mukherjee. Post-hoc analysis of machine learning model results was done by Purvasha Patnaik. Sina Salsabili applied the code developed to extract clinically relevant features from segmented placenta villi. This manuscript was prepared by Purvasha Patnaik and reviewed by Dr. Shannon Bainbridge-Whiteside, Dr. David Grynspan, Dr. Adrian D. C. Chan, Dr. Goutham Vasam, and Dr. Eranga Ukwatta.

Ethics Statement

The University of Ottawa Office Research Ethics board has examined and approved the ethics application for this thesis. Purvasha Patnaik was added as a student researcher on November 11, 2019, under the Ethics File Number of H-08-18-1023 for the project titled "Application of machine learning to identify distinct patterns of tissue damage in placenta histopathology images" (Appendix A).

Table of Contents

List of Abbreviations

AI	Acute Inflammation
AUC	Area Under the Curve
CAD	Computer-Aided Diagnosis
CHEO	Children's Hospital of Eastern Ontario
CI	Chronic Inflammation
EVT	Extravillous Trophoblast
FGR	Fetal Growth Restriction
FN	False Negatives
FP	False Positives
FVM	Fetal Vascular Malperfusion
H&E	Hematoxylin and Eosin
ML	Machine Learning
MVM	Maternal Vascular Malperfusion
PE	Preeclampsia
ROC	Receiver Operating Characteristic
RCWIH	Research Centre for Women's and Infants' Health
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machine
TN	True Negatives
TP	True Positives
WSI	Whole-Slide Image

List of Figures

Appendices

List of Tables

Chapter 3

Thesis Outline

Chapter 1 – Introduction

The introduction presents the research problem and the relevant background of the thesis.

Chapter 2 – Hypothesis and Research Aims

This chapter expands on the rationale, and provides the overarching hypothesis, as well as the research aims.

Chapter 3 – (Manuscript) Automated Detection of Maternal Vascular Malperfusion Lesions of the Placenta using Machine Learning

This manuscript describes the study which executes 3 different feature extraction methods to perform a support vector machine model. The motive of this study is to develop a model that is able of correctly classify placental histopathological images by MVM diagnosis.

Chapter 4 – General Discussion

This chapter provides the discussion and conclusions of the thesis. It also describes the significance of the discoveries and recommendations through future work.

Chapter 1

Introduction

1.1 Introduction

The human placenta is a transient organ of pregnancy essential to fetal growth and development. Disruptions in placental development can lead to several placenta-mediated diseases, including but not limited to preeclampsia, fetal growth restriction, and still-births [1]. They can also lead to long-term health consequences such as premature cardiovascular disease in mothers and metabolic dysfunction in offspring [1]. The placenta serves as a diary of in utero exposures across pregnancy, and as such detailed clinical pathology examination of this gestational organ following delivery can help identify underlying causes of obstetrical complications, and can help inform ongoing clinical management and counseling for mothers and babies [1]. This assessment must be done by a perinatal pathologist who specializes in placental pathology. The high degree of sub-specialty training required to effectively practice in this domain, coupled with the high resource demands of these exams, has resulted in smaller community healthcare settings being unable to leverage the utility of this practice and a high degree of inter-pathologist variability in diagnostic findings [1]. For this reason, it is proposed that automating some aspects of this examination will improve availability, reproducibility, and uniformity across all placental pathology examinations [1]. This thesis serves to investigate the performance of computer-aided diagnosis modalities within the placental pathology examination, specifically determining if a common and visually distinct form of placental disease (maternal vascular malperfusion, MVM) can be accurately classified in an automated fashion.

1.2 The Placenta: Function and Physiology

The successful development and function of the placenta are essential to a healthy pregnancy outcome. The main role of the placenta is to serve as a selective barrier between the mother and the fetus, where its function is to permit or prevent the exchange of certain nutrients, gases, hormones, and wastes [2]. The placenta permits the transport of essential macro- and micro-nutrients and gases to the fetal compartment required for fetal growth and development [2]. It also permits the passage of maternal IgGs, providing the fetus and neonate with immunity [2]. The placenta is also responsible for preventing the passage of many, but not all, pathogens from the maternal-fetal compartment to protect the fetus from infection [2]. It also generates hormones such as estrogen, progesterone, human chorionic gonadotrophin, and placental growth hormone that help promote fetal growth and maternal adaptation to pregnancy [3].

The placenta starts to develop in very early pregnancy at the implantation stage when the blastocyst attaches to the endometrial epithelium. The outer cells of the blastocyst, known as trophectoderm, form two distinct cell populations: the multinucleated syncytiotrophoblast cells on the outside and an underlying layer of progenitor cytotrophoblast cells (Figure 1.1) [4][5]. Small projections of the cytotrophoblasts start to expand into the syncytiotrophoblasts, forming the primary villi structures [6]. Subsequently, the extraembryonic mesoderm grows into the primary villi, forming the secondary villi structures [6]. The tertiary villi structures form when embryonic blood vessels start to form in the extra-embryonic mesoderm of the secondary villi [6]. These tertiary villi structures mature into chorionic villi trees, which encase the fetal vasculature and connects to the fetus via the umbilical cord [6].

There are two functional compartments to the human placenta: 1) the extravillous compartment – made up of specialized extravillous trophoblast (EVT) cells that have invaded into the decidua of the uterine wall and remodeled the uterine spiral arteries to ensure adequate maternal blood flow into the placenta; 2) the villous compartment – made up of highly branched chorionic villous trees that facilitate maternal-fetal exchange (Figure 1.1) [7]. The maternal spiral arteries remain plugged until ~ 12 weeks, after which the plugs break down and maternal blood flow into the placenta commences [8]. Importantly, no mixing of maternal and fetal blood takes place in the placenta. Instead, all exchange of gases, nutrients, and waste products takes place by passive or active diffusion and transport across the placental membranes of the chorionic villous trees. The chorionic villi trees are bathed in maternal blood which is brought into and fills the intervillous space surrounding the villi [5]. Importantly, when any aspect of placental development becomes compromised, an umbrella of 'placenta-mediated diseases' can arise, including two of the most common obstetrical complications: preeclampsia and fetal growth restriction[9].

1.3 Preeclampsia and Fetal Growth Restriction

Preeclampsia (PE) is a common and debilitating hypertensive disorder of pregnancy. It is defined as de novo onset of hypertension after 20 weeks of gestation, coupled with evidence of maternal end-organ dysfunction [10]. Three to eight percent of all pregnancies are affected by PE and it is responsible for 63,000 maternal deaths worldwide per year [11][12]. Currently, there is no cure for PE; the only solution is the delivery of the placenta [13]. Although the etiology is unknown, the pathophysiology of PE is commonly characterized by a two-stage

model of disease progression [14]. The first stage involves shallow invasion of the EVTs into the uterine decidua basalis, coupled with the insufficient remodeling of the uterine spiral arteries, resulting in inadequate blood flow into the placenta [14]. The second stage of pathophysiology is initiated when the malperfused and hypoxic placenta secretes several pro-inflammatory and anti-angiogenic factors [15] into the intervillous space of the placenta (Fig 1), transporting them into the systemic maternal circulation where they elicit a heightened inflammatory response and endothelial dysfunction [16]. The endothelial dysfunction and pro-inflammatory environment result in the classical signs and symptoms used to diagnose PE in the mother (i.e. hypertension, end-organ damage) [14]. While PE can have serious and long-lasting effects on the mother's health, approximately 20% of the fetuses from these pregnancies also demonstrate the comorbidity of fetal growth restriction – also recognized to put offspring at elevated risk for many chronic diseases across their life [17].

Fetal growth restriction (FGR) is the pathological failure of a fetus to achieve its genetic growth potential in utero [18]. A diagnosis of FGR includes low fetal/neonatal weight (typically <5-10%ile) accompanied by evidence of placental contribution to poor fetal growth, such as abnormal uterine artery Doppler measurements in utero and/or distinct placenta pathology findings at the time of delivery [19]. FGR affects 5-10% of all pregnancies and is a leading cause of perinatal morbidity and death, with long-term implications for cardiovascular, neurological, and metabolic systems [20]. Fetuses that fail to meet their growth potential in utero are at risk for stillbirth, preterm birth, and adverse neonatal and long-term health outcomes [21]. FGR can occur in both the presence and absence of PE, the latter of which being referred to as normotensive FGR [22]. While the exact pathophysiology of FGR is not entirely known, a

common finding of similar deficiencies in placentation has been widely described in both cases of PE and FGR [9][20].

There is strong evidence to suggest that the two-stage model of disease described above, which emphasizes the importance of placental malperfusion as central to disease development, is common to both PE and FGR. However, more recent investigations focused on detailed placental profiling indicate that these clinical diagnoses can also arise as a consequence of heightened inflammation at the maternal-fetal interface or poor villous development [23]. The fact that similar underlying placental disease can result in different maternal and fetal health outcomes, maternal hypertensive disease, and/or FGR, indicates that it is likely the physiological response of the mother and/or fetus to the placental disease that dictates the subsequent clinical outcome [19]. A significant challenge in understanding the specific disease processes that underly PE and/or FGR, and which have hindered our ability to identify effective predictive tests and treatments, is the high degree of clinical heterogeneity observed in these patient populations. Patients diagnosed with placenta-mediated diseases have a distinct variety of clinical differences relevant to their pregnancies, such as time of disease onset, disease severity, and individual or combined clinical outcomes for mother and infants [24][25][26]. Based on the patient's unique clinical profile and their subsequent response, these clinical outcomes could be that the mother is affected (PE), the baby is affected (FGR), or both are affected (PE + FGR).

1.4 Placental Pathology in Cases of PE and FGR

A diagnosis of PE and/or FGR at the time of delivery is an indication for submission of the placenta for pathological examination following delivery. Current practice includes a gross evaluation of the entire organ, as well as a histological evaluation of ~ four full-thickness sections of the placenta for the presence/absence and severity of a variety of lesions known to result from: developmental abnormalities; damage done by poor uteroplacental perfusion; maternal metabolic disorders; infectious pathogens; or chronic inflammation. These examinations, when done correctly, can be time and resource-intensive and are ideally performed by a specially trained perinatal pathologist. However, these examinations serve as an important component of maternal and neonatal care – as the collection of microscopic evidence from the placenta can serve as a diary of the intra-uterine environment and lead to appropriate diagnoses of placenta-mediated disease.

From the time the placenta begins to develop to the time it is removed at delivery, there are numerous points at which development and/or function can become compromised. These vulnerabilities are what lead to placenta-mediated diseases. Four broad forms of placenta-mediated diseases can be identified during a routine high-level placental pathology examination:

1. *Acute Inflammation (AI):* High-stage maternal or high-stage fetal acute inflammation is present.
2. *Chronic Inflammation (CI):* Chronic inflammation in the chorionic plate and membranes, basal plate, villi, intervillous space, and fetal vasculature.

3. *Fetal Vascular Malperfusion (FVM):* Presence of thrombus or intramural fibrin deposition in at least one fetal vessel and/or avascular villi.
4. *Maternal Vascular Malperfusion (MVM):* Presence of distinct placental lesions consistent with poor utero-placental perfusion, including infarcts, increased syncytial knots, villous agglutination, distal villous hypoplasia, etc. Relevant to the current thesis project, the majority of PE and FGR cases demonstrate common gross and microscopic placental findings of MVM placental disease [9][19][27].

1.5 Maternal Vascular Malperfusion of the Placenta

Maternal vascular malperfusion (MVM) refers to insufficient blood flow into the placenta via the uterine spiral arteries, leading to ischemia-reperfusion injuries to this vital organ of pregnancy [28]. As discussed above, the process of uterine spiral artery remodeling by the EVTs is required to establish a robust uteroplacental circulation required to sustain fetal growth. Insufficient remodeling of these vessels in early pregnancy (within the 1st trimester) compromises uteroplacental perfusion across gestation and is associated with many adverse pregnancies and neonatal outcomes [29]. Historically, MVM lesions were considered the hallmark of PE-specific placental pathology. However, it is now recognized that MVM lesions are likewise a common finding in cases of FGR, preterm labor, placental abruption, and stillbirth [29].

Establishing the diagnosis of MVM in the placenta at the time of the delivery via clinical placenta pathology examination informs the ongoing clinical care of the mother and infant, and

can be used to counsel families on the reasons for poor pregnancy outcomes and future pregnancy recurrence risks [27] – as patients with placenta-mediated diseases with MVM lesions have a 10-25% risk of recurrence in subsequent pregnancies [28]. Interestingly, these placental lesions have also been associated with poor long-term cardiometabolic health in the mothers, further allowing for preventative health counseling and intervention for these patients [30].

Several macro- and microscopic lesions identified during a placenta pathology examination are consistent with a diagnosis of MVM – all of which demonstrate a distinct visual appearance (Figure 1.2) [31]. Microscopic placental lesions of MVM include placental infarcts, distal villous hypoplasia, accelerated villous maturation, increased syncytial knots, and villous agglutination [32]. Placental infarcts are areas of placental tissue that have undergone necrosis due to the lack of sufficient blood supply, and can often be visualized grossly as white areas of necrotic tissue [32]. Their presence is often confirmed histologically through the presence of chorionic villous tree necrosis [28]. Distal villous hypoplasia is most commonly associated with low placental and fetal weights, and its pattern can be described as poorly developed distal chorionic villous tree structures, visualized as sparse and small villous trees encompassed by an extended intervillous space [28][32]. Accelerated villous maturation is defined as over-mature villi that are small in size for a given gestational age, and are often seen in parallel with an elevated number of syncytial knots [32]. The term syncytial knot is used to describe the tight accumulation of nuclei within the syncytial layer, often budding from the surface [33]. All placentas demonstrate some evidence of syncytial knotting as the placenta ages. However, in cases of MVM, the number and size of these knots drastically increase. Finally, villous

agglutination demonstrates a pattern of adherent terminal villi, held together by excessive fibrin deposition within the intervillous space, accompanied by fibrosis of the villous stroma [31]. Collectively, the different MVM lesions demonstrate distinct visual patterns, and findings of MVM lesions in a clinical pathology exam are indicative of ischemia-reperfusion injury to the placenta and compromised maternal-fetal exchange during the pregnancy. These lesions can be found in villous tissue samples, and holistically, decidual vessel changes would show evidence of MVM. It is estimated that ~60% of all cases of PE and FGR demonstrate evidence of MVM lesions within their placenta at the time of delivery [9], indicative of insufficient uteroplacental blood flow as a significant contributory factor to the disease pathogenesis.

1.6 Computer-Aided Diagnosis in Clinical Pathology

Clinical placenta pathology is a highly important component of effective maternal and neonatal care, particularly for patients with placenta-mediated diseases of pregnancy. Unfortunately, this clinical practice has historically been plagued by a lack of standardization in employed terminology and diagnostic criteria, limiting the clinical utility of this important practice. Despite landmark efforts undertaken to improve global standardization in this field, the degree of inter-pathologist variability in examination findings remains high, particularly in low resource settings where specialized training and expertise are lacking [32]. Considering these challenges, placental pathology has been identified as a key clinical domain that may benefit from the implementation of computer-aided diagnosis through the development of machine learning algorithms capable of identifying distinct visual patterns in placental histopathology specimens.

Computer-aided diagnosis (CAD) is the application of modern artificial intelligence to biological samples to classify them according to diagnostic labels. Classes can be binary (control vs diseased) and multiclass (control vs categories of disease). The objective of CAD is to employ a model that will input data (more commonly image data) and provide an output for clinicians to evaluate and use in their clinical decision-making. Deployments of these models involve curating a dataset of the appropriate classes, training the model, allowing the model to make predictions, and performing a clinical evaluation based on the model's performance accuracy. CAD was first applied to histological slides in the 1980s [34]. During that time, many issues hindered the progress of CAD in clinical pathologies, such as image resolution and quality and prohibitively high costs [34]. Over the following decades, digital pathology capabilities and CAD technology has advanced to a point where the implementation of CAD-based modalities in clinical pathology is beginning to take hold. Ongoing advancement of CAD tools in this domain is certain to improve clinical care while reducing diagnostic variability and physician workload burnout.

In other fields of clinical pathology, the integration of CAD into the diagnostic workflow is demonstrating tremendous promise – specifically in the domain of oncology pathology. Research and clinical implementation of CAD have shown success for breast, prostate, thyroid, ovarian, and lung cancer to date[35][36][37][38]. These models have been used to successfully classify benign vs malignant tumours, distinct subclasses of malignant tumours, and have even been capable of accurately identifying the prognosis (short- vs. long-term survival) of patients with lung cancer [35][36][38]. With these considerable advancements being made in the field

of oncology pathology, it is time for other pathology sub-specialties to take notice and begin testing and integrating CAD modalities across all fields of clinical pathology.

1.7 Application of Machine Learning for Digital Image Analysis

Many previously described CAD models developed and applied in the field of pathology are machine learning models. Machine learning (ML) is the development of automated systems that are capable of processing large amounts of complex data and learning from this data to carry out certain tasks – including predicting classifications of previously unseen data [39]. Computer vision, or computer-driven image recognition, is a branch of ML, in which the dataset used to teach the computer is digital image data, and the goal is to train the computer to interpret and classify previously unseen digital images [40]. Through ML, biomedical advances in computer vision have been progressing rapidly by using biomedical image datasets for recognition and classification. [39]. These models are the basis of CAD, where CAD envelopes the process of using ML models to the endpoint of supporting clinicians in treatment decisions for patients.

ML algorithms for digital image analysis can use both unsupervised and supervised approaches. Unsupervised ML involves feeding the computer unlabelled images [39], allowing the computer to analyze and identify clusters of images with similar inherent features – generating distinct groups of the most similar images [39]. Conversely, supervised ML is the process of feeding the computer a dataset of images that have been pre-labeled [39]. The computer then ‘learns’ the features and patterns of the images associated with each label and

can be tested by feeding in a new, unseen and un-labelled image dataset [39]. The aim is for the computer to be able to label these unseen and unlabelled images with a high degree of accuracy. Supervised ML approaches can be broadly classified according to the problems they are aiming to solve – either being used to predict a discrete class label (classification) or to predict a continuous quantity (regression). For the purposed of CAD tools within a healthcare setting, the classification category of machine learning models is most applicable.

The Support Vector Machine (SVM) algorithm is commonly used across various cases of automated classification. It uses plotted data points representing features of the images to find the best hyperplane that will segregate the classes [41]. The advantages of using SVM are that they are usually successful in finding the most optimal hyperplane, and they can work with high-dimensional data [41]. The algorithm is relatively simple and less exhaustive on computer memory space. A disadvantage is that although it is simple, its complexity is limited, and rather than analyzing the image itself, it analyzes the features of the image [41]. Common image datasets that have been used for successful SVM model development in past have included the MNIST dataset (a combination of handwritten numbers) and the iris dataset (different classes of flowers).

Accuracy is the main performance measure for supervised ML experiments. It is determined through comparison of the true classifier of the image to that which the algorithm estimates for the same image. True positives, true negatives, false positives, and false negatives are commonly presented in a confusion matrix to best demonstrate the performance of developed models. Confusion matrices display the distribution of correct classification and classification errors [42].

1.8 Application of Machine Learning to Placental Histopathology

Research focused on the application of ML-CAD modalities within the field of placenta pathology has been very limited in scope to date. With ML applications in their infancy in placental pathology, only a handful of models have been attempted and published. To-date, supervised ML models have been used successfully to predict gestational age from placental whole-slide digital images [43], and further have been used to accurately detect signs of suspicious invasive placentation in MRI image datasets [44]. These projects are certainly an exciting initial foray into the application of ML to better understand and help diagnose placenta-mediated diseases of pregnancy, but only scratch the surface of work that can be carried out in this domain.

Our group has begun making headway in this area, applying ML approaches to histological placental samples of patients with and without PE, classifying images based on the obstetrical diagnosis. ML models were applied to a placenta image dataset collated from 148 patients, with classification accuracies ranging from 41-73%, depending on the method of feature extraction used in the model development (unpublished data, Mukherjee et al.). This preliminary body of work demonstrates considerable promise for the use of supervised ML models to identify placenta-mediated disease, however, it also highlighted that the high degree of patient heterogeneity underlying the diagnosis of PE is a considerable challenge to overcome. Armed with this knowledge, along with lessons learned regarding dataset pre-processing and algorithm development, the current MSc thesis aims to develop and apply different supervised ML models for the classification of placental digital histopathology images according to the presence or absence of a common and visually distinct form of placental

damage common to PE and FGR – maternal vascular malperfusion. It is anticipated that training an algorithm with images labeled according to placenta pathology findings, rather than obstetrical outcomes, will greatly improve the training and performance of the developed algorithms, and will prove to have greater biological relevance in a clinical setting. The results of this study will serve as an important proof-of-concept, demonstrating the tremendous potential of CAD to improve the accessibility to consistent and high-quality diagnostic outputs of the placenta pathology exam while reducing the resource-intensive workload of this important clinical practice – ensuring that mothers and infants can receive the best possible health care afforded to them.

Figures

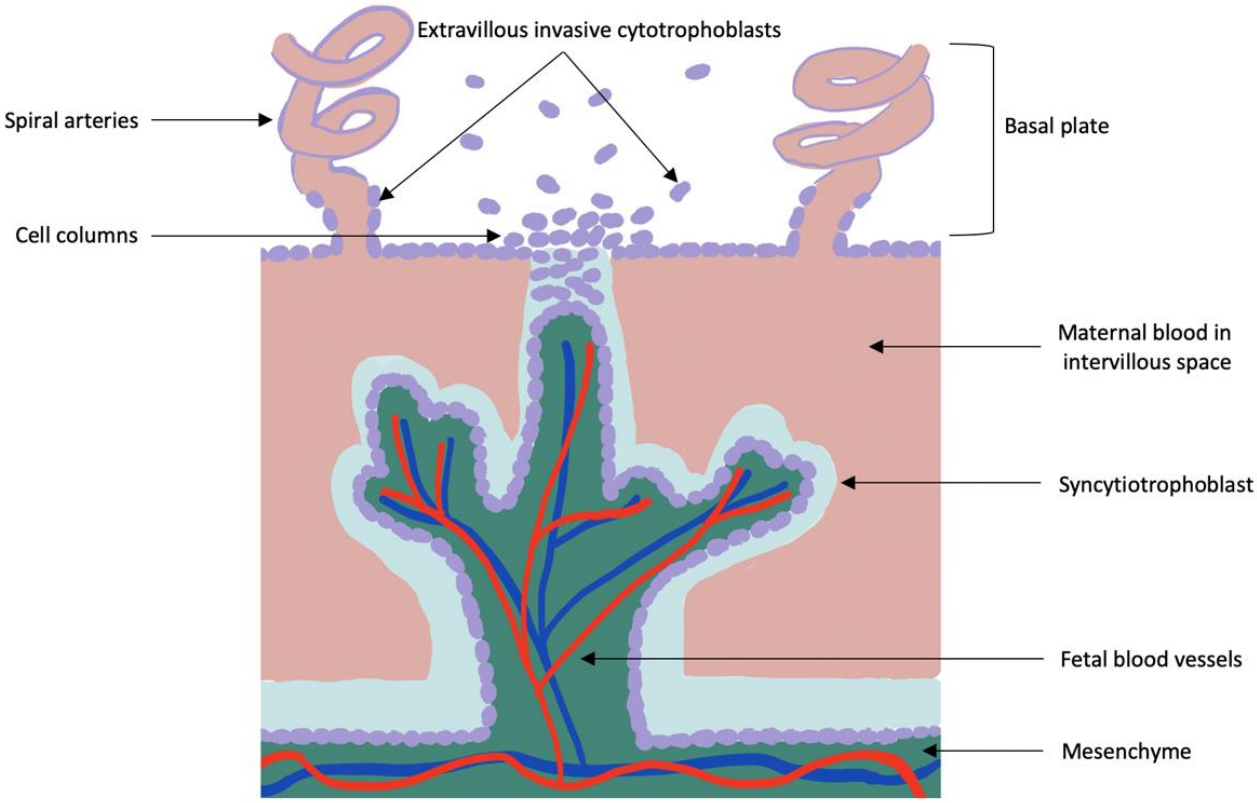


Figure 1.1 Labeled anatomy of a tertiary chorionic villi structure from the placenta.

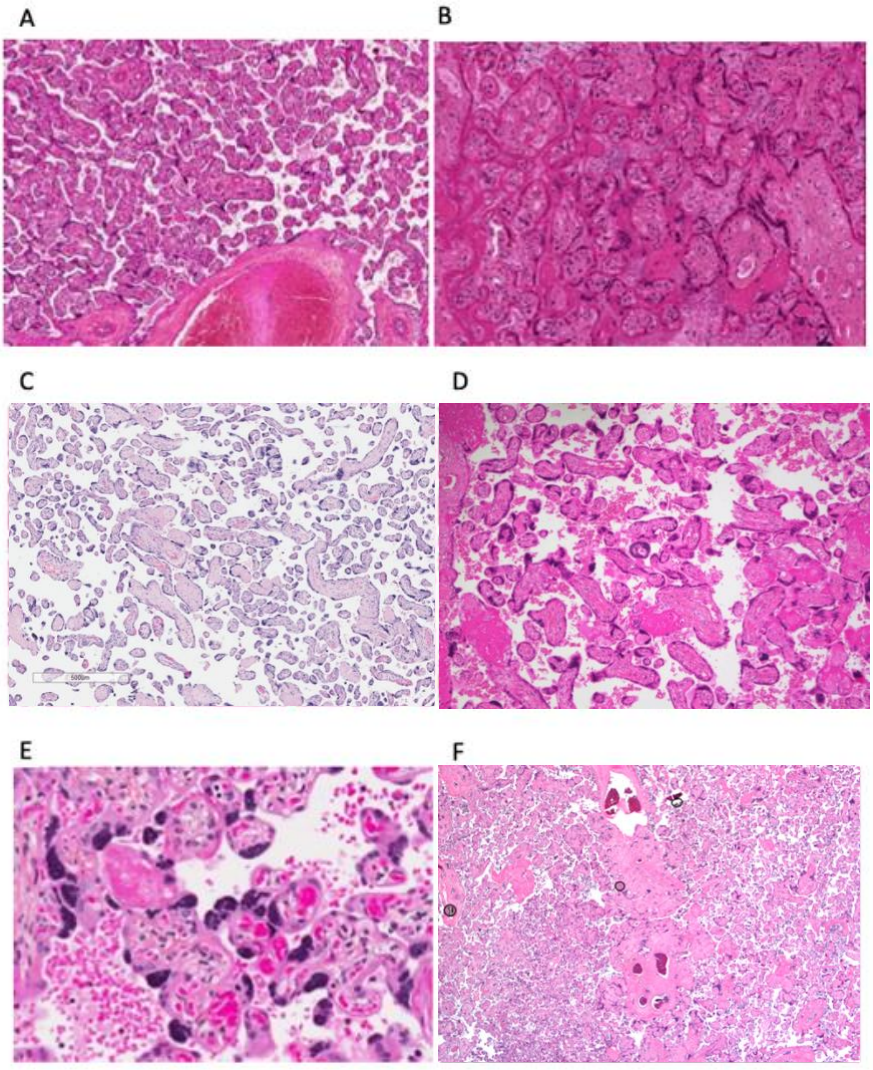


Figure 1.2. Placental lesions that define MVM. **A**: Normal placenta at term, **B**: Placental infarct, **C**: Distal Villous Hypoplasia, **D**: Accelerated Villous Maturation, **E**: Syncytial knots, **F**: Villous Agglutination. (All panels provided by Dr. David Grynspan).

Chapter 2

Hypothesis and Research Aims

2.1 Rationale

Considering that most obstetrical complications seen today result in part, or entirely as a result of placental dysfunction, the ability to identify and properly diagnose the placental cause of poor pregnancy outcomes is a priority [28]. The field of clinical placental pathology aims to address this need, using visual anatomical and histological cues in the placenta as a permanent record of intra-uterine damage and disease across pregnancy. The importance of this work cannot be overlooked, as understanding what happened to the placenta helps clinicians manage their patients in the postpartum/neonatal period and can help council patients regarding what went wrong during their pregnancy and what the risk of recurrence might be in subsequent pregnancies [28]. Despite the importance of this practice, the field of placental pathology has been plagued by a lack of standardization in diagnostic criteria, and the high degree of specialized training required to perform these time-consuming exams often leads to poor inter-pathologist agreement scores on diagnostic outcomes [45]. As such, this field is ripe for advancements aimed at reducing workload and simultaneously improving quality and reproducibility of this practice.

Artificial intelligence and machine learning are rapidly becoming integral to effective healthcare practice [46]. One medical field in which the application of these tools demonstrate considerable promise is clinical pathology diagnostics [46]. This potential is highlighted by success in the area of oncology, where machine learning has been applied with a high degree of accuracy in the detection of malignant vs benign tumors using histological specimens [47]. To date, this same approach has not been attempted in placental pathology. Our group has very recently cracked the surface, for the first time examining the potential of supervised machine

learning to identify placentas from pregnancies with and without PE (unpublished, Mukherjee et.al). However, this preliminary work highlighted the notion that classifying placentas strictly based on the final pregnancy outcome (i.e., PE vs. healthy) was insufficient to address the high degree of heterogeneity in the types of placental disease that may lead to the same clinical outcome (unpublished, Mukherjee et al.). As such, the current project aims to instead apply machine learning algorithms using classifiers reflective of the placental disease present rather than the outcome. Specifically, we aim to focus on a form of placental disease commonly seen in cases of PE and FGR – maternal vascular malperfusion.

2.2 Hypothesis

The overarching hypothesis of the current project is that by using machine learning we will be capable of easily and accurately classifying digital histological images of placentas with lesions of maternal vascular malperfusion, from patients with both PE and FGR similarly. We hypothesize that using a supervised machine learning approach, digital histological images of placentas with evidence of maternal vascular malperfusion will be accurately classified.

2.3 Research Aims

The **research aims** of this MSc thesis will aim to:

1. Work with a perinatal pathologist to blindly re-classify a series of 166 digital histological images of placental specimens from healthy control pregnancies and pregnancies afflicted with PE, FGR, or both, according to the presence or absence of lesions consistent with maternal vascular malperfusion.

2. Perform data augmentation on the digital image set to enhance the size of the training sets of images.

3. Test and compare three supervised machine learning models, differing according to methods of feature extraction, on a digital placental histopathology image set using a support vector machine (SVM) algorithm and determine the accuracy of classification for lesions of maternal vascular malperfusion.

Chapter 3

Automated Detection of Maternal Vascular Malperfusion Lesions of the Placenta using Machine Learning

Purvasha Patnaik[1,↟], Goutham Vasam[1,↟], Anika Mukherjee[1,2,3], Sina Salsabili[4], Eranga Ukwatta[4,5], David Grynspan[2,3,6], Adrian D. C. Chan[4], Shannon Bainbridge[1,7,*]

[1]Interdisciplinary School of Health Sciences, Faculty of Health Sciences, University of Ottawa, Ottawa, ON, Canada
[2]Department of Pathology and Laboratory Medicine, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
[3]Children's Hospital of Eastern Ontario, Ottawa, ON, Canada
[4]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada
[5]School of Engineering, University of Guelph, Guelph, ON, Canada
[6]Department of Pathology and Laboratory Medicine, Faculty of Medicine, The University of British Columbia, Vernon Jubilee Hospital, Vancouver, BC, Canada
[7]Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, ON, Canada

↟Joint first authorship
***Corresponding author:**
Shannon Bainbridge, Ph.D.
Associate Professor
Interdisciplinary School of Health Sciences, Faculty of Health Sciences
Department of Cellular and Molecular Medicine
University of Ottawa
2058, RGN building,
451 Smyth Road, Ottawa, ON, K1H 8M5

This manuscript was formatted for submission to the peer-reviewed journal *American Journal of Obstetrics & Gynecology.*

Abstract

Preeclampsia (PE) and fetal growth restriction (FGR) are common obstetrical complications, often with pathological features of maternal vascular malperfusion (MVM) in the placenta. Current placental clinical pathology methods involve a manual visual examination of histology sections, a practice that can be resource-intensive and demonstrate moderate-to-poor inter-pathologist agreement on diagnostic outcomes, dependant on the degree of pathologist sub-specialty training. The visual patterns of placental damage in MVM specimens can be used as features for machine learning (ML), which can minimize shortcomings of manual diagnosis. This study aims to apply different ML feature extraction methods to classify digital images of placental histopathology specimens, collected from PE, FGR, PE + FGR, and healthy pregnancies, according to the presence or absence of MVM lesions. 166 digital images were captured from histological placental specimens, manually scored for MVM lesions (MVM- or MVM+) and used to develop various support vector machine (SVM) classifier models, differing in feature extraction methods. Classification performance of each model was assessed through accuracy, precision, recall, specificity, F1-score, and AUC scores using confusion matrices. SVM models demonstrated accuracies between 47-73% in MVM classification, with poorest performance observed on images with borderline MVM presence, as determined through manual observation. The results are promising for the integration of ML methods into the placental histopathological examination process.

Keywords: support vector machine, placenta, preeclampsia, fetal growth restriction, maternal vascular malperfusion, supervised machine learning, computer-aided diagnosis, digital histopathology, image classification, whole-slide imaging

3.1 Introduction

The development and function of the placenta are essential to a healthy pregnancy outcome. When placental health becomes compromised, an umbrella of 'placenta-mediated diseases' can arise, including two of the most common obstetrical complications: preeclampsia and fetal growth restriction [1]. Preeclampsia (PE) is defined as de novo onset of hypertension after 20 weeks of gestation, coupled with the evidence of maternal end-organ dysfunction [2]. Fetal growth restriction (FGR) is the failure of a fetus to achieve its genetic growth potential in utero [3], typically diagnosed when a fetus falls below a critical size threshold (typically <5-10%ile) with some additional evidence of placental contribution, such as abnormal uterine artery Doppler measurements [4]. While the exact pathophysiology of these two obstetrical complications is not entirely known, a common finding of compromised uteroplacental blood flow has been widely described in both conditions [1].

In a healthy pregnancy, a subset of the placental trophoblast cells is tasked with invading the uterine wall and remodeling the uterine spiral arteries into large diameter vessels that are no longer responsive to vasopressors in the maternal circulation. If this process is disrupted, blood flow into the placenta can become compromised resulting in hypoxia-related placental injuries – observed visually as lesions of Maternal Vascular Malperfusion (MVM). It is estimated that ~ 60% of all cases of PE and FGR demonstrate evidence of placental MVM lesions at the time of delivery [1], indicative of insufficient uteroplacental blood flow as a significant contributory factor to the disease pathogenesis.

A diagnosis of PE and/or FGR at the time of delivery is an indication for submission of the placenta for pathological examination. Current practice includes a gross evaluation of the entire organ, as well as a histological evaluation of ~ four full-thickness sections of the placenta for the presence/absence and severity of a variety of lesions known to result from developmental abnormalities, damage done by poor uteroplacental perfusion, maternal metabolic disorders, infectious pathogens, or chronic inflammation. These examinations, when done correctly, can be time and resource-intensive and are ideally performed by a specialty-trained perinatal pathologist.

Over the years, perinatal pathologists have developed a detailed set of criteria that best determine the presence of MVM [5]. Microscopic placental lesions consistent with an MVM diagnosis have distinct visual appearances (Figure 3.1) and include: placental infarcts, distal villous hypoplasia, accelerated villous maturation, increased syncytial knots, and villous agglutination (Figure 3.1) [6]. Collectively, these MVM lesions demonstrate distinct visual patterns, and their detection in a clinical pathology exam is indicative of compromised placental development during the pregnancy. These examinations lead to appropriate diagnosis of placenta-mediated disease and serve as important components of maternal and neonatal care. Unfortunately, this clinical practice has historically been plagued by a lack of standardization in employed terminology and diagnostic criteria, limiting the clinical utility. Despite landmark efforts undertaken to improve global standardization in this field, the degree of inter-pathologist variability in examination findings remains high, particularly in low resource settings where specialized training and expertise are lacking [6]. Considering these challenges, placental pathology is a key clinical domain that may benefit most from the implementation of computer-

aided diagnosis (CAD) through the development of machine learning algorithms capable of identifying distinct visual patterns in placental histopathology specimens.

Computer vision, or computer-driven image recognition, is the ability of a computer to interpret and categorize images according to pre-defined labels, such as clinical diagnosis [7]. Through machine learning, biomedical advances in computer vision have been progressing rapidly by using biomedical image datasets for recognition and classification. Machine learning (ML) is the development of automated systems that can process large amounts of data to execute data mining and decision support [8]. Supervised ML is the process of feeding the computer a dataset of images that have been pre-labeled [8]. The computer then 'learns' the features and patterns of the images, which can be later tested using a new image dataset with unseen images that are not labeled [8]. The aim is for the computer to be able to label these unseen and unlabelled images with a high degree of accuracy.

As the first step in this direction, the current study aims to develop and test supervised ML algorithms that can identify and classify digital images of placental histopathology specimens according to the presence or absence of MVM in cases of PE and/or FGR and healthy samples. Three different algorithms were developed and compared, in which the method of feature extraction from the images differed. The results of this study will serve as an important proof-of-concept, demonstrating the potential of CAD to improve the accessibility to consistent and high-quality diagnostic outcomes of the placenta pathology exam while reducing the resource-intensive workload of this important clinical practice, ensuring that mothers and infants can receive the best possible healthcare afforded to them.

3.2 Materials and Methods

3.2.1 Digital Image Dataset Acquisition and Patient Characteristics

This study has received research ethics board approval from the University of Ottawa Research Ethics Board (H-08-18-1023) and Mount Sinai Hospital (13-0212-E) (Appendix A). Original paraffin-embedded placenta histopathology specimens of villous tissue samples were purchased from the Research Centre for Women's and Infants' Health (RCWIH) Biobank at Mount Sinai Hospital (Toronto, ON). A total of 166 specimens were collected from patients with PE (N=80 total; N= 47 PE alone, N= 33 PE + FGR), patients with normotensive FGR (N=20), and healthy controls (n=66), with each paraffin block containing four chorionic villous biopsies from each quadrant of the placenta (Table 3.1). PE was defined as the onset of hypertension (systolic pressure $\geq$ 140 mm Hg and/or diastolic pressure $\geq$ 90 mm Hg) after 20 weeks of gestation coupled with evidence of maternal end-organ dysfunction [9]. FGR was defined as a neonatal birthweight <10^{th} percentile for gestational age and sex, with evidence of placental dysfunction in histopathology specimens [4]. Sectioned tissue was mounted on glass slides and stained following a standard histopathology protocol with Hematoxylin and Eosin (H&E) at the Department of Pathology and Laboratory Medicine at the Children's Hospital of Eastern Ontario (CHEO). Slides were scanned at 20X magnification (0.5 micron/pixel) with an Aperio Scanscope digital scanner and saved in '.svs' format.

3.2.2 MVM Labelling of Images

All collated digital images underwent detailed histopathology examination by a perinatal pathologist (D. Grynspan, CHEO) blinded to all clinical information aside from gestational age at delivery and using a synoptic data collection tool, which groups placental lesions according to broad placental disease categories [10]. The first category included in this tool is MVM, which includes the following lesions: placental infarcts, distal villous hypoplasia, accelerated villous maturation pattern, syncytial knots, and villous agglutination (Figure 3.1) [6]. The pathologist scored each digital image for the absence (score = 0), presence (score = 1) and severity (if applicable, score = 2) of each MVM lesion. In addition to individual scores for each lesion, the pathologist generated a cumulative MVM score for the whole-slide image (WSI), by summing the pathology scores of all MVM lesions included in the synoptic tool (maximum score = 8). As many healthy placentas may demonstrate mild evidence of one or more of these lesions, which may not reflect true placental disease, all placentas which demonstrated a cumulative MVM score of ≥ 3 were classified as being diagnosed with MVM (MVM+). All placentas with a cumulative MVM score between 0-2 were classified as not having MVM (MVM+), both labels being irrespective of clinical outcome (PE or FGR). This classification threshold was determined based on the clinical guidance of the perinatal pathologist.

3.2.3 Supervised Machine Learning Model Development

Three different ML models were developed using support vector machine (SVM) algorithms, each differing in the method in which the features within each digital image were identified and extracted. The SVM algorithm is commonly used in binary classification schemes

and has been applied in the past to a diverse range of classification problems. Therefore, SVM was selected for this proof-of-concept study due to its potential ease of use in a clinical setting, requiring only modest computing power for execution. In model #1, the images underwent automated segmentation using a previously published method to identify all individual placental villi [11], and five *a priori* determined biological features of relevance (see below) were extracted from each segmented image. In model #2, the images underwent automated feature extraction using scale-invariant feature transform (SIFT). In model #3, the images underwent automated feature extraction using SIFT features in a visual bag of words model. A comparison between model development on the original dataset vs. an augmented dataset was performed. All coding used for image pre-processing, data augmentation, and ML model development and testing was scripted in Python 3 language (Table 3.2).

3.2.4 Image Pre-processing

The image dataset was pre-processed from 166 WSIs captured in '.svs' format from 166 patients. One or more images were cropped from each slide using Aperio ImageScope software, minimizing the amount of empty/white space included in the image and separating the tissue samples into separate images if there were multiple samples present from the same patient on the slide. These 263 cropped images underwent patch extraction using the 'sliding window' approach [12]. Briefly, image patches were extracted using a window of size 1000 x 1000 pixels (px) sliding from the top left corner to the right by 500 px and down by 500 px of the cropped images (Figure 3.2). The total number of patches extracted from one cropped image ranged from 3 – 328 patches, dependent on the amount of tissue on the WSI. This image dataset was

filtered to only include patches that had 500 or more SIFT keypoints to ensure the images had enough placental tissue present to provide sufficient data features. The final image dataset had a total of 17,364 image patches. Each image patch was converted to RGB and underwent colour normalization using OpenCV (Table 3.2).

3.2.5 Data Splitting and Augmentation

The image patch dataset was split into a training and testing set using the GroupShuffleSplit function (Table 3.2). The testing and training datasets were balanced according to the manual MVM diagnosis of the WSI (i.e. per patient) and all patches derived from each patient were assigned to either the test *or* training set to avoid bias in model performance (Table 3.3). The training and test set consisted of 13,349 images (approximately 75% of the images) and 4015 images (approximately 25% of the images), respectively (Table 3.3). The training set was used to train the models to identify patterns associated with the provided MVM label, whereas the testing set was fed into the models unlabelled (Table 3.3). Following splitting, images within the training dataset underwent data augmentation using Tensorflow (Table 3.2). This technique manipulates the patch images to produce numerous iterations of each patch image in the dataset [13]. Data augmentation techniques included shear by 0.2%, zoom by 0.2%, random rotations in a range of 0-30 degrees, random brightness in a range of 0.8-1.2, channel shift in a range of 0-150, horizontal flips, vertical flips, and fill mode set to 'reflect'. The data augmented training dataset therefore comprised of the original image in addition to three versions of the image to which the data augmentation techniques were randomly applied. Using this technique, our training image dataset was increased from

13,349 original image patches (approximately 75% of patches derived from the original WSIs) to 39,076 image patches (Table 3.3). For each model developed, classification performance was assessed when an augmented training set was used compared to the original training set. Considering large training datasets are vital to the development of a robust ML algorithm, it was hypothesized that this data augmentation step would increase model performance.

3.2.6 Feature Extraction

Model #1 - A Priori Determined Biologically Relevant Feature Extraction

Each image from the training and testing set underwent automated image segmentation to identify individual villous structures, as previously described by our group [11]. Once segmented, five pre-determined biological features of the villous exchange unit of the placenta were quantified using MATLAB: average villi size, average villi density, average villi circularity, average syncytial knot density, and villi count. These features were previously deemed to be of high biological relevance to MVM disease [14] and were used as the feature set used for training in this model.

Model #2 - Automated SIFT Feature Extraction

Feature extraction using the SIFT algorithm was applied to the training and testing set for extracting raw attribute values, known as keypoints, to describe edges, corners, blobs, and ridge detection that can be converted to feature vectors for the computer to analyze [15]. 500 SIFT keypoints were extracted, based on previously published protocols for histopathological

image data analysis [16]. Each of the 500 SIFT keypoints contains a 128-dimensional feature vector known as descriptors. The 500x128 feature matrix was flattened to a 1-dimensional feature vector by using the medians of the 128-dimensional feature vectors. These 1-dimensional feature vector sets were used for training in this model.

Model #3 - Automated Visual Bag of Words

The 'visual bag of words' model [17] treats image features like words [18], generating a codebook in which similar 'words' or features can be clustered. In keeping with published protocols, feature extraction using the SIFT algorithm was applied to the training and testing set, extracting 500 keypoints per image [16]. K-means clustering was applied to the 500 SIFT feature descriptors to capture the visual words [19]. The k-means clustering algorithm split our given labelled data set into a fixed number (k) of clusters. A fixed number of 500 k-means clusters for the original dataset and 600 k-means clusters for the augmented dataset was selected after testing the models with different k-values and comparing their accuracies [20]. The centroids of the clusters are considered the visual words, and the code vectors were extracted based on how close the local descriptors were to the centroid [21]. The output was a frequency histogram of vectors that represent the whole image, pooling the code vectors of the individual descriptors into one, which has a length of k [21].

3.2.7 Model Development and Testing

To apply equal weight to the magnitude of each feature before being passed into the ML model, extracted features from all three feature extraction models were normalized using a

Min-Max Scaler (Table 3.2). Features extracted via all methods with and without data augmentation (the variables X_train represented features for training, X_test represented features for testing) and their corresponding labels (the variables y_train to train the model, y_test to calculate the true positives, true negatives, false positives and false negatives against the predictions) were organized into NumPy arrays and fed through an SVM using scikit-learn by the SVC function with an 'RBF' kernel (Table 3.2).

3.2.8 Classification Performance Metrics

The classification results obtained when the test image set was analyzed by each of the three developed models were based on the ground truth labeling scheme of MVM+ and MVM-, as scored by the blinded perinatal pathologist. Classification performance metrics were visualized using normalized confusion matrices to provide the ratios of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Figure 3.3). Weighted averages for accuracy, precision, recall (sensitivity), and F1-score were obtained using classification reports (Table 3.4), and specificity was manually calculated for each model. Weighted averages for each performance metric were calculated to account for any class imbalance observed in the training and testing sets. Area under the curve (AUC) values were also obtained from scikit-learn, measuring how well the model can distinguish between the binary classes.

The following equations represent how these values were calculated:

(1) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

(2) $Precision = \frac{TP}{TP+FP}$

(3) $Recall\ (Sensitivity) = \frac{TP}{TP+FN}$

(4) $Specificity = \frac{TN}{TN+FP}$

(5) $F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$

(6) $weighted_x = \frac{(x_{c1} * |c1|)+(x_{c2} * |c2|)}{|c1|+|c2|}$, *where x is precision, recall(sensitivity), specificity, or F1-Score per class.*

A comparison between the classification of each feature extraction model was performed, along with a comparison of the performance of each model with and without an augmented training set. A sub-analysis was also performed to assess which sample characteristic was most often associated with image misclassification in each of the three developed models, specifically examining the effect of raw MVM score (from 0 to 8) and obstetrical outcome (PE, FGR, PE+FGR, Control) on classification performance (Figure 3.4).

3.3 Results

3.3.1 Sample Demographics and Class Imbalance

Patient demographics according to MVM diagnosis are summarized in Table 3.1. Equal number of patients with MVM- and MVM+ were randomly assigned to either the training or testing set, however due to differences in the number of patches derived from each patient WSI, a minor imbalance was observed in the testing set (Table 3.3), with 2359/4015 (59%) MVM- patches and 1656/4015 (41%) MVM+ patches. Unfortunately, due to the random allocation of images according to patient MVM status, no FGR cases labelled as MVM+ were found within the testing set.

3.3.2 Classification Performance Metrics

Confusion matrices and performance metrics are documented by the feature extraction model tested and training dataset type (original vs. augmented) in Fig 3.3, Table 3.4, and Fig S1. Collectively, SVM models demonstrated MVM classification with moderate to good accuracies, ranging from 61-73% (Fig 3.3, Table 3.4, Fig S1). Feature extraction model #2 demonstrated the highest degree of classification accuracy (73%), highest F1-score (73%), and highest AUC (77%) of all models tested, irrespective of whether the testing dataset was augmented or not (Table 3.4, Fig S2).

3.3.3 Model #1 - A Priori Determined Biologically Relevant Feature Extraction

The first model tested, using extracted *a priori* biological features, demonstrated fair classification performance (accuracy = 66%; F-1 = 62%; AUC = 75%) when the original dataset was used for training. Surprisingly, the classification performance decreased (accuracy = 47%; F-1 = 48%; AUC = 50%) when the augmented training set was used in model development (Fig 3.3 A-B; Table 3.4). Misclassification in this model was driven in large part by false negatives [False negative rate (FNR) =72% original, FNR = 61% augmented], with the model performing well for the classification of MVM- images (specificity, original = 66%, augmented = 47%) and for MVM+ images (recall/sensitivity, original = 66%, augmented = 47%). When the model was run with the original training set, false negative misclassification was highest in cases with PE and PE + FGR diagnosis (Fig 3.4 B, D). When the model was run with the augmented training set, false positive misclassification of images was highest when the MVM cumulative score was 0, in cases with normotensive FGR and control samples (Fig 3.4 A, C). In this same model, false negative

misclassification was highest when the cumulative MVM score hovered close to the threshold cut-off value (score = 3) and in cases with PE and PE + FGR diagnosis (Fig 3.4 B, D).

3.3.4 Model #2 - Automated SIFT Feature Extraction

The second model tested, using automated SIFT feature extraction, demonstrated the highest classification performance, with an accuracy and F1-score of 73% and an AUC of 77%, indicating the model classified image data well considering any class imbalances found in the training dataset. All model performance metrics were unchanged by training set augmentation (Fig 3.3 C-D; Table 3.4). This model demonstrated good classification of MVM- images (specificity = 73%) and modest classification of MVM+ images (sensitivity/recall = 73%), with the model slightly overcalling the MVM- label (false negatives). For both the original and augmented training set models, the false negative misclassification of images was highest for images with the highest cumulative MVM scores (most severe form of the disease) and in cases with PE and PE + FGR diagnosis (Fig 3.4 B, D).

3.3.5 Model #3 - Automated Visual Bag of Words

The third model tested, using an automated visual bag of words feature extraction, demonstrated a modest classification performance (accuracy = 61%, F1-score = 61%, AUC = 65%), with similar classification performance with training set augmentation (accuracy = 63%, F1-score = 61%, AUC = 66%) (Fig 3.3 E-F, Table 3.4). Misclassification in this model was equally distributed across false positives and false negatives, with the model performing equally well in

the classification of MVM+ and MVM- images (recall/sensitivity: original = 61%, augmented = 62%; specificity: original = 61%, augmented = 62%). False positive misclassification of images was highest when the MVM scores hovered near the classification threshold (score = 1,2) and in cases of normotensive FGR (Fig 3.4 A, C), whereas false negative misclassification was highest at the cumulative MVM score extremes (close to threshold, score = 3-4; severe pathology, score = 7-8) and in cases with PE and PE + FGR diagnosis (Fig 3.4 B, D).

3.4 Discussion

In the current proof-of-concept study, three different supervised ML models were tested and compared for their ability to classify placental histopathology images according to the presence/absence of MVM lesions – placental damage associated with poor utero-placental perfusion across pregnancy. Collectively, the developed models demonstrated moderated to good MVM classification performance. The use of an SVM model that employs an automated SIFT feature extraction method demonstrated the strongest classification performance, and surprisingly the method of data augmentation used in the current study did not enhance the performance of any developed model. A sub-analysis of misclassified images unveiled the underperformance of the models when the cumulative MVM scores straddled the ground truth cut-off threshold and at the extreme ends of the cumulative score spectrum. Expectedly, the misclassified images more frequently originated from cases of PE and PE + FGR – obstetrical conditions which are known to be highly heterogeneous in clinical presentation and placenta pathology.

Model #2 provided the highest MVM classification performance metrics of all three models tested, likely because of the feature extraction method employed. In this model, the medians of 500 SIFT descriptors were extracted from each image patch, generating many features that densely cover the entire image. SIFT features are considered highly distinctive, making it easier for the model to match these unique points across images in the training and testing sets. Although SIFT features were also used in the visual bag of words model (model #3), our results reflect that using the SIFT descriptors medians directly as the features worked best. The visual bag of words model involves a strict assignment of a feature to a codeword, and there is information loss associated with this encoding process [22]. Additionally, the higher the number of codewords, the more computation effort it takes for codebook generation. This model is also often criticized for ignoring the spatial arrangement of features within the images [22]. The poorer performance of model #1, in comparison to model #2, may in large part be driven by the inherent biases associated with *a priori* selection, by humans, of biological features of relevance. While the structural features chosen may reflect logical visual cues associated with our current understanding of the pathophysiology of the disease, this method likely underestimated the high degree of variability both in anatomical structure of the placenta and the heterogeneity of placental disease. For example, using average villi density as a feature to define an MVM lesion (such as villous agglutination) requires the assumption that normally the villi are relatively diffuse across the placental tissue. However, this is not necessarily the case from a clinical standpoint; villi can be observed as mildly clustered in healthy placenta. Ultimately, model #2 demonstrated to the best classification performance, and future work

should be focused on further optimizing the training of the model to minimize misclassifications.

In the sub-analysis performed to determine misclassified image characteristics, it was determined that misclassified images often had a manual cumulative MVM score at either extreme – either low scores that straddled the manual MVM+ diagnosis threshold (score = 2-4) or high scores indicative of severe placental disease (score = 7-8). When interpreting these results, it is important to acknowledge that the ground truth MVM+ diagnosis threshold used in the current study to convert a continuous MVM score variable into binary classes (cumulative score $\geq$3) was selected by a perinatal pathologist, based on his clinical expertise and experience. In reality, placental disease demonstrates a spectrum of lesion presence and severity, with increasing pathology more likely associated with adverse impacts on the mother and fetus. As such, it is recognized that the selected threshold is subjective, and it was not surprising that those images with scores surrounding this threshold were more likely to be misclassified. Future work in this area should specifically address this study limitation, either through comparison of model performance at different MVM+ thresholds and/or moving past a binary classifier, rather attempting to develop a multi-class model that may better capture the spectrum of pathology seen clinically. On the other end of the spectrum, cases with high MVM scores indicative of severe placental disease were also more often misclassified. This finding may simply be the result of a small sample size, in which only 6% (N=11/166) of the original WSIs demonstrated this severe form of placental disease. Further, a post-hoc analysis revealed a modest class imbalance of these high MVM score images between the training and testing set, with only 2% of the training set containing an MVM score of 7-8 while 12% of the testing

set had these elevated scores. Future studies should be designed to include larger sample sizes with equal numbers of MVM cases across the pathology spectrum, with specific attention to balancing for this characteristic at the dataset splitting stage.

When assessing the performance of these models, one must also consider the issue of heterogeneity of pathology across the placenta. Some MVM lesions, such as distal villous hypoplasia, can certainly be present in a diffuse pattern across the placenta, and hence will be present in both the original WSI and all digital image patches used for model training. However, other MVM lesions, such as placental infarcts, may be more localized in nature and while the WSI may be classified as MVM+ due to the presence of this lesion, it may not in fact be visible in all or even many of the digital image patches used to train the models. The degree of heterogeneity of MVM lesions across the original WSI gives the possibility for the mislabeling of patches, leading to poorer training and hence performance of the developed model. Additional image features which may compromise model training could include imaging artifacts or remnants of maternal red blood cells within the intervillous space, which could both skew feature extraction and affect the prediction of that image patch [11][23].

In addition to the MVM score characteristic, the obstetrical outcome characteristic of each image was also associated with image misclassification by the developed models. Specifically, false negative misclassification was most often observed for patches originating from PE and PE+FGR specimens, for all tested models. From a clinical perspective, false negatives are a concern, as this would result in patients at highest risk of MVM in future pregnancies being missed. As such, future work must be focused on this area for improvement. A plausible explanation for these misclassifications is the high degree of heterogeneity of

placental pathology observed in cases of PE and PE + FGR, with the likelihood of multiple disease processes (diverse combination of lesions) being present. Further, as discussed above, some lesions associated with these complications may be focal in nature, leading to several MVM+ labelled image patches with no disturbance to the visual features present. To tackle this problem moving forward, an effort to curate a larger image bank of placental disease specimens will be required to ensure robust training of newly developed models. Further, attempting to re-group linked patches (from the same original WSI) and perform a weighted classification strategy may be warranted.

Unexpectedly, there was no improvement in the classification performance of any developed model with the added element of data augmentation. This was surprising given the applied data augmentation method employed is widely described as a method to boost model training and classification performance [21]. Although reasons remain unclear for this finding, particularly in the case where model performance decreased with augmentation (model #1), an increased number of similar-kind training data points from the 'reflect' type of fill mode of the data augmentation might have distorted the support vectors and thereby the hyperplane such that there was a suboptimal separation between two binary classes after data augmentation [24]. For method #2, which uses 500 SIFT descriptor medians as the input, data augmentation did not appear to have a strong effect in either direction. One explanation of this could be that the SIFT algorithm itself is invariant to some of the methods of augmentation used (ex. flips, random rotations), however, it is unlikely to be invariant to all forms of augmentation (ex. changes in intensity). This implies that by using the non-augmented model, you can get similar results without the extra computational force needed to train augmented images, even with an

imbalanced dataset. An alternate explanation is that some methods of data augmentation used may have altered the extracted features to an extent in which the features are no longer reflective of true MVM pathology, causing no boost in classification performance with augmentation. This study contained traditional data augmentation techniques that are typically used in deep learning models. One approach to enhancing the data augmentation technique would be to augment certain images more than others across the MVM scores to tackle class imbalance. Accordingly, a different type of data augmentation that can be explored is changing the intensities of the RGB channels or being selective with the data augmentation techniques such as removing flips and random rotation in case these augmentations are invariant to SIFT features [25]. Therefore, a modification to this model would involve exploring the methods of data augmentation so it will noticeably add to the strength of the accuracy.

As a first foray, the results of the current study certainly highlight the potential utility of ML models to automate aspects of placental histopathology evaluations, while identifying aspects of image pre-processing and model optimization that can be explored further to bring this modality to the forefront. Using this study as a foundation to start, our group and others can aim to continue improving these models and validating them for clinical use.

3.5 Limitations

The success of ML model development for use with clinical imaging data has in large part been driven by access to robust training datasets. In domains of clinical pathology where the biggest gains in ML implementation have been made (i.e. onco-pathology), successful model development has been achieved using training datasets that include 500 – 10,000

original WSIs, which have then been augmented for training purposes [21][26][27]. Comparatively, the digital image sample sizes used in the current study are quite small (166 WSIs). Compounding the issue of sample size is the high degree of heterogeneity in clinical presentation and underlying placental pathology in cases of PE, PE + FGR, and FGR. Collectively, this results in a training set that is small with varying presentations of disease, likely hampering the ability to develop a robust ML model. While not a serious limitation, as discussed above, the dataset used did demonstrate an element of class imbalance, where the image patches in training and testing datasets were not equally distributed across MVM scores and obstetrical outcomes. To address the above-mentioned limitations, the clinical and research communities must work together to generate, and make available for research purposes, a robust clinical image dataset of reproductive tissues, including the vital organ of pregnancy - the placenta. Access to such a research-ready image database would allow for timely and robust ML model development to address clinical needs unique to the health and wellbeing of mothers and their offspring.

3.6 Future Work

The current study not only shows promise in using supervised ML in placental pathology but also uncovers the next steps that should be taken to improve the models. While continued work to further advance traditional ML models such as those tested in the current study is warranted, additional investigation into the use of deep learning models may also yield important contributions to this field. Convolutional neural networks (CNN) are an ideal model for image classification that learns from discerning features rather than handcrafted features as

a part of its training process, and transfer learning has further shown improvement in accuracy rates [28][29][30]. We anticipate that in future studies, a CNN will provide us with results of higher accuracy than the SVM, as CNNs are known to perform well for image analysis (see preliminary data in Appendix B). However, it is important to consider that each model has its pros and cons; where CNNs are prevalent in object recognition and requires a much larger dataset than SVMs, SVMs are highly prevalent in classification analysis [31]. Therefore, testing and comparing different paradigms in computer vision such as classical ML (ex. SVM, random forest), deep learning (CNNs) with and without transfer learning (ex. VGG-16, ResNet-50) is essential to gaining perspective on the most optimal ways these images can be classified. The ability to automate the detection of MVM and other forms of placental disease would open the door to the integration of ML approaches into the placental pathology workflow in the future. It would also be a proof-of-concept in support of performing similar models for other forms of placental disease.

References

[1] E. Wright *et al.*, "Maternal Vascular Malperfusion and Adverse Perinatal Outcomes in Low-Risk Nulliparous Women:," *Obstet. Gynecol.*, vol. 130, no. 5, pp. 1112–1120, Nov. 2017, doi: 10.1097/AOG.0000000000002264.

[2] American College of Obstetricians and Gynecologists and American College of Obstetricians and Gynecologists, Eds., *Hypertension in pregnancy*. Washington, DC: American College of Obstetricians and Gynecologists, 2013.

[3] L. M. M. Nardozza *et al.*, "Fetal growth restriction: current knowledge," *Arch. Gynecol. Obstet.*, vol. 295, no. 5, pp. 1061–1077, May 2017, doi: 10.1007/s00404-017-4341-9.

[4] I. Gibbs, K. Leavey, S. J. Benton, D. Grynspan, S. A. Bainbridge, and B. J. Cox, "Placental transcriptional and histologic subtypes of normotensive fetal growth restriction are comparable to preeclampsia," *Am. J. Obstet. Gynecol.*, Oct. 2018, doi: 10.1016/j.ajog.2018.10.003.

[5] R. W. Redline, "Classification of placental lesions," *Am. J. Obstet. Gynecol.*, vol. 213, no. 4, pp. S21–S28, Oct. 2015, doi: 10.1016/j.ajog.2015.05.056.

[6] T. Y. Khong *et al.*, "Sampling and Definitions of Placental Lesions: Amsterdam Placental Workshop Group Consensus Statement," *Arch. Pathol. Lab. Med.*, vol. 140, no. 7, pp. 698–713, Jul. 2016, doi: 10.5858/arpa.2015-0225-CC.

[7] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, 2010.

[8] I. G. Maglogiannis, *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press, 2007.

[9] S. J. Benton, K. Leavey, D. Grynspan, B. J. Cox, and S. A. Bainbridge, "The clinical heterogeneity of preeclampsia is related to both placental gene expression and placental histopathology," *Am. J. Obstet. Gynecol.*, vol. 219, no. 6, p. 604.e1-604.e25, Dec. 2018, doi: 10.1016/j.ajog.2018.09.036.

[10] S. J. Benton, A. J. Lafreniere, D. Grynspan, and S. A. Bainbridge, "A synoptic framework and future directions for placental pathology reporting," *Placenta*, Jan. 2019, doi: 10.1016/j.placenta.2019.01.009.

[11] S. Salsabili, A. Mukherjee, E. Ukwatta, A. D. C. Chan, S. Bainbridge, and D. Grynspan, "Automated segmentation of villi in histopathology images of placenta," *Comput. Biol. Med.*, vol. 113, p. 103420, Oct. 2019, doi: 10.1016/j.compbiomed.2019.103420.

[12] P. Yang *et al.*, "GSWO: A programming model for GPU-enabled parallelization of sliding window operations in image processing," *Signal Process. Image Commun.*, vol. 47, pp. 332–345, Sep. 2016, doi: 10.1016/j.image.2016.05.003.

[13] N. G. Polson and S. L. Scott, "Data augmentation for support vector machines," *Bayesian Anal.*, vol. 6, no. 1, pp. 1–23, Mar. 2011, doi: 10.1214/11-BA601.

[14] L. M. Ernst, "Maternal vascular malperfusion of the placental bed," *APMIS*, vol. 126, no. 7, pp. 551–560, Jul. 2018, doi: 10.1111/apm.12833.

[15] J. Lampinen and E. Oja, “Distortion tolerant pattern recognition based on self-organizing feature extraction,” *IEEE Trans. Neural Netw.*, vol. 6, no. 3, pp. 539–547, May 1995, doi: 10.1109/72.377961.

[16] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, “A Dataset for Breast Cancer Histopathological Image Classification,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, Jul. 2016, doi: 10.1109/TBME.2015.2496264.

[17] J. Liu, S. Zhang, W. Liu, C. Deng, Y. Zheng, and D. N. Metaxas, “Scalable Mammogram Retrieval Using Composite Anchor Graph Hashing With Iterative Quantization,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 11, pp. 2450–2460, Nov. 2017, doi: 10.1109/TCSVT.2016.2592329.

[18] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual Categorization with Bags of Keypoints,” p. 16, 2004.

[19] L. Fei-Fei and P. Perona, “A Bayesian hierarchical model for learning natural scene categories,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, Jun. 2005, vol. 2, pp. 524–531 vol. 2. doi: 10.1109/CVPR.2005.16.

[20] J. Feng, Y. Liu, and L. Wu, “Bag of Visual Words Model with Deep Spatial Features for Geographical Scene Classification,” *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–14, 2017, doi: 10.1155/2017/5169675.

[21] S. Winters-Hilt, A. Yelundur, C. McChesney, and M. Landry, “Support vector machine implementations for classification & clustering,” *BMC Bioinformatics*, vol. 7 Suppl 2, p. S4, Sep. 2006, doi: 10.1186/1471-2105-7-s2-s4.

[22] H. Chougrad, Z. Hamid, and A. Omar, “Bag of Features Model Using the New Approaches: A Comprehensive Study,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, Jan. 2016, doi: 10.14569/IJACSA.2016.070132.

[23] D. Komura and S. Ishikawa, “Machine Learning Methods for Histopathological Image Analysis,” *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 34–42, 2018, doi: 10.1016/j.csbj.2018.01.001.

[24] D. Bardou, K. Zhang, and S. M. Ahmad, “Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks,” *IEEE Access*, vol. 6, pp. 24680–24693, 2018, doi: 10.1109/ACCESS.2018.2831280.

[25] J. Fan, J. Lee, and Y. Lee, “A Transfer Learning Architecture Based on a Support Vector Machine for Histopathology Image Classification,” *Appl. Sci.*, vol. 11, no. 14, Art. no. 14, Jan. 2021, doi: 10.3390/app11146380.

[26] L. Ma, X. Liu, L. Song, C. Zhou, X. Zhao, and Y. Zhao, “A new classifier fusion method based on historical and on-line classification reliability for recognizing common CT imaging signs of lung diseases,” *Comput. Med. Imaging Graph.*, vol. 40, pp. 39–48, Mar. 2015, doi: 10.1016/j.compmedimag.2014.10.001.

[27] M. A. Kahya, W. Al-Hayani, and Z. Y. Algamal, “Classification of Breast Cancer Histopathology Images based on Adaptive Sparse Support Vector Machine,” p. 22.

[28] N. Bayramoglu and J. Heikkilä, “Transfer Learning for Cell Nuclei Classification in Histopathology Images,” in *Computer Vision – ECCV 2016 Workshops*, Cham, 2016, pp. 532–539. doi: 10.1007/978-3-319-49409-8_46.

[29] V. G. Buddhavarapu and A. A. J. J, “An experimental study on classification of thyroid histopathology images using transfer learning,” *Pattern Recognit. Lett.*, vol. 140, pp. 1–9, Dec. 2020, doi: 10.1016/j.patrec.2020.09.020.

[30] J. Chang, J. Yu, T. Han, H. Chang, and E. Park, “A method for classifying medical images using transfer learning: A pilot study on histopathology of breast cancer,” in *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, Oct. 2017, pp. 1–4. doi: 10.1109/HealthCom.2017.8210843.

[31] W. Borges Sampaio, E. Moraes Diniz, A. Corrêa Silva, A. Cardoso de Paiva, and M. Gattass, “Detection of masses in mammogram images using CNN, geostatistic functions and SVM,” *Comput. Biol. Med.*, vol. 41, no. 8, pp. 653–664, Aug. 2011, doi: 10.1016/j.compbiomed.2011.05.017.

Figures

Figure 3.1

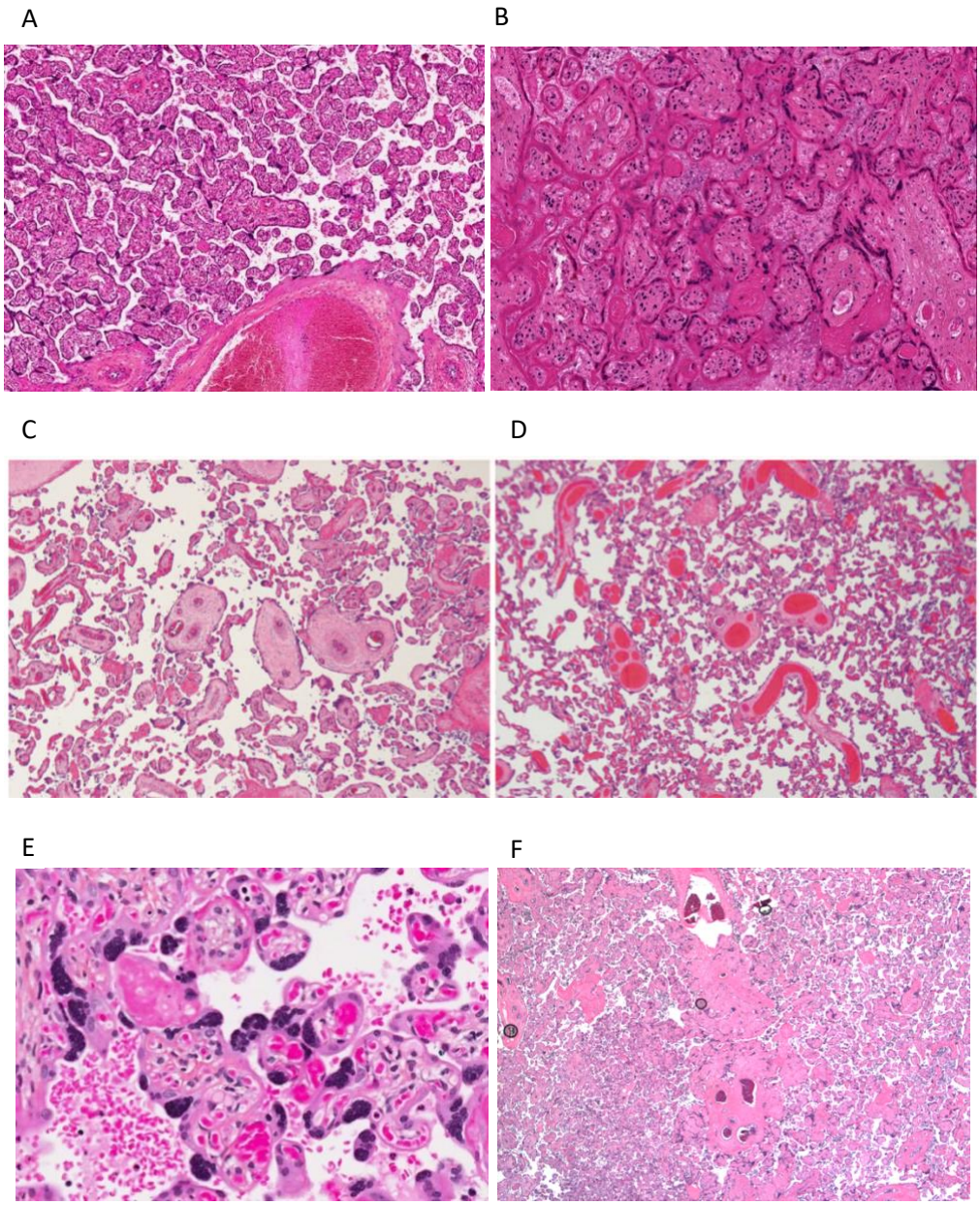


Figure 3.2

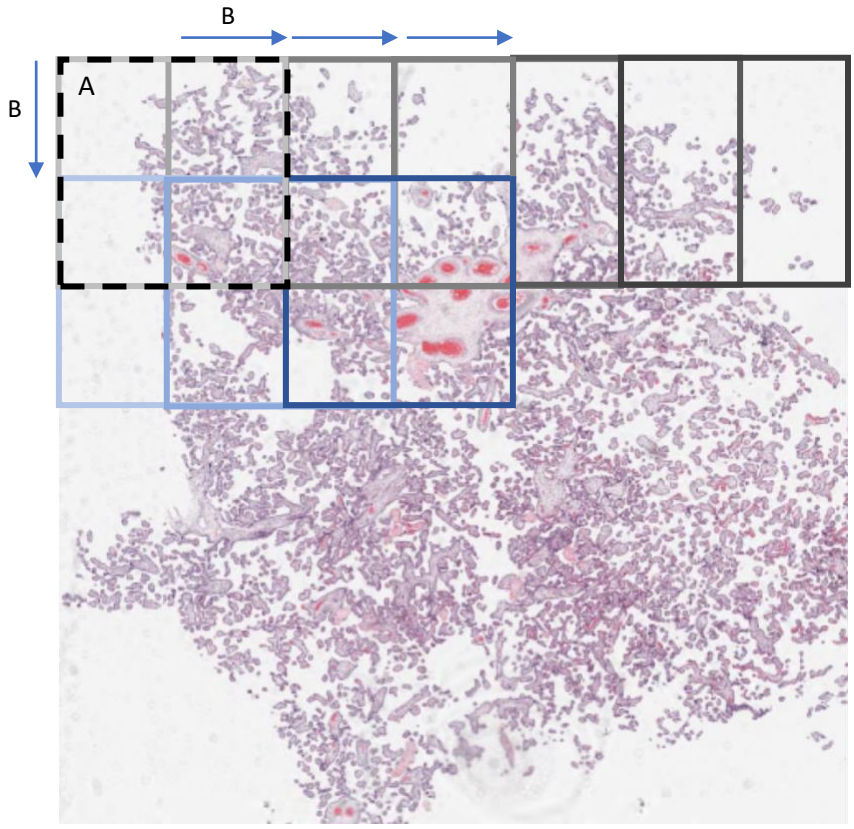


Figure 3.3

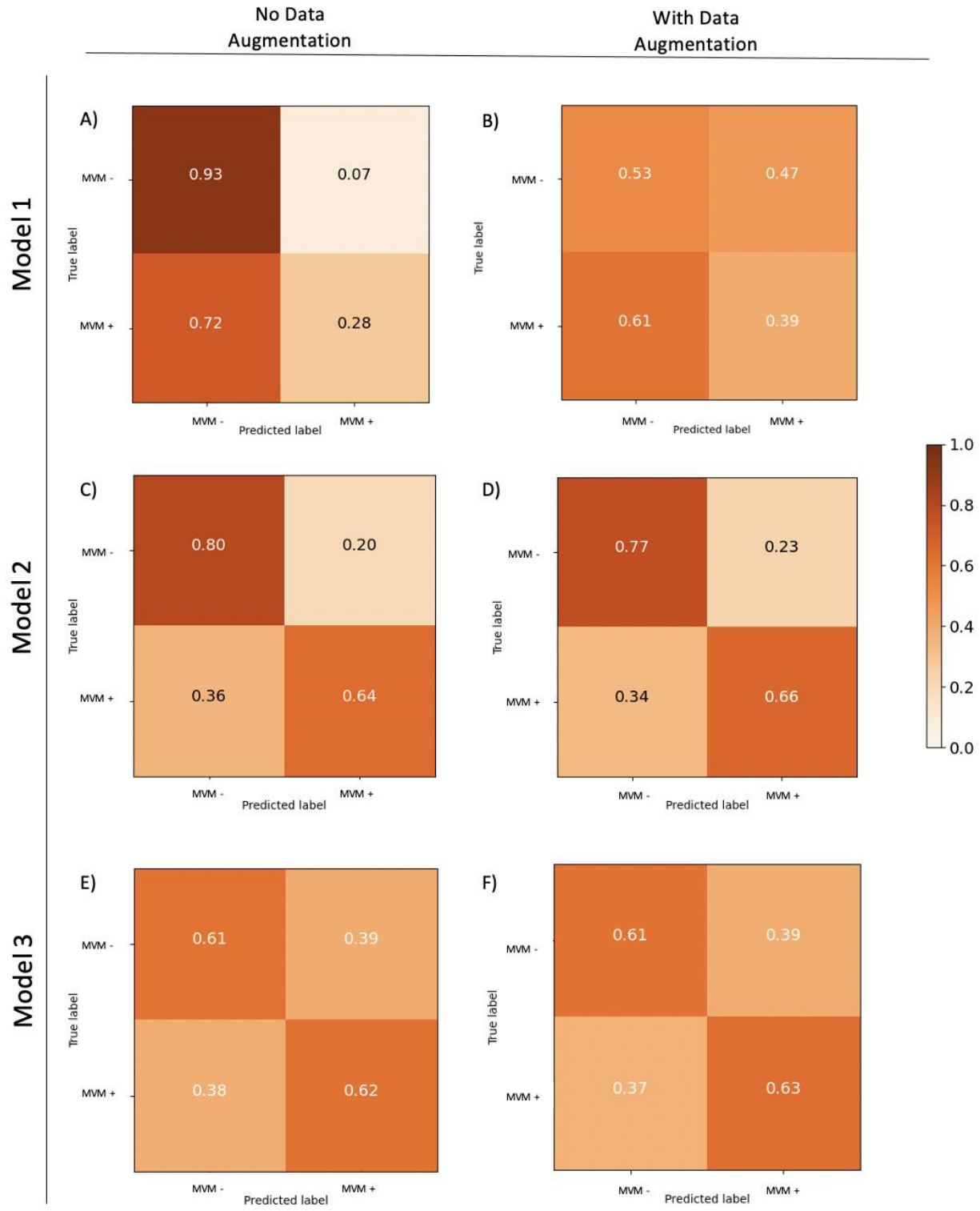

No Data Augmentation
With Data Augmentation
Model 1
Model 2
Model 3
A)
B)
C)
D)
E)
F)
MVM -
MVM +
True label
Predicted label
0.93
0.07
0.72
0.28
0.53
0.47
0.61
0.39
0.80
0.20
0.36
0.64
0.77
0.23
0.34
0.66
0.61
0.39
0.38
0.62
0.61
0.39
0.37
0.63
1.0
0.8
0.6
0.4
0.2
0.0

Figure 3.4

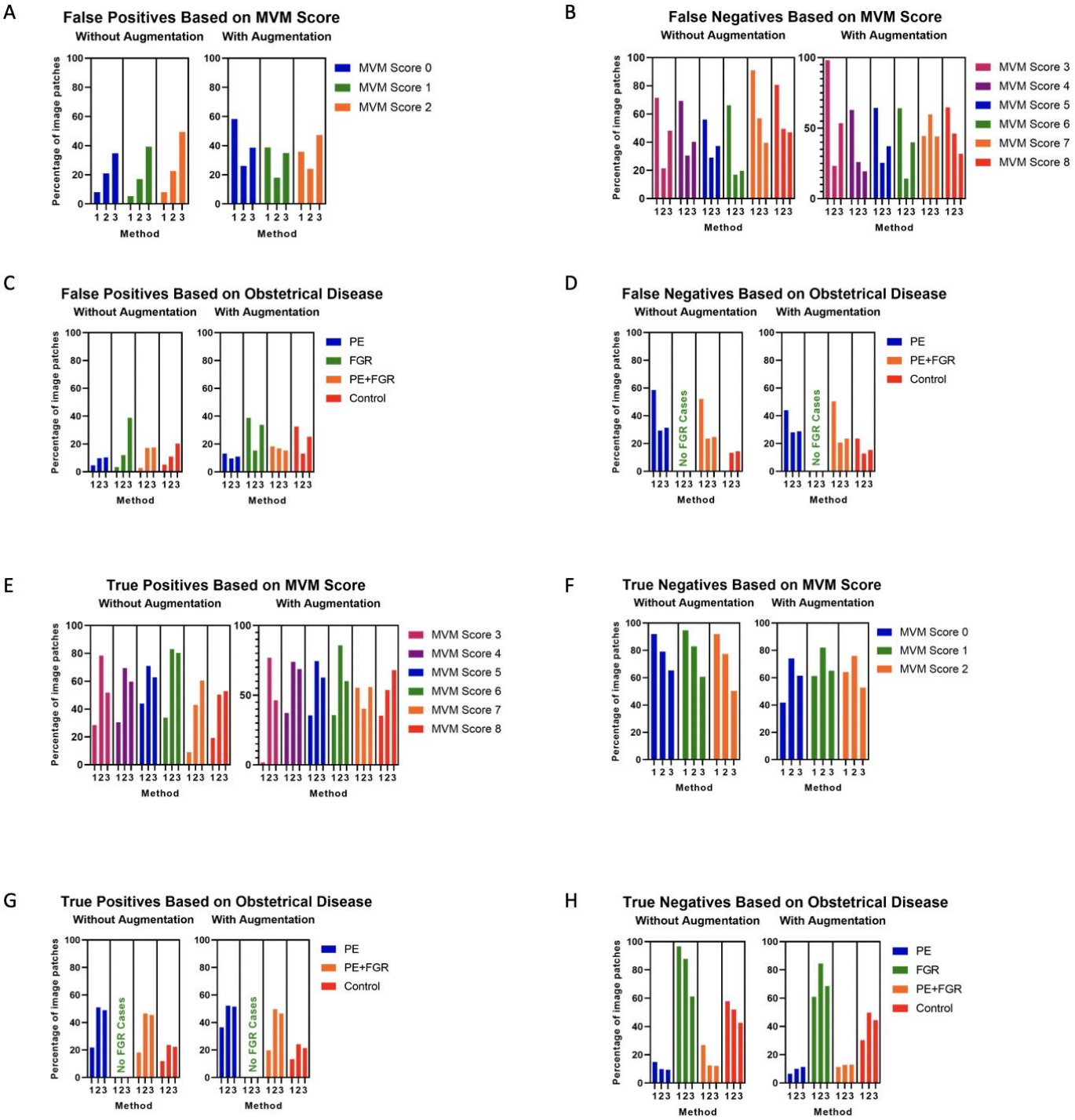

A
False Positives Based on MVM Score
Without Augmentation
With Augmentation
Percentage of image patches
Method
MVM Score 0
MVM Score 1
MVM Score 2
B
False Negatives Based on MVM Score
Without Augmentation
With Augmentation
Percentage of image patches
Method
MVM Score 3
MVM Score 4
MVM Score 5
MVM Score 6
MVM Score 7
MVM Score 8
C
False Positives Based on Obstetrical Disease
Without Augmentation
With Augmentation
Percentage of image patches
Method
PE
FGR
PE+FGR
Control
D
False Negatives Based on Obstetrical Disease
Without Augmentation
With Augmentation
Percentage of image patches
No FGR Cases
Method
PE
PE+FGR
Control
E
True Positives Based on MVM Score
Without Augmentation
With Augmentation
Percentage of image patches
Method
MVM Score 3
MVM Score 4
MVM Score 5
MVM Score 6
MVM Score 7
MVM Score 8
F
True Negatives Based on MVM Score
Without Augmentation
With Augmentation
Percentage of image patches
Method
MVM Score 0
MVM Score 1
MVM Score 2
G
True Positives Based on Obstetrical Disease
Without Augmentation
With Augmentation
Percentage of image patches
No FGR Cases
Method
PE
PE+FGR
Control
H
True Negatives Based on Obstetrical Disease
Without Augmentation
With Augmentation
Percentage of image patches
Method
PE
FGR
PE+FGR
Control

Figure Legends

Figure 3.1 Placental lesions that define MVM. **A**: Normal placenta at term, **B**: Placental infarct, **C**: Distal Villous Hypoplasia, **D**: Accelerated Villous Maturation, **E**: Syncytial knots, **F**: Villous Agglutination. (All panels provided by Dr. David Grynspan).

Figure 3.2 Patch Extraction - Cropping in a sliding window in pre-processing for a manual crop of a tissue sample on a WSI. **A**: Window indicated by dashed lines of 1000 x 1000 px, **B:** Arrows indicating how the window slides by 500x and extracts patches of size 1000 x 1000 px. Squares shown by grey and blue colour gradients are patches extracted from the sliding window method.

Figure 3.3 Normalized confusion matrices comparing the three different feature extraction models, with or without data augmentation. Quadrants in the confusion matrices represent true negatives (top left), false negatives (bottom left), true positives (bottom right), and false positives (top right). The number of images classified for these four groups is represented by a colour scale; where white represents low numbers and maroon represents high numbers.

Figure 3.4 Sub-analysis of image characteristics for misclassified images. Images that were misclassified as MVM+ (false positives) or MVM – (false negatives) were examined according to manual cumulative MVM score (0-8; panel A and B, E and F) and according to the obstetrical outcome (PE, FGR, or PE+FGR; Panel C and D, G and H).

Tables

Table 3.1 Demographic characteristics of participants in the study.

		MVM- (n=85)	MVM+ (n=81)
Maternal Age[a]		33(5)	33(6)
Gestational Age at Delivery (weeks)[a]		36(4)	32(3)
Fetal Sex (F)[b]		41/85	40/81
Birthweight (g)[a]		2417(907)	1364(579)
Birthweight Centile (<10th Percentile)[b]		24/85	54/81
Placental Weight Centile (<10th Percentile)[b]		25/85	67/81
Maternal Systolic Blood Pressure[a]		148(20)	165(19)
Maternal Diastolic Blood Pressure[a]		92(15)	104(12)
Mode of Delivery(C-section)[b]		47/85	68/81
Obstetrical Outcome[b]	**Control**	51/85	16/81
	Normotensive FGR	13/85	7/81
	PE	15/85	31/81
	PE+FGR	6/85	27/81

[a]Results presented as “mean (standard deviation)” for continuous variables

[b]Results presented as “count/clinical subset” for categorical variables.

Table 3.2 Python libraries used for SVM models.

Phase		Task	Python3 Library	Module
Pre-processing		Reading .svs files	OpenSlide	.read_region
		Cropping and Convert to RGB	from PIL import Image	.crop .convert('RGB')
		Data Augmentation	Keras	ImageDataGenerator()
		Minimum Keypoint Threshold	OpenCV	cvtColor normalize
				cv2.xfeatures2d.SIFT_create() sift.detectAndCompute()
Feature Extraction	**Method #1**	Patch extraction and feature calculation		MATLAB
		Feature Normalization	SciKit Learn	preprocessing.MinMaxScaler()
	Method #2	Extract SIFT descriptors	SciKit Learn	cv2.xfeatures2d.SIFT_create() sift.detectAndCompute()
		Stack descriptors and calculate medians	SciKit Learn	dsc[0:500]
			Numpy	np.median()
		Feature Normalization	SciKit Learn	preprocessing.MinMaxScaler()
	Method #3	Extract SIFT descriptors	SciKit Learn	cv2.xfeatures2d.SIFT_create() sift.detectAndCompute()
		K-means clustering for the visual bag of words dictionary codebook	SciPy	kmeans(()
		Feature Normalization	SciKit Learn	preprocessing.MinMaxScaler()
Classification		Data splitting (train/test)	SciKit Learn	GroupShuffleSplit
		Support Vector Machine	SciKit Learn	svm.SVC
Evaluation		ROC Curves	SciKit Learn	metrics.roc
		Confusion Matrices	SciKit Learn	metrics.confusion_matrix

Table 3.3 Breakdown of datasets used in the study. Patches (X=17,364) were extracted from 166 WSIs.

Train Set – Original Patches (Total=13,349)	**MVM Threshold**	**MVM-**			**MVM+**					
	MVM Score	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	Number of Patches	3297	1660	1820	1689	1683	1486	1420	222	72
	Obstetrical Outcome	**Control**			**PE**		**FGR**		**PE + FGR**	
	Number of Patches	4397			2317		4561		2074	
Train Set – Data Augmented Patches (Total=39,076)	**MVM Threshold**	**MVM-**			**MVM+**					
	MVM Score	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	Number of Patches	9576	4859	5310	4941	4981	4370	4178	645	216
	Obstetrical Outcome	**Control**			**PE**		**FGR**		**PE + FGR**	
	Number of Patches	13,119			6835		13 027		6095	
Test Set (Total=4015)	**MVM Threshold**	**MVM-**			**MVM+**					
	MVM Score	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	Number of Patches	1054	845	460	56	611	357	148	365	119
	Obstetrical Outcome	**Control**			**PE**		**FGR**		**PE + FGR**	
	Number of Patches	1152			1106		1272		485	

Table 3.4 Summary of classification performance metrics for each developed model, both with and without data augmentation.

Model #	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1-Score	AUC of ROC
Model 1) Biologically Relevant Features: Original Image Set	0.66	0.68	0.66	0.66	0.62	0.75
Model 1) Biologically Relevant Features: Data Augmented Image Set	0.47	0.48	0.47	0.47	0.48	0.50
Model 2) SIFT Descriptors: Original Image Set	0.73	0.73	0.73	0.73	0.73	0.77
Model 2) SIFT Descriptors: Data Augmented Image Set	0.73	0.73	0.73	0.73	0.73	0.76
Model 3) Visual Bag of Words: Original Image Set	0.61	0. 62	0. 61	0.61	0. 61	0.65
Model 3) Visual Bag of Words: Data Augmented Image Set	0. 63	0. 63	0. 62	0.62	0. 61	0.66

*Precision, recall, and F1-scores are weighted averages that consider the proportion for each label in the dataset.

Supplementary Figures

Figure S1.

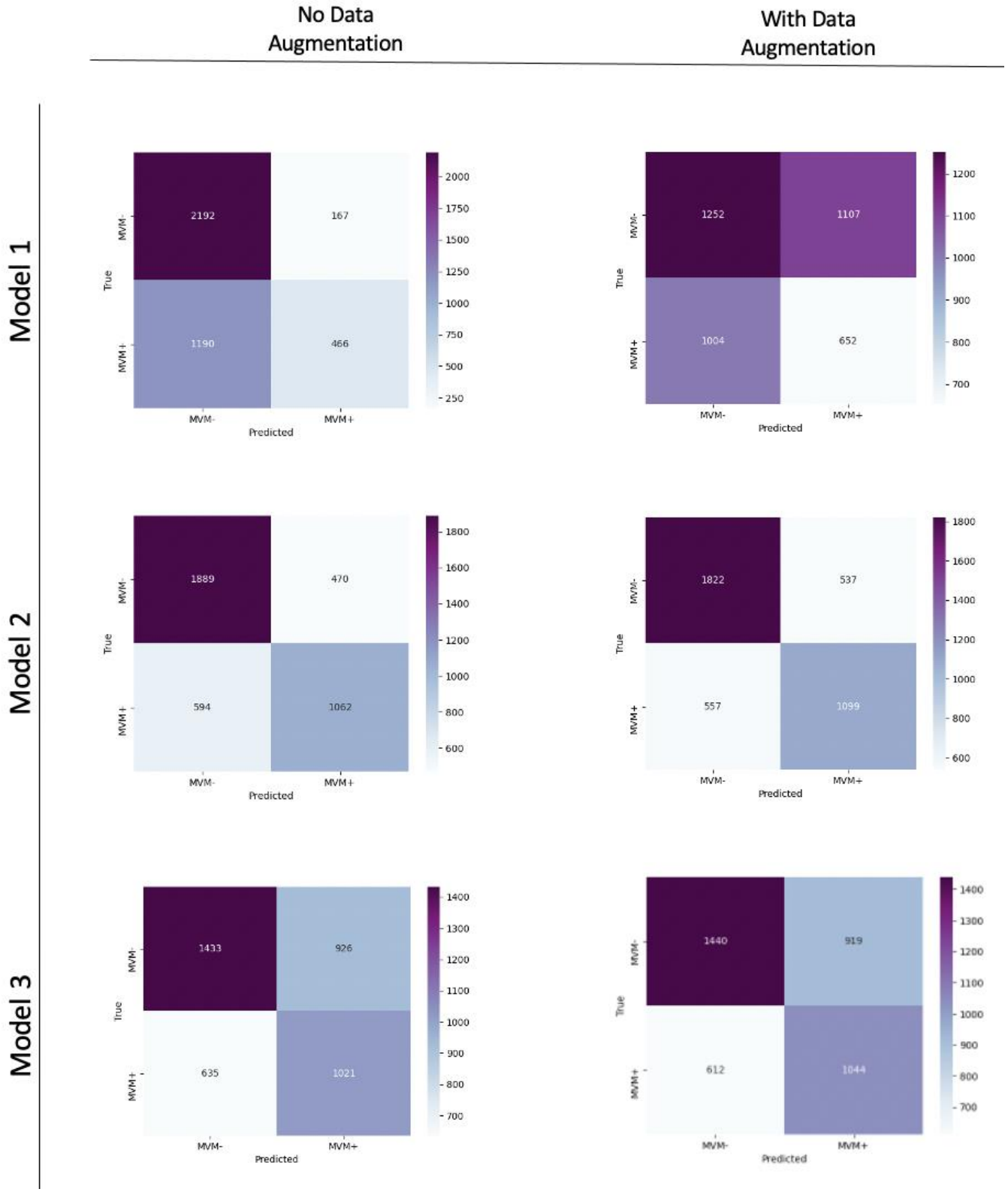
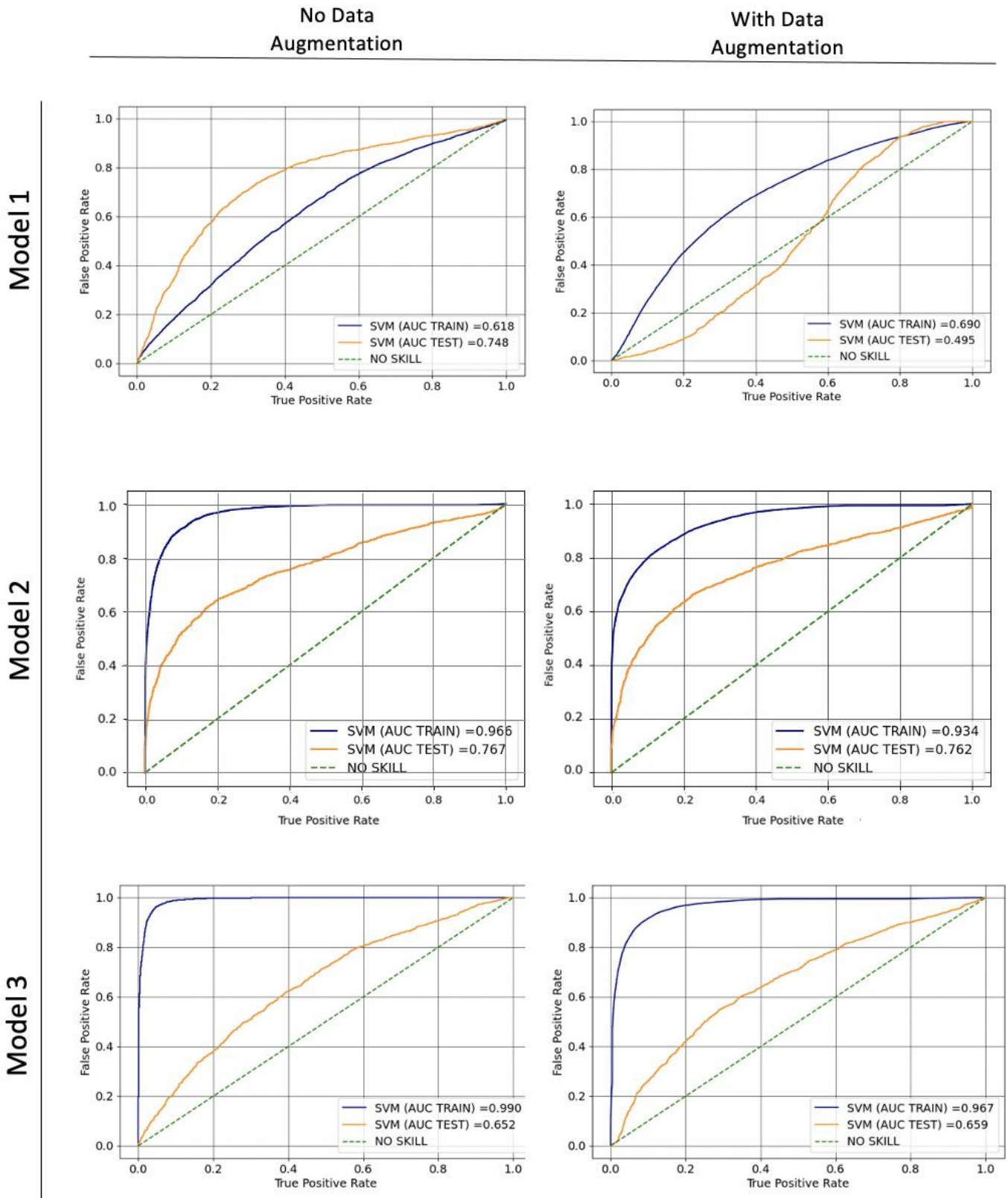


Figure S2.


No Data Augmentation
With Data Augmentation
Model 1
Model 2
Model 3
False Positive Rate
True Positive Rate
SVM (AUC TRAIN) =0.618
SVM (AUC TEST) =0.748
NO SKILL
SVM (AUC TRAIN) =0.690
SVM (AUC TEST) =0.495
NO SKILL
SVM (AUC TRAIN) =0.966
SVM (AUC TEST) =0.767
NO SKILL
SVM (AUC TRAIN) =0.934
SVM (AUC TEST) =0.762
NO SKILL
SVM (AUC TRAIN) =0.990
SVM (AUC TEST) =0.652
NO SKILL
SVM (AUC TRAIN) =0.967
SVM (AUC TEST) =0.659
NO SKILL

Supplementary Figure Legend

S1. Confusion matrices comparing the three different feature extraction models, with or without data augmentation. Quadrants in the confusion matrices represent true negatives (top left), false negatives (bottom left), true positives (bottom right), and false positives (top right). The number of images classified for these four groups is represented by a colour scale; where white represents low numbers and purple represents high numbers.

S2. Receiver Operating Characteristic (ROC) curves and AUC scores for three different feature extraction models, with and without data augmentation.

Chapter 4

Integrated Discussion

4.1 Discussion

Tremendous advances are happening across the healthcare landscape with CAD-based technology, yet the field of clinical pathology – and even more so placental pathology – remains in the early discovery phase of this health technology domain. In fact, only a handful of small-scale, research-focused studies have explored the application of CAD-tools to the topic of placental health and disease. The current thesis directly tackles this research gap, aiming to examine the performance of ML models to accurately identify placenta-mediated disease. Specifically, using a highly phenotyped dataset of digital placenta histopathology images, three ML models that employed different feature extraction methods were tested for their ability to accurately predict the presence of a visually distinct, common, and recurring placental disease – MVM.

In this thesis, the ML models tested used an SVM architecture. SVMs are classical supervised machine learning models, which treat extracted features as class-labelled data points in a multi-dimensional space. The SVM trains to identify an optimal hyperplane of separation that can split extracted feature data points according to their class label, and in the testing phase uses this hyperplane to 'predict' what class the unseen and unlabelled datapoints belongs to. This model has shown considerable promise when applied to different types of histopathology specimens, and continues to be a standard model tested for applications of image recognition and histological image analysis [48][49][50].

While SVM models have been a primary model examined for utility in the domain of clinical pathology, the high degree of heterogeneity in research question/aim, type and sample

size of specimens used, study design and methodology employed have made direct comparison of bodies of work a challenge and have limited the clinical uptake of this technology. For example, a recent systematic review of the literature carried out by our group [Mukherjee et al. (co-author status), 2021 under review; Appendix C] identified that the majority of work completed to date was carried out using oncology histopathology specimens (N= 18 studies total), however this body of work spanned 9 different types of cancer, with sample sizes ranging from 7 - 1,000 tissue specimens (15 - 21,860 digital images). While a common overarching work pipeline was identified across studies, consisting of 5 distinct steps [1) image acquisition, 2) pre-processing, 3) feature extraction, 4) classifier training and 5) model testing], there were considerable discrepancies across studies regarding the specific methodologies employed. Specifically, different types of normalization techniques in image pre-processing [51] and various forms of patch extraction for patch-based classification approaches are used in histopathological ML models [52]. Data augmentation techniques vary across studies; for example some studies used selective techniques while others apply several techniques [53][54].

A primary motivation in carrying out this systematic review was to identify a widely used and easily comparable ML workflow that could be applied to the field of placenta pathology, as very little work had been carried out to date in this domain. Armed with a highly thought-out workflow, our group first attempted to develop both an SVM and CNN model with the aim of classifying placental histopathology specimens according to obstetrical outcome (PE with and without FGR) using the same digital image dataset used in the current study. This first attempt generated model classification accuracies ranging from 41-73%, depending on which diagnostic criteria were used to define PE and which ML model was used (unpublished, Mukherjee et al.).

The interdisciplinary group working on this project quickly realized that the highly heterogenous nature of PE, particularly in the types of placental disease that can underlie this diagnosis, was limiting the success of ML model development and performance. The current thesis was born out of this discovery, re-focusing these preliminary attempts at the identification of specific forms of placental disease with distinct visual appearance – as is seen in MVM. In addition to a classification re-focus, these initial experiments were tremendously helpful in optimizing aspects of the ML model development and testing workflow, with improvements to the methods of patch extraction and feature extraction resulting in fewer sources of bias, more specificity, and improved codes. In the data science pipeline, it is common to develop ML models, analyze the results, and readjust the model – in some cases several times over. The current project is a nice example of this evolution of health technology, having been informed by past attempts and being used to inform future attempts.

One major contribution of the current thesis in moving this field forward was the direct comparison of three different methods of feature extraction from the highly complicated microscopic structure of the human placenta. Biologically relevant features, 500 SIFT keypoints, and visual bag of words (using SIFT keypoints) were the 3 different methods of feature extraction compared in this thesis. While all three methods showed promise for the classification of placental disease, the direct use of 500 SIFT keypoint features provided the highest degree of classification accuracy (73%). As discussed in Chapter 3, this is likely the result of limitations in the types of feature extractions and ML models used in model #1 and #3 in the context of placental histopathological image classification. For example, model #1 involves a high degree of user bias in the selection of features that are being extracted. For model #3, the

method of feature extraction in the visual bag of words model involves the loss of information in its strict encoding process, and uses a high computational effort for large codebook generation [55]. Unfortunately, the SIFT feature extraction method has recently become patented, making the implementation of SIFT extraction using Python3 modules a challenge moving forward. However, it should be noted that there are newer published methods for feature extraction for image recognition purposes, such as BRIEF, ORB, and SURF features [56][57][58][59], which should be assessed for their compatibility for placenta histopathology purposes moving forward. This clearly demonstrates how rapidly the ML modules and programs are currently evolving. However, as ML application for placental pathology purposes is still in its infancy, our team and others are well posited to build upon our foundational work with these classical ML models, evolving them as new advances are made in this field.

The most surprising finding of this project was the lack of model enhancement when data augmentation methods were applied to our models. Data augmentation techniques are frequently described as an effective method to increase training dataset size, particularly in situations with small sample sizes, allowing for the development of a more robust ML model with higher classification performance [60][61][62][63]. As employed in the current project, data augmentation provided no added benefit to the classifier. In fact, in the case of model #1 (*a priori* biological feature extraction), the use of data augmentation *decreased* the classification performance of the model. While we are not entirely certain why this might be, we speculate this may be the result of an increased number of similar-kind training data points from the 'reflect' type of fill mode of the data augmentation. These similar data points might have distorted the support vectors and thereby the hyperplane such that the separation

between the two classes after data augmentation was worse than training the model without data augmentation [64]. Due to both model #2 and #3 using SIFT features, one explanation of why the data augmentation did not improve the accuracy could be because the SIFT algorithm is invariant to some forms of the data augmentation (ex. flips, rotations). The next steps will be to manually review the augmented images as an interdisciplinary approach to determine if any specific form of image manipulation used in the augmentation process was distorting the visually distinct MVM damage such that model training became compromised, and if so possibly modify or limit the types of image manipulations employed. Although adding data augmentation to the training dataset provided little to no improvement to classification performance, it did highlight the importance of having a robust dataset of unique WSI images. Moving forward, we will further explore whether data augmentation may be useful to address issues of class imbalance across training and testing sets – particularly for rarer forms of severe placental disease (i.e., MVM score of 7-8). For example, in the analyzed dataset, there were only 72 patches (out of 13,349) that had an MVM score of 8. Overall, this thesis confirmed that feeding the model too many augmentations of the same image in a small unique image dataset will not improve accuracies.

Given that MVM is the most common form of placental disease underlying PE and FGR, the ability to automate the identification of this type of placental disease would pave the way for the future incorporation of machine learning into the placental pathology process. The current study serves as a proof-of-concept for the development and future incorporation of CAD-based tools into the placenta pathology exam. It further serves as an important

contribution to data science workflow optimization for CAD development in histopathology image analysis.

4.2 Limitations and Future Work

The current study not only shows promise in using supervised ML in placental pathology but also uncovers the next steps that should be taken to improve the models. One of the most limiting parameters is the small number of images readily available for training and testing models. Future development of ML models in placental pathology will require the curation and maintenance of publicly available digital image datasets of placental histopathology specimens, including both healthy and diseased tissue with a range of disease severity. Curating databases of these images, having them validated by a placental pathologist, and ensuring that they are utilizable in ML models is a massive undertaking. These databases could not only hold whole-slide images but could further include pre-extracted and pre-processed image patches (with and without augmentation) that are readily available for ML model development and testing. With contributions from placental biologists, obstetricians, perinatal pathologists, and data scientists alike, this database could become a tremendously useful resource (much like https://www.cancerimagingarchive.net/), allowing for the development of CAD-based tools for a variety of placenta-mediated diseases. The development of such a crowdsourced database will no doubt require considerable resources – both financial and man-power – but is a future goal for our group. This large-scale future endeavour, and the CAD-tools that can be developed as a result, will certainly help to improve the clinical care for mothers and babies, particularly in low-resource settings with limited access to robust placenta pathology examinations.

Access to a large image repository would further allow our group, and others, to explore the potential of deep learning models for classification purposes in the domain of placenta pathology. Convolutional neural networks are an ideal ML model for image classification, and when coupled to transfer learning has further tremendous potential for image recognition purposes [65][66][67]. We anticipate that results of on-going efforts to develop a CNN model for placenta pathology classification, and mores specifically MVM classification, will provide us with results of higher accuracy than the SVM, as CNNs are known to perform very well for digital image data (see Appendix B for preliminary results). However, it is important to consider that each model has its pros and cons; where CNN is prevalent in object recognition, SVM is highly prevalent in classification analysis [68]. Therefore, testing and comparing different paradigms in computer vision such as classical ML (ex. SVM, random forest, logistic regression, decision trees), deep learning (CNNs), and transfer learning (VGG-16, ResNet50) is essential to gaining perspective on the most optimal ways these images can be classified.

4.3 Interdisciplinarity

This thesis is highly interdisciplinary and bridges a crucial research gap between the worlds of artificial intelligence (computer science), medical imaging, and clinical placental pathology. Computer programing, machine learning, advanced statistics, clinical placental pathology, and translational researchers came together to identify a clinical problem that requires a solution, propose a potential solution to the problem, design and implement the current study and importantly interpret the results collectively and identify next steps for the project. Each member of our team comes to the table with a strong background in their

respective disciplines that were relevant and required to mature this project – including a perinatal pathologist, computer engineer, and placental biologist. Approaching our objectives in an interdisciplinary manner allowed us to answer innovative research questions and helped move the field of perinatal pathology forward in a profound way. All members of the team, including myself as the MSc student working on this project, found the experience of working together to be highly rewarding, and recognize that the work accomplished would not have been possible without the collaborative and integrated approach taken by all team members.

4.4 Significance & Conclusion

The significance of this research is clear – improvement of healthcare and health outcomes for mothers and their babies. The placenta pathology exam offers important insight into the intra-uterine environment over the course of pregnancy, identifying underlying causes of obstetrical disease and adverse pregnancy outcomes. These CAD tools are envisioned as a supplementary tool that could aid a perinatal pathologist in coming to final diagnosis - helping to reduce the high-resource burden faced by pathologists. The results of the placenta pathology exam are sent to the clinical care team of the mother and infant (in cases where infants are admitted to NICU). Their health teams use these results to plan post-partum care and importantly these results can be used in the context of post-partum counseling for families who have had an adverse pregnancy outcome, allowing them to understand what went wrong in the current pregnancy and what their risk of recurrence might be in future pregnancies. More recent research suggests that the placenta pathology exam can further be used to identify mothers and/or infants at higher risk of chronic disease across the lifespan, resulting from

exposure to, and damage from, disease pathophysiology across pregnancy. Despite the importance of this clinical practice, placenta pathology exams have historically been plagued by a lack of standardization in diagnostic terminology and criteria, and poor inter-pathologist agreement on diagnostic outcomes. Further, a placental pathology exam that can have the most impact clinically needs to be carried out by a highly trained perinatal pathologist and is highly resource intensive. As a result, low resource and community settings are often unable to benefit from this clinical practice. Collectively, these limitations make the placenta pathology exam an ideal candidate for CAD-based technology advances, where automation of some aspects of this exam may reduce workload burden on pathologists and improve healthcare for mothers and babies.

This proof-of-concept study is an important first step in getting CAD technology integrated into the placenta pathology exam, demonstrating the potential of ML models to accurately identify distinct forms of placental disease. The foundational contributions of this thesis are: 1) An ML model that demonstrates good classification performance for the identification of MVM in digital histopathology images, and 2) Improved ML model development workflow methodology, with identified solutions for further model improvement. These are contributions that will serve to move this field of study forward, and ultimately lead to improved care for mothers and their babies.

References

[1] S. Salsabili, A. Mukherjee, E. Ukwatta, A. D. C. Chan, S. Bainbridge, and D. Grynspan, "Automated segmentation of villi in histopathology images of placenta," *Comput. Biol. Med.*, vol. 113, p. 103420, Oct. 2019, doi: 10.1016/j.compbiomed.2019.103420.

[2] V. Kapila and K. Chaudhry, "Physiology, Placenta," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2021. Accessed: Jul. 20, 2021. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK538332/

[3] S. Oratz, "The Hormones of the Placenta," *Sci. J. Lander Coll. Arts Sci.*, vol. 8, no. 1, Jan. 2014, [Online]. Available: https://touroscholar.touro.edu/sjlcas/vol8/iss1/7

[4] N. Gleicher, J. Metzger, G. Croft, V. A. Kushnir, D. F. Albertini, and D. H. Barad, "A single trophectoderm biopsy at blastocyst stage is mathematically unable to determine embryo ploidy accurately enough for clinical use," *Reprod. Biol. Endocrinol.*, vol. 15, no. 1, p. 33, Apr. 2017, doi: 10.1186/s12958-017-0251-8.

[5] J. B. Gordon et al., *Anatomy and Physiology*, vol. Fertilization. Rice University, OpenStax, 2016. [Online]. Available: https://opentextbc.ca/anatomyandphysiology/chapter/28-2-embryonic-development/

[6] "Growth and function of the normal human placenta," *Thromb. Res.*, vol. 114, no. 5–6, pp. 397–407, Jan. 2004, doi: 10.1016/j.thromres.2004.06.038.

[7] G. J. Burton and E. Jauniaux, "What is the placenta?," *Am. J. Obstet. Gynecol.*, vol. 213, no. 4, p. S6.e1-S6.e4, Oct. 2015, doi: 10.1016/j.ajog.2015.07.050.

[8] G. Weiss, M. Sundl, A. Glasner, B. Huppertz, and G. Moser, "The trophoblast plug during early pregnancy: a deeper insight," *Histochem. Cell Biol.*, vol. 146, no. 6, pp. 749–756, 2016, doi: 10.1007/s00418-016-1474-z.

[9] E. Wright *et al.*, “Maternal Vascular Malperfusion and Adverse Perinatal Outcomes in Low-Risk Nulliparous Women:,” *Obstet. Gynecol.*, vol. 130, no. 5, pp. 1112–1120, Nov. 2017, doi: 10.1097/AOG.0000000000002264.

[10] American College of Obstetricians and Gynecologists and American College of Obstetricians and Gynecologists, Eds., *Hypertension in pregnancy*. Washington, DC: American College of Obstetricians and Gynecologists, 2013.

[11] N. Bhattacharjee *et al.*, “A randomised comparative study between low-dose intravenous magnesium sulphate and standard intramuscular regimen for treatment of eclampsia,” *J. Obstet. Gynaecol.*, vol. 31, no. 4, pp. 298–303, May 2011, doi: 10.3109/01443615.2010.549972.

[12] J. Park, D. H. Cha, S. J. Lee, Y. N. Kim, Y. H. Kim, and K. P. Kim, “Discovery of the serum biomarker proteins in severe preeclampsia by proteomic analysis,” *Exp. Mol. Med.*, vol. 43, no. 7, pp. 427–435, Jul. 2011, doi: 10.3858/emm.2011.43.7.047.

[13] G. Yosten, G. Kolar, L. Stein, C. Haddock, G. A. Pereira, and W. Samson, “Signaling of Phoenixin via GPR173 and the Paradox of Increased Vasopressin Secretion in Preeclampsia,” *FASEB J.*, vol. 35, no. S1, 2021, doi: 10.1096/fasebj.2021.35.S1.04204.

[14] J. M. Roberts and C. A. Hubel, “The two stage model of preeclampsia: variations on the theme,” *Placenta*, vol. 30 Suppl A, pp. S32-37, Mar. 2009, doi: 10.1016/j.placenta.2008.11.009.

[15] S. I. Gilani, T. L. Weissgerber, V. D. Garovic, and M. Jayachandran, “Preeclampsia and Extracellular Vesicles,” *Curr. Hypertens. Rep.*, vol. 18, no. 9, p. 68, Sep. 2016, doi: 10.1007/s11906-016-0678-x.

[16] F. T. Spradley, A. C. Palei, and J. P. Granger, “Increased risk for the development of preeclampsia in obese pregnancies: weighing in on the mechanisms,” *Am. J. Physiol.-Regul. Integr. Comp. Physiol.*, vol. 309, no. 11, pp. R1326–R1343, Dec. 2015, doi: 10.1152/ajpregu.00178.2015.

[17] B. Huppertz, “Placental Origins of Preeclampsia,” *Hypertension*, vol. 51, no. 4, pp. 970–975, Apr. 2008, doi: 10.1161/HYPERTENSIONAHA.107.107607.

[18] L. M. M. Nardozza *et al.*, “Fetal growth restriction: current knowledge,” *Arch. Gynecol. Obstet.*, vol. 295, no. 5, pp. 1061–1077, May 2017, doi: 10.1007/s00404-017-4341-9.

[19] I. Gibbs, K. Leavey, S. J. Benton, D. Grynspan, S. A. Bainbridge, and B. J. Cox, “Placental transcriptional and histologic subtypes of normotensive fetal growth restriction are comparable to preeclampsia,” *Am. J. Obstet. Gynecol.*, Oct. 2018, doi: 10.1016/j.ajog.2018.10.003.

[20] L. Youssef *et al.*, “Paired maternal and fetal metabolomics reveal a differential fingerprint in preeclampsia versus fetal growth restriction,” *Sci. Rep.*, vol. 11, no. 1, p. 14422, Jul. 2021, doi: 10.1038/s41598-021-93936-9.

[21] S. R. Easter *et al.*, “Fetal growth restriction: Case definition & guidelines for data collection, analysis, and presentation of immunization safety data,” *Vaccine*, vol. 35, no. 48Part A, pp. 6546–6554, Dec. 2017, doi: 10.1016/j.vaccine.2017.01.042.

[22] “O114. Maternal cardiovascular changes in pregnancies complicated by small for gestational age neonate with or without maternal hypertension,” *Pregnancy Hypertens. Int. J. Womens Cardiovasc. Health*, vol. 5, no. 3, p. 234, Jul. 2015, doi: 10.1016/j.preghy.2015.07.065.

[23] R. B. Ness and B. M. Sibai, “Shared and disparate components of the pathophysiologies of fetal growth restriction and preeclampsia,” *Am. J. Obstet. Gynecol.*, vol. 195, no. 1, pp. 40–49, Jul. 2006, doi: 10.1016/j.ajog.2005.07.049.

[24] J. Stubert, S. Ullmann, M. Dieterich, D. Diedrich, and T. Reimer, “Clinical differences between early- and late-onset severe preeclampsia and analysis of predictors for perinatal outcome,” *J. Perinat. Med.*, vol. 42, no. 5, pp. 617–627, Sep. 2014, doi: 10.1515/jpm-2013-0285.

[25] “Placental histopathological lesions in correlation with neonatal outcome in preeclampsia with and without severe features,” *Pregnancy Hypertens.*, vol. 12, pp. 6–10, Apr. 2018, doi: 10.1016/j.preghy.2018.02.001.

[26] T. Wang *et al.*, “Epigenome-wide association data implicate fetal/maternal adaptations contributing to clinical outcomes in preeclampsia,” *Epigenomics*, May 2019, doi: 10.2217/epi-2019-0065.

[27] S. J. Benton, K. Leavey, D. Grynspan, B. J. Cox, and S. A. Bainbridge, “The clinical heterogeneity of preeclampsia is related to both placental gene expression and placental histopathology,” *Am. J. Obstet. Gynecol.*, vol. 219, no. 6, p. 604.e1-604.e25, Dec. 2018, doi: 10.1016/j.ajog.2018.09.036.

[28] L. M. Ernst, “Maternal vascular malperfusion of the placental bed,” *APMIS*, vol. 126, no. 7, pp. 551–560, Jul. 2018, doi: 10.1111/apm.12833.

[29] C. M. Scifres, W. T. Parks, M. Feghali, S. N. Caritis, and J. M. Catov, “Placental maternal vascular malperfusion and adverse pregnancy outcomes in gestational diabetes mellitus,” *Placenta*, vol. 49, pp. 10–15, Jan. 2017, doi: 10.1016/j.placenta.2016.11.004.

[30] J. M. Catov *et al.*, “Preterm birth with placental evidence of malperfusion is associated with cardiovascular risk factors after pregnancy: a prospective cohort study,” *BJOG Int. J. Obstet. Gynaecol.*, vol. 125, no. 8, pp. 1009–1017, 2018, doi: 10.1111/1471-0528.15040.

[31] R. W. Redline, “Classification of placental lesions,” *Am. J. Obstet. Gynecol.*, vol. 213, no. 4, pp. S21–S28, Oct. 2015, doi: 10.1016/j.ajog.2015.05.056.

[32] T. Y. Khong *et al.*, “Sampling and Definitions of Placental Lesions: Amsterdam Placental Workshop Group Consensus Statement,” *Arch. Pathol. Lab. Med.*, vol. 140, no. 7, pp. 698–713, Jul. 2016, doi: 10.5858/arpa.2015-0225-CC.

[33] K. Loukeris, R. Sela, and R. N. Baergen, “Syncytial Knots as a Reflection of Placental Maturity: Reference Values for 20 to 40 Weeks’ Gestational Age,” *Pediatr. Dev. Pathol.*, vol. 13, no. 4, pp. 305–309, Jul. 2010, doi: 10.2350/09-08-0692-OA.1.

[34] S. Waheed, R. A. Moffitt, Q. Chaudry, A. N. Young, and M. D. Wang, “Computer Aided Histopathological Classification of Cancer Subtypes,” in *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, Boston, MA, USA, Oct. 2007, pp. 503–508. doi: 10.1109/BIBE.2007.4375608.

[35] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, “A Dataset for Breast Cancer Histopathological Image Classification,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, Jul. 2016, doi: 10.1109/TBME.2015.2496264.

[36] A. BenTaieb, M. Nosrati, H. Li-Chang, D. Huntsman, and G. Hamarneh, “Clinically-inspired automatic classification of ovarian carcinoma subtypes,” *J. Pathol. Inform.*, vol. 7, no. 1, p. 28, 2016, doi: 10.4103/2153-3539.186899.

[37] J. Angel Arul Jothi and V. Mary Anita Rajam, “Effective segmentation and classification of thyroid histopathology images,” *Appl. Soft Comput.*, vol. 46, pp. 652–664, Sep. 2016, doi: 10.1016/j.asoc.2016.02.030.

[38] K.-H. Yu *et al.*, “Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features,” *Nat. Commun.*, vol. 7, no. 1, Dec. 2016, doi: 10.1038/ncomms12474.

[39] I. G. Maglogiannis, *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press, 2007.

[40] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, 2010.

[41] N. Cristianini, J. Shawe-Taylor, D. of C. S. R. H. J. Shawe-Taylor, and J. (Royal H. S.-T. London) University of, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[42] L. Ma, X. Liu, L. Song, C. Zhou, X. Zhao, and Y. Zhao, “A new classifier fusion method based on historical and on-line classification reliability for recognizing common CT imaging signs of lung diseases,” *Comput. Med. Imaging Graph.*, vol. 40, pp. 39–48, Mar. 2015, doi: 10.1016/j.compmedimag.2014.10.001.

[43] “GestAltNet: aggregation and attention to improve deep learning of gestational age from placental whole-slide images | Laboratory Investigation.” https://www-nature-com.proxy.bib.uottawa.ca/articles/s41374-021-00579-

5?s=09&fbclid=IwAR1W5YdQMDMuBi2BczI5wIroMJeFE9QV3lC19RGNJY5JpankHJRspUv3xmE (accessed Aug. 11, 2021).

[44] H. Sun *et al.*, "Identification of suspicious invasive placentation based on clinical MRI data using textural features and automated machine learning," *Eur. Radiol.*, vol. 29, no. 11, pp. 6152–6162, Nov. 2019, doi: 10.1007/s00330-019-06372-9.

[45] C.-C. J. Sun, V. O. Revell, A. J. Belli, and R. M. Viscardi, "Discrepancy in Pathologic Diagnosis of Placental Lesions," *Arch Pathol Lab Med*, vol. 126, p. 4, 2002.

[46] Z. Obermeyer and E. J. Emanuel, "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine," *N. Engl. J. Med.*, vol. 375, no. 13, pp. 1216–1219, Sep. 2016, doi: 10.1056/NEJMp1606181.

[47] U. Djuric, G. Zadeh, K. Aldape, and P. Diamandis, "Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care," *Npj Precis. Oncol.*, vol. 1, no. 1, Dec. 2017, doi: 10.1038/s41698-017-0022-1.

[48] W. Jiang *et al.*, "A Nomogram Based on a Collagen Feature Support Vector Machine for Predicting the Treatment Response to Neoadjuvant Chemoradiotherapy in Rectal Cancer Patients," *Ann. Surg. Oncol.*, Jun. 2021, doi: 10.1245/s10434-021-10218-4.

[49] D. T. Blumenthal, M. Artzi, G. Liberman, F. Bokstein, O. Aizenstein, and D. B. Bashat, "Classification of High-Grade Glioma into Tumor and Nontumor Components Using Support Vector Machine," *Am. J. Neuroradiol.*, vol. 38, no. 5, pp. 908–914, May 2017, doi: 10.3174/ajnr.A5127.

[50] J. Dong and M. Xu, "A 19-miRNA Support Vector Machine classifier and a 6-miRNA risk score system designed for ovarian cancer patients Corrigendum in

/10.3892/or.2019.7385," *Oncol. Rep.*, vol. 41, no. 6, pp. 3233–3243, Jun. 2019, doi: 10.3892/or.2019.7108.

[51] H. M. Ahmad, S. Ghuffar, and K. Khurshid, "Classification of Breast Cancer Histology Images Using Transfer Learning," in *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, Jan. 2019, pp. 328–332. doi: 10.1109/IBCAST.2019.8667221.

[52] W. Zhi, H. W. F. Yueng, Z. Chen, S. M. Zandavi, Z. Lu, and Y. Y. Chung, "Using Transfer Learning with Convolutional Neural Networks to Diagnose Breast Cancer from Histopathological Images," in *Neural Information Processing*, vol. 10637, D. Liu, S. Xie, Y. Li, D. Zhao, and E.-S. M. El-Alfy, Eds. Cham: Springer International Publishing, 2017, pp. 669–676. doi: 10.1007/978-3-319-70093-9_71.

[53] R. Yan *et al.*, "Breast cancer histopathological image classification using a hybrid deep neural network," *Methods*, vol. 173, pp. 52–60, Feb. 2020, doi: 10.1016/j.ymeth.2019.06.014.

[54] B. Wei, Z. Han, X. He, and Y. Yin, "Deep learning model based breast cancer histopathological image classification," in *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Apr. 2017, pp. 348–353. doi: 10.1109/ICCCBDA.2017.7951937.

[55] H. Chougrad, Z. Hamid, and A. Omar, "Bag of Features Model Using the New Approaches: A Comprehensive Study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, Jan. 2016, doi: 10.14569/IJACSA.2016.070132.

[56] E. Karami, S. Prasad, and M. Shehata, “Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images,” p. 5.

[57] C. Kaushal and A. Singla, “Analysis of Breast Cancer for Histological Dataset Based on Different Feature Extraction and Classification Algorithms,” in *International Conference on Innovative Computing and Communications*, Singapore, 2021, pp. 821–833. doi: 10.1007/978-981-15-5113-0_69.

[58] S. D. Roy, S. Das, D. Kar, F. Schwenker, and R. Sarkar, “Computer Aided Breast Cancer Detection Using Ensembling of Texture and Statistical Image Features,” *Sensors*, vol. 21, no. 11, Art. no. 11, Jan. 2021, doi: 10.3390/s21113628.

[59] R. Pal and M. Saraswat, “A new weighted two-dimensional vector quantisation encoding method in bag-of-features for histopathological image classification,” *Int. J. Intell. Inf. Database Syst.*, vol. 13, no. 2–4, pp. 150–171, Jan. 2020, doi: 10.1504/IJIIDS.2020.109453.

[60] N. G. Polson and S. L. Scott, “Data augmentation for support vector machines,” *Bayesian Anal.*, vol. 6, no. 1, pp. 1–23, Mar. 2011, doi: 10.1214/11-BA601.

[61] S. T. Mpinda Ataky, J. de Matos, A. de S. Britto, L. E. S. Oliveira, and A. L. Koerich, “Data Augmentation for Histopathological Images Based on Gaussian-Laplacian Pyramid Blending,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1–8. doi: 10.1109/IJCNN48605.2020.9206855.

[62] K. Faryna, J. van der Laak, and G. Litjens, “Tailoring automated data augmentation to H&E-stained histopathology,” presented at the Medical Imaging with Deep Learning, Feb. 2021. Accessed: Sep. 28, 2021. [Online]. Available: https://openreview.net/forum?id=JrBfXaoxbA2

[63] Y. W. Jin, S. Jia, A. B. Ashraf, and P. Hu, “Integrative Data Augmentation with U-Net Segmentation Masks Improves Detection of Lymph Node Metastases in Breast Cancer Patients,” *Cancers*, vol. 12, no. 10, Art. no. 10, Oct. 2020, doi: 10.3390/cancers12102934.

[64] D. Bardou, K. Zhang, and S. M. Ahmad, “Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks,” *IEEE Access*, vol. 6, pp. 24680–24693, 2018, doi: 10.1109/ACCESS.2018.2831280.

[65] N. Bayramoglu and J. Heikkilä, “Transfer Learning for Cell Nuclei Classification in Histopathology Images,” in *Computer Vision – ECCV 2016 Workshops*, Cham, 2016, pp. 532–539. doi: 10.1007/978-3-319-49409-8_46.

[66] V. G. Buddhavarapu and A. A. J. J, “An experimental study on classification of thyroid histopathology images using transfer learning,” *Pattern Recognit. Lett.*, vol. 140, pp. 1–9, Dec. 2020, doi: 10.1016/j.patrec.2020.09.020.

[67] N. Dimitriou, O. Arandjelović, and P. D. Caie, “Deep Learning for Whole Slide Image Analysis: An Overview,” *Front. Med.*, vol. 6, 2019, doi: 10.3389/fmed.2019.00264.

[68] J. Chang, J. Yu, T. Han, H. Chang, and E. Park, “A method for classifying medical images using transfer learning: A pilot study on histopathology of breast cancer,” in *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, Oct. 2017, pp. 1–4. doi: 10.1109/HealthCom.2017.8210843.

Appendices

Appendix A. Renewed Certificate of Ethics Approval from the University of Ottawa Office of Research Ethics and Integrity

16/11/2020

Université d'Ottawa
Bureau d'éthique et d'intégrité de la recherche

University of Ottawa
Office of Research Ethics and Integrity

CERTIFICAT D'APPROBATION ÉTHIQUE I CERTIFICATE OF ETHICS APPROVAL

Numéro du dossier / Ethics File Number	H-11-19-5230
Titre du projet / Project Title	Application of machine learning to identify distinct patterns of tissue damage in placenta histopathology images
Type de projet / Project Type	Recherche de professeur / Professor's research project
Statut du projet / Project Status	Renouvelé / Renewed
Date d'approbation (jj/mm/aaaa) / Approval Date (dd/mm/yyyy)	11/11/2019
Date d'expiration (jj/mm/aaaa) / Expiry Date (dd/mm/yyyy)	10/11/2021

Équipe de recherche / Research Team

Chercheur / Researcher	Affiliation	Role
Shannon BAINBRIDGE-WHITESIDE	École interdisciplinaire des sciences de la santé / Interdisciplinary School of Health Sciences	Chercheur Principal / Principal Investigator
Purvasha PATNAIK	École interdisciplinaire des sciences de la santé / Interdisciplinary School of Health Sciences	Étudiant-chercheur / Student-researcher
Adrian CHAN	École des sciences de l'activité physique / School of Human Kinetics	Co-chercheur / Co-investigator
Anika MUKHERJEE	École interdisciplinaire des sciences de la santé / Interdisciplinary School of Health Sciences	Étudiant-chercheur / Student-researcher

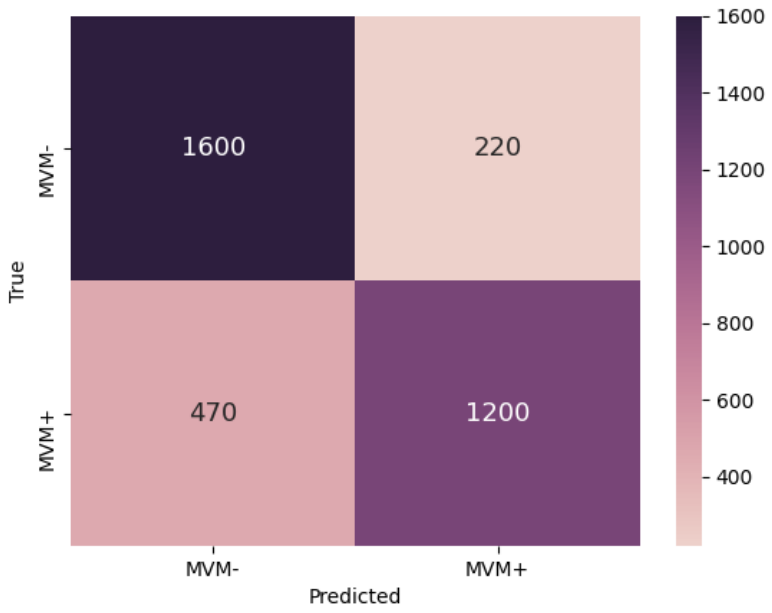
Conditions spéciales ou commentaires / Special conditions or comments

550, rue Cumberland, pièce 154 Ottawa (Ontario) K1N 6N5 Canada
550 Cumberland Street, Room 154 Ottawa, Ontario K1N 6N5 Canada

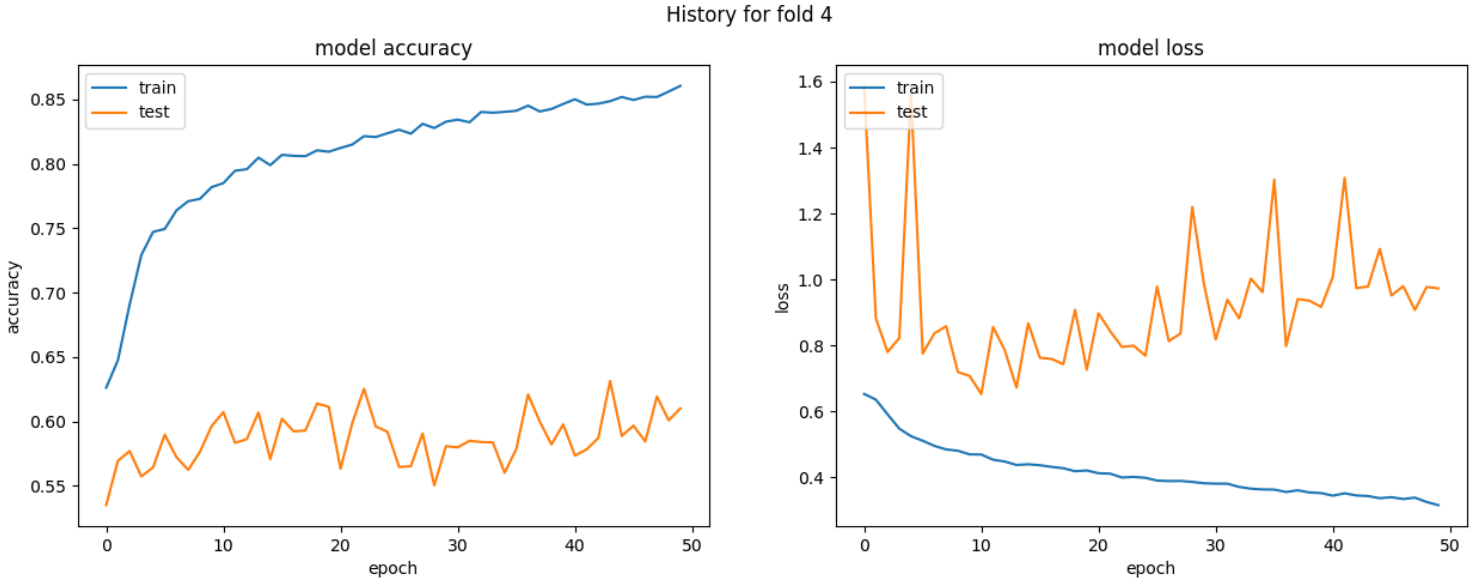
613-562-5387 • 613-562-5338 • ethique@uOttawa.ca / ethics@uOttawa.ca
www.recherche.uottawa.ca/deontologie I www.recherche.uottawa.ca/ethics

Appendix B. Preliminary Results from Deep Learning Experiment

B1. Confusion matrix for a deep learning model with data augmentation and k-fold cross validation for a MVM-/MVM+ binary classification scheme using a threshold between the cumulative score 2 and 3. Quadrants in the confusion matrices represent true negatives (top left), false negatives (bottom left), true positives (bottom right), and false positives (top right). The number of images classified for these four groups is represented by a colour scale; where light pink represents low numbers and dark purple represents high numbers.



B2. Accuracy and loss per epoch for a deep learning model with data augmentation and k-fold cross validation for a MVM-/MVM+ binary classification scheme using a threshold between the cumulative score 2 and 3. Values across k-fold cross validation (k=5) averaged 0.80 (standard deviation = 0.01).



Appendix C. Computer Aided Classification of Histopathology Images: A Systematic Review to Inform Best Practices

[1]Anika Mukherjee, BSc; [2,3]Michal Leckie, BSc; **[1]Purvasha Patnaik, BSc**; [4]Adrian D.C. Chan, PhD, PEng; [3,5]David Grynspan, MD, FRCPC; [1,6]Shannon A. Bainbridge, PhD

[1]Interdisciplinary School of Health Sciences, University of Ottawa, Ottawa, Ontario, Canada

[2]Department of Physiology, University of Toronto, Ontario, Canada

[3]Department of Pathology and Laboratory Medicine, Children's Hospital of Eastern Ontario, Ottawa, ON, Canada

[4]Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada

[5]Department of Pathology, Vernon Jubilee Hospital, Vernon, British Columbia, Canada

[6]Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario, Canada

Corresponding author (email address and complete address): Shannon Bainbridge, Interdisciplinary School of Health Sciences, University of Ottawa, Roger Guindon Hall, Rm 2116, 451 Smyth Rd, Ottawa, Canada K1H 8M5; Tel.: 1-613-562-5800 ext. 8569; E-mail: shannon.bainbridge@uottawa.ca.

Funding: D.G and S.A.B are funded by the Physicians Services Inc (PSI), S.A.B is funded by the Canadian Institute of Health Research (CIHR, grant 130543 CIA and 142298 MOP), A.M. and P. P are funded by the Faculty of Health Sciences at the University of Ottawa.

Abstract:

Context: Computer-aided diagnosis (CAD) has revolutionized several domains of medical image analysis over the past few decades but has yet to be adopted into clinical pathology practice. Clinical histopathology examination is a prime candidate for CAD-based applications, as issues of procedural and diagnostic standardization have plagued this field for years due to the subjective nature and high degrees of intra- and inter-personal/institutional discrepancies in reported diagnoses. A significant deterrent to clinical adoption in pathology is the lack of a standardized protocol for CAD model development for applications to histopathology images, which makes safety and efficacy testing a challenge. Having a standard protocol for the development of such tools is necessary to provide a gold-standard against which models can be evaluated, and once in place, it would facilitate their transition to clinical use.

Objective: To conduct a systematic review of the current methods of CAD development and application to histopathology image datasets and provide recommendations for best practice.

Data Sources: The databases Medline, Embase, Scopus, and IEEE Explore were used to identify all applicable articles.

Methods: A total of 230 articles were retrieved from a MeSH term search in the 4 academic databases listed above. After three authors conducted full-text screening with moderate agreement (weighted Kappa scores of 0.47-0.71), 18 articles remained to be analyzed. Articles were assessed for common methodological elements and for the performance of machine learning classifiers on histopathology images. The results reported, discussion in the articles, and information from current literature were used to propose a set of recommendations to

guide researchers in applying the 5 steps to develop machine learning tools on histopathology images.

Results: All studies that passed through screening had 5 common methodological elements: image acquisition, image preprocessing, tissue feature extraction, pattern recognition and classifier training, and algorithm testing. One or more recommendations were proposed for each of the 5 elements based on the findings in the articles and the available literature.

Conclusions: While a wide variety of approaches were used, all studies followed a 5-step workflow to develop classification models for histopathology images. The articles assessed focused uniquely on onco-histopathology images, but due to the adaptive nature of machine learning, the techniques presented can be applied to datasets relevant to other disease states. The guidelines proposed should be used to inform experimental design in future machine learning projects on histopathology images across pathology domains.

Introduction:

Histopathology analysis is one of the primary methods for assessing disease progression in a variety of tissues, but continues to face challenges in an ever evolving clinical environment[1,2]. Analysis of whole slide images (WSI) captured from histological specimens requires pathologists to detect subtle visual patterns in complex tissues, which can be challenging, time-consuming, and lead to high rates of inter-observer reporting differences[3,4]. As such, a strong case has been made for the development and integration of computer aided diagnostic (CAD) tools for clinical histopathology image analysis[3,5,6]. CAD applications in the field of pathology use mathematical models to classify an input image according to a clinical output of interest, such as diagnosis or prognosis, helping to standardize the evaluation of clinical specimens and provide decision support for pathologists[3–5]. Historically, developing such mathematical models on histopathology images was nearly impossible due to the complexity and size of high-resolution WSI[4,7,8]. However, the increasing popularity and accessibility of machine learning (ML) model development has rendered CAD applications within reach for this clinical domain[9].

Supervised machine learning algorithms allow for automated model development with minimal human intervention, with the model "learning" inherent characteristics in the training dataset.[10–12] In the case of histopathology image analysis, a branch of machine learning termed Computer Vision is applied, in which the model is trained to recognize distinct *visual* features in the images that are unique to the labeled clinical outcome of interest (i.e. diagnosis, prognosis, etc). [13–15] Once the model is adequately trained, it can predict (i.e. classify) the clinical outcome of interest in a set of unseen and ultimately unlabeled images, based on the absence or presence of these same visual characteristics[16]. Machine learning model development and

assessment, in the context of Computer Vision applications, typically follows 5 fundamental steps [4] (Figure 1):

Step 1) *Image acquisition and data preparation*[4,8,14,17,18] – Digital images are captured from clinical specimens of interest. Tissue samples can be acquired from large public databases or from private hospital collections. Images are annotated or labeled according to the desired clinical categories to be predicted by the model. At this stage, all original annotated image data can be split into model training, validation and testing datasets.

Step 2) *Image pre-processing*[4,14] - Image transformations and data augmentation are conducted on all images. These transformations help to normalize non-biological visual variations across specimens and to augment the size of the original dataset to ensure robust model development.

Step 3) *Tissue feature extraction*[4,14,17] – Automated extraction of visual features of interest is conducted, ideally extracting features that may be different across the groups of clinical interest[17]. Dimension reduction is often applied at this step[4,19], aiming to reduce the computational power required to process all of the data and remove redundant features.

Step 4) *Pattern recognition and classifier training (**model development**)*[4,14,17,18] – An iterative process in which the algorithm "learns" visual characteristics in each image, unique to the clinical labels applied to the training dataset. The model is punished for errors in classification until the rate of correct classification improves and reaches a stable classification accuracy[4,20]. The resulting product is a trained classifier (model) that can predict a clinical classification label based on the rules adopted during training. Cross-validation can be applied to fine-tune model training and can provide a reliable estimate of classifier performance [21,22].

Step 5) *Model testing*[4,14] - The developed classifier is tested on an unseen image dataset – the testing subset[22]. Model classification performance is assessed through comparison of model outputs to ground truth image labels.

Despite considerable advances in CAD-based research focused on clinical applications, to date we have yet to see a translation of this technology into clinical practice. The highly technical language used in the majority of papers on this topic, coupled to a lack of standardized protocols for the development, application and assessment of histopathology-based CAD tools have certainly been limiting factors for clinical translation. A 2013 survey on CAD support systems declared insufficient testing and validation as the primary barrier for clinical adoption[23], with the identification of published systematic methods for developing and assessing the utility of these tools identified as key to overcoming this barrier. As such, the purpose of this systematic review was to provide a comprehensive review of the CAD model development and validation protocols that have been reported to date for clinical histopathology image analysis. It is anticipated that by consolidating this information and proposing an accessible standard guideline for best practices, the barriers for clinical adoption of CAD technology in histopathology can be overcome.

In the current review, a systematic analysis of 18 articles which applied the above-mentioned Computer Vision workflow for purposes of histopathology analysis is presented, including a comparison of model performance according to methods used. The findings are an important first step in the identification of the best practice for applying commuter-aided diagnosis in the clinical pathology environment.

Methods:

Search strategy

The York protocol for systematic reviews[24] was employed, using a search of 4 primary biomedical, clinical science, and engineering databases (Medline, Embase, Scopus, and IEEE). A combination of relevant MeSH terms, keywords, and word variants were used (Supplementary Table 1). A scan of all identified articles was performed to remove any duplicate articles.

Inclusion/Exclusion Screening

All articles were assembled in reference manager software (Mendeley), and 3 reviewers (A.M., D.G., and M.L.) performed the initial screening of titles and abstracts for inclusion/exclusion criteria. Primary inclusion/exclusion criteria were: 1) a primary research aim which included image segmentation or image classification; 2) methods that were being applied to digital histopathology specimens from humans; 3) primary objective related to clinical utility of CAD technology (not solely for the purpose of basic scientific investigation). All tissue and disease subtypes were included provided they met the above criteria. Conference proceedings, provided all other criteria were met, were included, however reviews, book chapters, and papers discussing methods that are not routinely used in clinical pathology practice (ex. focused exclusively on nuclear texture, immunohistochemistry quantification, or prognostic immunohistochemistry) were excluded. The level of reviewer agreement was assessed by comparing screening decisions from each reviewer and measured using a weighted Kappa score[25]. Disagreements were resolved through joint discussion until a consensus was reached. Subsequently, a full-text screening was performed by 2 reviewers (A.M and M.L.) on articles

that passed the initial screening to ensure they met the inclusion/exclusion criteria listed above.

Data Extraction

Authors A.M. and M.L. systematically extracted methods presented in each article according to the five standard steps of a computer vision workflow (Figure 1). An excel data extraction template was devised to include the following categories found to be common to all articles: tissue type, pre-processing methods, feature extraction methods, dimension reduction methods, classifier(s) used, cross-validation methods, training/testing data split, performance of model. Each reviewer used this template to collect raw data from each article, which was later refined to fit into the 5 standard steps of a computer vision workflow.

Quality Assessment

A risk of bias and assessed applicability analysis of each article was performed using the QUADAS-II tool for diagnostic tests of accuracy (under the Cochrane Guidelines for Systematic Reviews)[26]. To our knowledge, there are no quality assessment tools unique to CAD-based diagnostic application, however the flexibility with QUADAS-II tool allowed for customization of this analysis to the current literature dataset. Reviewers A.M. and P.P. assessed the following categories for each paper: 1) Patient Selection, 2) Index Test (new diagnostic test), 3) Reference Standard (original validated protocol for diagnosis), and 4) Flow and Timing (process patient undergoes to compare reference standard and index test). Risk of bias for articles for the Reference Standard and Flow and Timing criteria were judged to be unclear when there was no information on how the pathologists assigned ground truth labels.

Results:

Article Selection

The initial article search yielded 230 articles, with 209 removed following screening of the title and abstract and another 3 removed following full text screening for inclusion/exclusion criteria. Reviewer agreement on the inclusion of articles during screening was moderate to good based on weighted kappa scores[27]. Weighted kappa scores between reviewers were: 0.71 (A.M. - D.G.), 0.55 (D.G. - M.L.), and 0.47 (A.M. - M.L). 10 articles were discussed for discrepancies in screening to reach a consensus on articles to be included. The final dataset comprised of 18 articles that met the specific criteria outlined above (Figure 2). A list of the papers included for analysis by author and year is included in Table 1. All 18 articles applied their methods to cancer-related histopathology specimens including brain, breast, colon, kidney, larynx, ovary, prostate, skin, and thyroid cancer (Table 2).

Computer Vision Methodology

i. Primary Research Aim

All articles analyzed had a primary research aim of automating diagnostic classification of clinical histopathology images. Small differences in identified research objectives stemmed from differences in tissue and/or disease subtypes included in the datasets. For example, prostatic adenocarcinoma severity is measured using the Gleason scale[28], so training labels/classifiers used were in reference to the corresponding Gleason scores[29,30]. All articles used either a non-biased approach, using textural patterns in the histopathological images[31], or relied on explicit segmentation of biologically relevant features (i.e. nuclei)[32]. In individual instances, articles explored both approaches described above, along with many different classifiers, and reported and compared the results from each[33]. Most papers used visual

elements of an image to train an image classifier, with the goal of automating prediction of clinical diagnosis in unseen images (ex. Benign vs malignant tumor). However, Ninos et al.[34] and Yu et al.[35] used their classifiers to predict patient survival rates, and were included as the methods used followed the same 5-step computer vision workflow described (Figure 1). Image classification categories (i.e. labels) reported in the included papers were: Gleason score[29,30], benign vs malignant tumor[33,36], glioblastoma stages[37,38], American Joint Committee on Cancer (AJCC) melanoma stages[31], ovarian carcinoma subtypes[39], renal carcinoma subtypes[40], tumor vs non-tumor[35,41–43], thyroid follicular adenoma vs. thyroid follicular carcinoma vs. healthy thyroid[36], colorectal cancer subtypes[44], good vs. poor prognosis[34,35], low vs. high grade laryngeal cancer[45], and healthy thyroid vs. papillary thyroid carcinoma[32].

ii. Image Acquisition (Computer Vision Workflow Step 1)

A number of studies used large publicly available histopathology image databases for image acquisition. These included the BreakHis dataset[33,38], The Cancer Genome Atlas (TGCA)[35,46,47], the Trans Canadian Ovarian Cancer Dataset[39], the Stanford TMA database[35], the Cooperative Human Tissue Network (CHTN)[31] and the PETAACC-3 dataset[44]. The articles that did not use a public database retrieved and digitized a small number of archived samples from a local hospital or university[29,34,36,45,48] or did not specify the origin of the histopathology images used[40–42,49,50]. Sample sizes varied widely across studies (Table 1), ranging from 7-742 individual patients and 15-21860 individual images, with some groups failing to indicate the number of patient samples used[37,43]. Only Xu et al[43]. mentioned using a sample size for each class that reflects the true prevalence of the lesion(s) in question.

Histopathology specimens were prepared from original tissue samples using standard procedure: tissue fixation, embedding in paraffin wax, and sectioning with a microtome[51]. Embedded tissue sections were then mounted on glass slides, stained with H&E, p63, Feulgen stain, or DAB, covered with a glass slip, and digitally scanned. Images were taken with a light microscope paired with an RGB digital camera or with a high-resolution glass tissue slide scanner, most commonly at a magnification of 40X. Some databases had previously acquired images at multiple resolutions (40X, 100X, 200X, 400X). The digital image file types most commonly used were Whole Slide Images (WSI) and Tissue Microarrays (TMA), and were in the SVS file format which contains a pyramid of images at multiple magnifications[52]. Images included in the pyramid usually range from 240x240 pixels to 1024x768 pixels. SVS files of scanned histopathology tissue were often greater than 1 Gb in size[7].

iii. Image Pre-processing (Computer Vision Workflow Step 2)

Image pre-processing is the process of applying a series of transformations to images before feeding them to a classifier. This step is often done in preparation for feature extraction to ensure all images are uniform[4]. Four pre-processing techniques were identified in the article dataset: 1) augmentation; 2) colour transformations; 3) data cleaning; and, 4) automatic segmentation. All articles used some combination of these four methods. Yang et al.[36] did not report using any preprocessing on their data. Table 3 contains a summary of image pre-processing methods used in each article. Selective discarding of poor-quality tissue sections or images is included as a separate category in Table 3, however, for simplicity it is categorized under data cleaning in the description below.

Data augmentation involves increasing the number of examples a classifier can learn from. Data augmentation is achieved through random image transformations such as rotation, flipping, and stretching at each round of training and/or by extracting multiple image patches from the original image prior to training, increasing the size of the training dataset[18,53]. Data augmentation is particularly crucial in computer vision applications, allowing developers to address the requirement for large training datasets in the face of ethical and/or financial constraints which may limit the number of original patient samples available for ML development[53,54]. This technique also provides an opportunity to address inherent problems in a dataset, such as class imbalance. In this case, under-represented clinical classes can be specifically augmenting in the dataset, removing any possible bias introduced in the model training phase through uneven class distribution[54,55]. In articles reviewed, fifteen of eighteen articles reported using data augmentation in their workflow. The primary method of data augmentation was multiple image patch extraction. Extracted image patchs ranged in size from 32x32 to 2000x2000 pixels, sometimes capturing hundreds of small overlapping patches from across the original image, or randomly selecting a small number (10-20) of representative patches from each original image. In some cases, the augmentation method used included the extraction of patches from the same image at multiple magnifications[33]. Bayramoglu et al.[38] also described rotating and flipping images as part of their augmentation method, further helping a classifier learn more variations of image features.

Many articles reported manipulating the image colour in the pre-processing step. Either the images were converted to greyscale, or they were converted to HSI (Hue, Saturation, Intensity), HSV (Hue, Saturation, Value), or CIE L*a*b* (Commission internationale de l'éclairage –

Lightness*, Red-Green*, Blue-Yellow*) colour spaces[55,56]. Converting images to greyscale reduces the complexity of an image, reducing the image from 3-4 channels (ex. red, green, blue (RGB)) to 1 channel (grey)[55]. Greyscaling images also helps to highlight dark areas, which can facilitate segmenting objects of interest, such as nuclei[37,55]. Non-greyscale colour transformations, such as converting an RGB image to HSV format, were used in one article that tried to mimic the process a pathologist would use to analyze images[42]. In this article, colours closer to those interpreted by the human eye were chosen, rather than simply RGB colours[55]. Three groups deconvolved the images into separate hematoxylin and eosin channels because they found less redundancy in the features extracted from those channels as opposed to the RGB channels. DiFranco et al.[30] tested several colour channels to determine which (alone or in combination) yielded better classification results and found that b* channel features from a CIE L*a*b* colour space provided optimal results.

Data cleaning is another pre-processing step that involves inspection of the data in order to remove outliers and images containing little or unusable data[4]. In the case of histopathology images this could include artefacts on slides, folds in tissue, areas of inconsistent staining[57], or poor scanning quality[52]. To mitigate this problem, several authors chose to have a pathologist manually select regions of interest within slides[30,33,34,38,42,48], discard images or patches of poor quality[33,38,42,48], automatically discard images containing too little information[39,48], or exclude objects based on size threshold[45]. In the context of computer vision, data cleaning may also include reducing the complexity of images. Articles in this search reported using smoothing, noise reduction, and increasing contrast of image datasets to amplify important patterns and facilitate feature selection. For histopathology images, normalization during the data cleaning

stage of pre-processing can specifically address a common issue unique to histopathology specimens - differences in staining intensity[58]. Normalization involves modifying the intensity of the pixels in an image by a scaling factor to achieve consistency between different images[18,58,59]. A common approach was Min-Max scaling, which shifts the possible pixel values to a consistent range for all images[60].

While some articles chose to classify images based on the global appearance of images, others felt it more pertinent to segment objects of interest prior to classification. Automatic segmentation is used to pre-select and extract objects or regions of interest within an image that are relevant to the clinical outcome (ex. nuclei from malignant cancer cells)[61]. Seven articles reported using segmentation to isolate regions of interest in images prior to classifier training. Open source and commercial software packages, such as CellProfiler, OpenCV, and MATLAB, are currently available to aid in image segmentation, and were used in several of the reviewed articles[33–35,37,39,40,62]. Algorithms used to segment nuclei included random field graph cut[63], active contours[64], fuzzy c-means clustering[65], PSO-based Otsu level thresholding[66], density-based clustering[67], and curvelet transform[29]. Most of the algorithms described above are available through free software such as ImageJ[68], or can be implemented with common coding languages such as Python. However, Curvelet transform was a novel method developed by Lin et al.[29]

iv. Tissue Feature Extraction (Computer Vision Workflow Step 3)

Feature extraction involves converting an image from a raw set of pixels to a numeric representation of an image with real-world relevance (ex. number of cancerous nuclei in an image), or that is less computationally expensive to analyze (ex. quantifying textures found in

an image)[4]. In the reviewed articles, global texture feature extraction[30,31,33,36,41,42,44] or object level (clinically relevant) feature extraction[48,49] methods were employed, with some groups testing and comparing both approaches[29,34,35,37,39,40]. As well, two articles extracted keypoint descriptors, which are points of interest in an image that can be used to identify it at any scale[33,43]. Ninos et al.[34] additionally included patient information as a feature to train their classifier.

To carry out global textural and spatial feature extraction there are several programs available to researchers, with the feature extraction toolboxes provided in MatLab and CellProfiler[62] used in some of the reviewed articles to extract over 9000 features[30,31,34,35,40,43]. Global features described in the articles retrieved can be categorized into: Local Binary Pattern (LBP)[69], Local Phase Quantization (LPQ)[70], Grey-Level Co-occurrence Matrix (GLCM)[71], Haralick features[71], wavelet features[72], and Gabor features[73], and keypoint descriptors[74,75]. The most commonly used Keypoint descriptors include SIFT[76], SURF[77], ORB[78], FAST[79], and BRIEF[33,80]. We direct the reader to Spanhol et al.[33] for a description of LBP, LPQ, GLCM, and keypoint features, and to the Appendix in Wang et al.[48] for a brief description of Haralick features, wavelet features, and Gabor features.

A total of 9 articles instead used an object level feature extraction approach, with common extracted features including area, convexity, and circularity[29,34,35,37,39,40,45,48,49]. Object level features are attractive for analysis on histopathology images because they are computed on objects pre-determined to be clinically relevant in an image (ex. nuclei, tumour regions)[4,81]. The drawback is that this step may introduce human bias or error into a classifier if objects of interest are manually segmented, or if thresholds for automatic thresholding are inappropriate.

Object level features include elliptical, convex hull, bounding box, boundary, and other shape features[37]. Table 1 in Fukuma et al.[37] presents a comprehensive list of object level features and their organization into the five categories listed above. These shape features are commonly used to distinguish cancer and healthy samples in digital pathology, and they describe shape elements such as area, perimeter, convexity, and eccentricity of an object[48].

The articles we examined reported extracting between 10 – 9878 features per image. When too many features are used, some may be redundant and no longer help the classification algorithm discriminate between classes. In these cases, dimensionality reduction is used to reduce the length of the feature vectors that describe each image and increase the efficiency of classification[82]. Dimension reduction methods reported in the articles retrieved include: Greedy Hill Climbing[83], Pearson's Correlation[84], mRMR[85], Linear Discriminant Analysis (LDA)[86], Dissimilarity Matrix[87], Fisher Score[88], Wilcoxon's Correlation[89], Variable Precision Rough Sets (VPRS)[90], Mean Rank Plots (MRP)[91], Principal Component Analysis (PCA)[92], Multiple-Dimensional Scaling (MDS)[93], Local Linear Embedding (LLE)[94], Isometric Feature Mapping (ISOMAP)[95], Local Tangential Space Embedding (LTSA)[96], Point Biserial Correlation[97], and Information Gain Ratio[98]. Yang et al.[36] tested MDS, PCA, LLE, ISOMAP, and LTSA with 3 to 4000 dimensions and reported that ISOMAP performed the best. Jothi et al.[99] described and tested the VPRS based b-reduct methods with a closest matching rule classifier and achieved 100% classification accuracy. Rough Set Theory[100] (the basis of VPRS) has previously been used to reduce image features in various medical image types such as mammograms, Computed Tomography, and Magnetic Resonance Images[99]. An alternative approach described by Yu et al.[35] and Orlov et al.[31] was to determine the optimal number of features using the cross-validation results. DiFranco et al.[30]

also used a random forest classifier, which provides feature importance by counting the number of times a feature is selected in each decision tree. (See Results – Classifiers and cross-validation). Table 4 includes a summary of feature extraction and dimensionality reduction methods included in each paper.

v. Pattern Recognition and Classifier Training (Computer Vision Workflow Step 4)

Classifier training is the fundamental stage in developing an ML model where it sets rules by which the algorithm will distinguish between classes[4]. Fine-tuning of the training is carried out using cross-validation methods which can provide a reliable estimate for how well the model is performing by plotting a learning curve that shows the rate of improvement of performance on a portion of the training data over rounds of training. A summary of classifiers and cross-validation methods used in the selected articles is included in Tables 5. Types of classifiers applied in the selected articles were: Support Vector Machine (SVM)[101], Nearest Neighbors[102], Decision Tree[103] and Random Forests[104], Bayesian[105], Boosting[106], Neural Networks[107], Multiple Instance Learning (MIL)[108], Discriminant Functions[109], and novel classifiers[32]. Differences between classifiers lie in the mathematical approach to creating a decision boundary. Some articles chose to apply ensemble-based learning, where the output from multiple classifiers was used to improve the overall accuracy[30]. Support vector machine (SVM) was a popular choice of classifier in the article dataset, appearing in at least one experiment in 11/18 articles[29,30,33,35–37,39,42–44,48]. The SVM classifier finds a set of hyperplanes in the n-dimensional feature space that has the longest distance to the nearest training points of other classes[48]. They can be paired with different kernels (linear, quadratic, radial based Gaussian) to fit the data better. This

classifier works well for large feature spaces, which usually is the case for histopathological image data.

In 8 articles reviewed[30,33–36,43,45,48], cross validation methods were used to fine tune the parameters used during the training phase and assess computational needs using a learning curve, although some articles[29,31,32,37–40,44] only reported using cross validation to evaluate model performance (*Computer Vision Workflow Step 5*) or employed both uses of cross validation[34,36,43,45,48]. Broadly, this method involves splitting the training dataset into smaller training/validation sets (ex. training set split into 80% for training and 20% for validation) and evaluating the rate of improvement of performance on the portion of data reserved for validation in order to tune parameters and assess how much computing resources are required for a project. Types of cross-validation methods reported in the reviewed articles included K-fold[110], Leave-One-Out (LOO)[111], Sequential Backwards Selection[112], Jackknife[113], and Bootstrap[113]. Differences between cross-validation methods stem from how training dataset is split at each round of training. LOO, Bootstrap and Jackknife cross-validation involve testing every possible split of data to obtain optimal results, whereas the others involve splitting the data into a discrete number of portions for validation.

vi. Algorithm Testing (Computer Vision Workflow Step 5)

To assess the performance and generalizability of a developed ML model, an unseen test dataset is fed into the model and the predicted output labels are compared against the "Ground Truth" labels of the dataset. All articles reviewed were retrospective studies comparing image classification based on manual pathology evaluation (ground truth) vs. ML

model evaluation on the same task. ML model performance was measured as a combination of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These can be combined in many ways to punish errors that are less acceptable in a given problem. In the articles analyzed, accuracy ((TP + TN) / number of images) was most commonly used as the primary performance measure[29,31,33,34,36–42,44,45,48,49]. Reported accuracies for the methods presented ranged from 65% - 100%. Orlov et al.[31], Wang et al.[48], and Jothi et al.[32] reported achieving 100% accuracy in at least one developed model. They used a nearest neighbor classifier, an SVM, and an SVM and Decision tree combination respectively to achieve these results. A summary of the optimal model performance achieved in each article is summarized in Table 6.

Risk of Bias

The QUADAS-II tool was used to assess the risk of bias and applicability of the included articles, with the results summarized in Figure 3. Overall, the risk of bias was low for all articles, except for rare cases where one of four categories of assessment were considered high risk[37,40,42–44]. In most cases, the reason for the high risk of bias was that it was unclear whether or not patient selection was biased[40,42–44]. Several of the articles accessed cancer histopathology images from a database of samples previously collected, and authors did not disclose whether they recruited representative patients, whether patients were randomly enrolled, or whether there were any inappropriate exclusions of patients. The articles by Budinska et al.[44], Mete et al.[42], and Yu et al.[35] were deemed to have a high risk of bias for the Patient Selection criterion, as they did not describe how patients were selected and did not include a clinically representative distribution of cases. Yu et al.[35] explicitly stated in their discussion that because they obtained their cases

from a public database (TCGA), there was an overrepresentation of cases where morphological patterns of disease were obvious and might not reflect what a pathologist would see in clinical practice[35].

For the most part, we found that articles scored low on the risk of bias for the Index Test criterion. Contrary to a standard diagnostic test of accuracy, where a clinical researcher would be applying a new test and comparing it to previous standards, in the articles retrieved, a computer performed the new test analysis. An automated tool removes subjectivity and ensures the reproducibility of results. In some articles, it was unclear whether regions of interest were selected or discarded systematically, so the risk of bias for the index test criterion could not be definitively assessed. The article by Fukuma et al.[37] demonstrated a high risk of bias, as they did not split their data into training and test sets and they did not report results from all classifiers. If their classifier was evaluated on the same data it was trained on, the accuracy would be falsely high. The assignment of ground truth labels was often unspecified in cases where images were obtained as a secondary source from a database. Finally, there were few concerns regarding the applicability of the articles to the original research question for this study, as very stringent inclusion and exclusion criteria were applied to article selection in respect to this criterion.

Discussion:

The application of computer vision as a CAD aid in the field of clinical histopathology has tremendous potential, yet the complexity and wide array of methods available in this domain make the translation of these tools into the clinical environment a daunting task. One of the primary challenges in bringing histopathology CAD tools to the market is the lack of standardization in the methods used to develop the models, making it difficult to determine which models would be best suited and have the highest performance metrics in a clinical setting[23]. While a large variety of methods were observed in the systematic review of the literature on this topic, a clear 5-step workflow was identified as the common thread to all studies (Figure 1). Interestingly, the application of a 5-step workflow was not used as an explicit inclusion criterion for this review, but instead was an inherent characteristic of the all reviewed study designs. While a clear consensus on best practices to be used within the 5-step workflow framework was not apparent in the individual studies reviewed, the high-level overview provided through this systematic evaluation has allowed for the generation of recommendations of current best practices for each of the 5 workflow steps, as described below (Table 7).

Image Acquisition (Computer Vision Workflow Step 1)

The articles reviewed had a common research aim of developing a computerized model that can take a histopathology image as an input and output a clinically-relevant label of interest, such as diagnosis or prognosis. Despite not having excluded any type of histopathology samples in the screening criteria, the articles represented in the search were overwhelmingly focused on

cancer diagnostic aids. This was not surprising considering the large publicly available digital image databases dedicated to cancer-specific pathologies (ex. The Cancer Genome Atlas), facilitating these types of CAD-application initiatives. A considerable strength of using image datasets from these public repositories is access to large datasets with minimal costs and often a high degree of quality control as it relates to image quality and annotation – aspects that can be highly variable in smaller, institution-based collections. This can allow for better model comparison across research groups as well. Unfortunately similar large-scale, highly annotated public image datasets for tissue types and pathologies outside of cancer are not currently available, a considerable limitation that has no doubt hindered paralleled progress in the development of CAD-based tools for other important clinical domains, such as perinatal pathology, dermatology, gastrointestinal and hepatic pathology.

While there is no gold standard for sample size calculations that can be easily applied to determine optimal numbers of patient samples required for ML model development, it is generally acknowledged that model accuracy improves with increasing sample sizes[114,115 116]. The complexity of the machine learning algorithm being used is an important consideration, with traditional machine learning models, such as SVM, requiring fewer input images compared to more complex deep learning models. And finally, in the context of histopathology datasets, the baseline degree of variability that may be observed in the tissue of interest in both healthy and sick populations need to be considered, ensuring the dataset used for training adequately represents this variability. A small number of studies have proposed statistical methods for calculating optimal sample size, using either characteristics of the data such as number of classes and number of features (model-based approach)[117] or by learning from the rate at

which the model improves over training (inverse power law approach)[118,119]. However, even these methods yield considerable variations in estimated sample sizes, in the magnitude of hundreds of images, providing little clarity to a researcher requiring guidance in experimental design. In the reviewed articles, digital image sample sizes ranged from 15 to > 21,000, originating from 7-742 distinct patient samples – demonstrating a clear lack of guidance and consensus on sample size determination for ML model development in this domain.

A clear description of how patient samples are selected for inclusion in a study and what parameters were used for ground truth sample classification are also important considerations. To generate a robust CAD model with widespread generalizability in the clinical context, one must be sure that the patient samples being used for model development were truly representative of the clinical population of interest and that the methods used to classify each sample were clear, appropriate, and reproducible. Unfortunately, these descriptions were often lacking in the reviewed articles on this topic, properly described in only 4 of the 18 articles reviewed (Figure 3). Further, the highly subjective nature of pathological evaluations – used as ground truth classification of images – must be considered. While in the field of onco-pathology the degree of inter- and intra- evaluation variability may be lower, in other fields of study this variability can be quite high and would therefore greatly impact the performance metrics of the developed ML model.

Recommendation 1a: The use of highly annotated, publicly available digital image databases is the current gold standard for image acquisition in the development of CAD-based machine learning model development, allowing for wide access to large, high-quality datasets. Consolidated efforts should be undertaken to generate large scale, highly annotated digital

image databases for tissue types and pathologies outside of cancer to ensure the progress being made in the field of oncology can likewise be realized in other domains of pathology.

Recommendation 1b: Researchers should aim to maximize the number of individual samples included when developing machine learning models, applying data augmentation techniques to further increase training sample size and assess data size impact on model development through the generation of learning curves. Further efforts in this field should be focused on developing reliable methods of samples size determinations for ML model development experiments.

Recommendation 1c: Groups developing CAD-based machine learning models for histopathology purposes need to be explicit in how the image dataset was sourced and annotated.

Image Pre-processing (Computer Vision Workflow Step 2)

Image pre-processing includes 3 major activities, each of which was included to varying degrees in the reviewed articles: 1) data augmentation, 2) image normalization, and 3) data cleaning. Of these three activities, the biggest challenges in protocol standardization efforts will likely be focused on the first two. As indicated previously, the development of a robust model requires a considerably large image dataset in order to properly train the algorithm to detect class-specific differences in extracted features. Unique to the realm of medical imaging, access to large, highly annotated image datasets of clinical specimens can be a major limitation to CAD model development. An important technique used to overcome this challenge is data augmentation. A common approach to data augmentation is the use of a sliding window image patch extraction,

in which a number of smaller overlapping images are captured from across the entire original image. This method was employed in a total of 3 articles reviewed[29,42,43], augmenting dataset up to 7,000 fold. For researchers with limited access to samples, a sliding window approach offers a fast and systematic way to augment data. A similar approach, which attempts to mimic a pathologist's approach, extracts a smaller number of image patches from distinct or representative regions of the original image, and in the case of the reviewed articles, augmented datasets by a factor of 10-3,000 fold[40,44,45]. Importantly, the process of data augmentation through patch extraction also addresses another major barrier currently faced in histopathology image analysis – the large file sizes of these high-resolution digital images. Scanned images used in histopathology analysis often contain over a gigabyte of information[7], presenting an enormous computational task. The use of image patch extraction can reduce the size of these files to the magnitude of hundreds of kilobytes (or hundreds- thousands of pixels squared), sizes that are compatible with current computing power availability for most users. Further, opting for the development of a traditional machine learning classifier (vs. a deep learning classifier) with pre-extracted image features, as seen in 15 of 18 articles review, can also minimize the required computational power needed. Bayramoglu et. al. was the only article to mention using image modifications such as rotation and flipping to their dataset as a form of data augmentation. This is somewhat unsurprising, as it is a technique more commonly used in deep learning where it is typically embedded in the training process as a random image generator that would produce different transformations at each round of training[31,38,45]. Although this technique is useful in increasing variability and size in a smaller dataset[120], one

must be careful when applying some image manipulations (i.e. stretching), always keeping in mind the biological importance of morphological features in the tissue of interest.

Image normalization is an important consideration when working with histopathology specimens. Differences in tissue staining intensity and/or lighting differences in the captured digital images can have a considerable impact on model classification accuracy. The articles reviewed combatted this problem using a variety of image normalization strategies. In some of the articles reviewed, color transformation processes were compared[30,31,42,48], indicating that there was no clear choice for which colour space to use. Orlov et. al. compared classifier performance using RGB, CIELab, and deconvolved H&E colour spaces and found that the H and E colour space outperformed the other two by far, achieving an accuracy of 95% compared to 75% and 72% for the RGB and LAB spaces, respectively. However, the method of color normalization applied for any CAD model development will need consideration to ensure that that the features of highest relevance are being adequately highlighted for classifier training. In addition to image normalization, data cleaning includes routinely scanning for artifacts and poor quality until automating the detection and removal of histopathology-specific artifacts becomes commonplace and efficient. Data cleaning with manual removal of images or regions in 11/18 articles and was used to ensure that the features subsequently extracted represent the true morphological differences between groups, unconfounded by differences in staining or artifacts[30–35,37,41,42,45,48].

Recommendation 2a: The adoption of sliding window image patch extraction is a reasonable data augmentation approach to considerably increase the size of the training dataset, while minimizing computational power required.

Recommendation 2b: Image modifications incorporated through a random image data generator during classifier training can be used for deep learning experiments to increase size and variability of image datasets.

Recommendation 2c: While model-specific considerations must be used, H&E channel and l*a*b* color normalization methods demonstrate optimal results in their ability to best highlight distinct tissue morphology in histopathology specimens used for ML model development.

Tissue Feature Extraction (Computer Vision Workflow Step 3)

Tissue feature extraction demonstrated the highest degree of methodological variability in the review of the literature. If a fully automated approach is used for feature extraction, the introduction of human bias into model development can be minimized. However, the mathematical descriptors used to identify and extract features from each image may not necessarily be driven by the biology and/or pathophysiology underlying disease mechanism. On the contrary, if manual or semi-automatic feature extraction was guided by the expertise of a clinical pathologist - highlighting tissue areas or feature of highest relevance - clarity on how classification decisions were being mediated by the ML model can better be achieved. However, this method can be highly time-consuming and heavily influenced by human biases. Interestingly, the performance of machine learning models was similar across both feature extraction approaches in the reviewed literature, with some reviewed articles reporting classification accuracy of 100% for both approaches. This would suggest that both approaches demonstrate promise, and ultimately the developer will need weigh the pros and cons of each

approach considering the research question at hand. That being said, in order to move this technology into clinical practice, a move towards full automation of feature extraction would likely be required to address clinical work-load management. Moving forward, removing the feature extraction phase entirely, and instead having feature extraction integrated into classifier development will likely yield the most robust and clinically relevant CAD models. Historically, integrated feature extraction which is used in deep learning, has demonstrated improved model performance when applied to non-medical image datasets (ex. for facial recognition or self-driving cars) compared to traditional machine learning algorithms[121–125].

Recommendation 3: Both manual/semi-automated and fully automated feature extraction methods work well in histopathology-focused CAD tool development. A move towards fully automated, and ultimately fully integrated, approaches to feature extraction is recommended to best accommodate translation of these tools into a clinical setting.

Pattern Recognition and Classifier Training (Computer Vision Workflow Step 4)

Two distinct architectures for machine learning models exist, each with advantages and disadvantages: 1) traditional machine learning classifiers with extracted features, 2) deep learning classifiers with integrated feature extraction. In the reviewed articles, multiple classifying algorithms were tested and compared for both machine learning and deep learning algorithms. Support vector machine (SVM) and probabilistic neural network were by far the most commonly used, and while SVM demonstrated the highest ML model performance convolutional neural network (CNN) demonstrated the highest deep learning performance metrics. Although deep learning algorithms have been shown to outperform traditional

classifiers on image classification tasks, a current limitation of this approach is the inability to determine which image features were weighted more heavily in classification decisions. Deep learning is suitable in practical situations where a very high level of accuracy is essential but may pose a challenge in a clinical setting where it is crucial to be able to justify a diagnosis. However this limitation of deep learning is the primary focus of explainable artificial intelligence (XAI), a relatively new but increasingly important field of study, which will allow for the integration *and* understanding of deep learning models in a healthcare setting [126,127].

Eight of 18 articles used cross-validation to assess the rate of improvement of classification accuracy over the course of training, or the "learning", by plotting a curve of error decrease over training epochs[30,33–36,43,45,48]. Although several articles also reported using cross validation to evaluate model performance (Table 5), the results are reported and a recommendation is provided in *Computer Vision Workflow Step 4* because the topic of whether it should be used for parameter tuning or model evaluation appears to still be under debate, but a 2006 study addressing this problem found that using cross validation to evaluate a model and test its generalizability to new data gives a "significantly biased estimate of error" when it is also used to tune training algorithm parameters [128]. Moreover, two of the articles using cross-validation to estimate model performance highlighted the a major limitations to this approach which is the potential for the model to overfit to the training data and fail at generalizing to new data[30,40]. In the reviewed articles, the most widely used method of cross-validation for parameter tuning, but not model evaluation were 5-fold cross validation and leave-one-out cross validation.

Recommendation 4a: Deep learning models represent optimal model architecture when working with complex histopathology image data. However, due to current lack of transparency with deep learning models and the computing resources required, traditional machine learning algorithms are currently better for use with histopathology image data, with SVM classifiers shown to produce equally strong performance metrics.

Recommendation 4b: 5-fold and leave-one out cross validation are common and reasonable methods for fine tuning ML model development for histopathology application, helping to optimize model parameters such as number of rounds of training, sample size, and computing resource allocation.

Algorithm Testing (Computer Vision Workflow Step 5)

Standard practice for ML model development is to reserve a portion of the available dataset as the "test dataset", used strictly for the purposes of final model performance assessment. In total, only 3 of 18 articles explicitly stated following this standard practice, with the size of the testing dataset ranging from 0.016%-50% of the total dataset [40,43], in line with the industry standard of reserving at least 50% of the data for training with room for adjustment if the dataset is too small[129]. Performance was reported as accuracy in all articles except for 3, with other reported metrics including F-1 score and area under the curve. As all these metrics are permutations of the true positives, true negatives, false positives, and false negatives and can be interconverted, they are all reasonable to report. However standardizing reporting guidelines would help with quick comparisons of results across studies.

Surprisingly, in 13 of the 18 articles reviewed, a separate testing dataset was not used to assess model performance, or it was unclear in the text whether or not the data was split. Instead, these articles (aside from the two where the data structure was unclear) reported using cross-validation to evaluate model performance. While cross-validation can provide a reasonable estimate, true model performance metrics are measured when the model is tested on a previously unseen dataset – allowing the determination of a model's generalizability. A likely cause for the adapted use of cross-validation for evaluation purposes was small image dataset sizes[128,129], however this is not typical of standard of ML modelling workflows due to problems related to data overfitting, and ultimately inflated model performance accuracy estimates.

Recommendation 5a: Ideal experimental design for ML model development for histopathology application includes reserving 20% of the original dataset for model performance testing.
Recommendation 5b: Accuracy and F-1 score are standard model performance metrics that should be routinely presented in articles on histopathology CAD-tool development. Standardization in reporting practices will allow for head-to-head comparisons of developed models and expedite translation of these tools into the clinical setting.

Conclusion:

The findings of the current systematic review highlight expanding and promising development of CAD-based tools that can be applied in the domain of clinical histopathology. To date, this work has been focused entirely in the realm of cancer pathologies, identifying the expansion of this technology into other branches of pathology as important opportunities for the future.

While CAD-based advances show tremendous promise for clinical utility, rigorous research protocols must be developed which can clearly demonstrate the value machine learning in a clinical setting in order to realize this clinical utility of these tools. Recommendations provided in this research article should provide orientation for researchers developing CAD tools and for clinical pathologists interested in adopting this technology into their practice, and should serve as a basis for the creation of evidence-based guidelines.

References:

1. Kamel HM. Trends and Challenges in Pathology Practice: Choices and necessities. *Sultan Qaboos Univ Med J*. 2011;11(1):38-44. https://www.ncbi.nlm.nih.gov/pubmed/21509206.
2. Rabe K, Snir OL, Bossuyt V, Harigopal M, Celli R, Reisenbichler ES. Interobserver variability in breast carcinoma grading results in prognostic stage differences. *Hum Pathol*. 2019. doi:https://doi.org/10.1016/j.humpath.2019.09.006
3. Wang S, Yang DM, Rong R, et al. Artificial Intelligence in Lung Cancer Pathology Image Analysis. *Cancers (Basel)*. 2019;11(11). doi:10.3390/cancers11111673
4. Jimenez-del-Toro O, Otálora S, Andersson M, et al. Chapter 10 - Analysis of Histopathology Images: From Traditional Machine Learning to Deep Learning. In: Depeursinge A, S. Al-Kadi O, Mitchell JRBT-BTA, eds. Academic Press; 2017:281-314. doi:https://doi.org/10.1016/B978-0-12-812133-7.00010-7
5. Hipp J, Flotte T, Monaco J, et al. Computer aided diagnostic tools aim to empower rather than replace pathologists: Lessons learned from computational chess. *J Pathol Inform*. 2011;2:25. doi:10.4103/2153-3539.82050
6. Bhargava R, Madabhushi A. Emerging Themes in Image Informatics and Molecular Analysis for Digital Pathology. *Annu Rev Biomed Eng*. 2016;18:387-412. doi:10.1146/annurev-bioeng-112415-114722
7. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. *Comput Struct Biotechnol J*. 2018;16:34-42. doi:https://doi.org/10.1016/j.csbj.2018.01.001

8. Huisman A, Looijen A, van den Brink SM, van Diest PJ. Creation of a fully digital pathology slide archive by high-volume tissue slide scanning. *Hum Pathol*. 2010;41(5):751-757. doi:10.1016/j.humpath.2009.08.026

9. Ilse M, Tomczak JM, Welling M. Chapter 22 - Deep multiple instance learning for digital histopathology. In: Zhou SK, Rueckert D, Fichtinger GBT-H of MIC and CAI, eds. Academic Press; 2020:521-546. doi:https://doi.org/10.1016/B978-0-12-816176-0.00027-2

10. Michie D. “memo” functions and machine learning. *Nature*. 1968;218(5136):19-22. doi:10.1038/218019a0

11. Langley P. Machine Learning as an Experimental Science.(Author abstract)(Editorial). *Mach Learn*. 1988;3(1):5. doi:10.1007/BF00115008

12. Baldi P. *Bioinformatics : The Machine Learning Approach*. (Brunak S, ed.). Cambridge, Mass.: Cambridge, Mass. : MIT Press, c1998.; 1998.

13. Ikeuchi K, ed. Computer Vision BT - Computer Vision: A Reference Guide. In: Boston, MA: Springer US; 2014:145. doi:10.1007/978-0-387-31439-6_100073

14. Klette author R. *Concise Computer Vision : An Introduction into Theory and Algorithms*. London : Springer, 2014.; 2014.

15. Montoya JC, Muñoz CQG, Márquez FPG. Chapter 9 - Remote condition monitoring for photovoltaic systems. In: Papaelias M, Márquez FPG, Karyotakis ABT-N-DT and CMT for REIA, eds. Boston: Butterworth-Heinemann; 2020:133-142. doi:https://doi.org/10.1016/B978-0-08-101094-5.00009-5

16. Kandemir M, Hamprecht FA. Computer-aided diagnosis from weak supervision: A benchmarking study. *Comput Med Imaging Graph*. 2015;42:44-50.

doi:https://doi.org/10.1016/j.compmedimag.2014.11.010

17. Ritter author GX, Ritter GX. *Handbook of Computer Vision Algorithms in Image Algebra*. Second edi. (Wilson JN, ed.). Boca Raton: Boca Raton : CRC Press, 2001.; 2001.

18. Khan author S. *A Guide to Convolutional Neural Networks for Computer Vision*. Vol # 15. (Rahmani author H, Shah author SAA, Bennamoun author M (Mohammed), eds.). San Rafael, California : Morgan & Claypool, 2018.; 2018.

19. Tsagkatakis G. Dimensionality Reduction and Sparse Representations in Computer Vision. Savakis A, Amuso V, Cahill N, Rhody H, eds. 2011.

20. Masnadi-Shirazi H, Mahadevan V, Vasconcelos N. On the design of robust classifiers for computer vision. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. ; 2010:779-786. doi:10.1109/CVPR.2010.5540136

21. Lopez-del Rio A, Nonell-Canals A, Vidal D, Perera-Lluna A. Evaluation of Cross-Validation Strategies in Sequence-Based Binding Prediction Using Deep Learning. *J Chem Inf Model*. 2019;59(4):1645-1657. doi:10.1021/acs.jcim.8b00663

22. Gutierrez author D. *Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R*. 1st editio. (Safari an OMC, ed.). Technics Publications, 2015.; 2015.

23. Belle A, Kon MA, Najarian K. Biomedical Informatics for Computer-Aided Decision Support Systems: A Survey. *Sci World J*. 2013;2013:1-8. doi:10.1155/2013/769639

24. Tacconelli E. Book: Systematic reviews: CRD's guidance for undertaking reviews in health care. *Lancet Infect Dis*. 2010;10(4):226.

25. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas*. 1960;20:37.

26. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009

27. McHugh ML. Interrater reliability: the kappa statistic. *Biochem medica*. 2012;22(3):276-282. https://www.ncbi.nlm.nih.gov/pubmed/23092060.

28. Hansel DE. The Gleason Grading System: The Approach that Changed Prostate Cancer Assessment. *J Urol*. 2017;197(2):S140-S141. doi:10.1016/j.juro.2016.11.028

29. Lin WC, Li CC, Epstein JI, Veltri RW. Curvelet-based texture classification of critical Gleason patterns of prostate histological images. *2016 IEEE 6th Int Conf Comput Adv Bio Med Sci*. 2016:1-6. doi:10.1109/ICCABS.2016.7802768

30. DiFranco MD, O'Hurley G, Kay EW, Watson RWG, Cunningham P. Ensemble based system for whole-slide prostate cancer probability mapping using color texture features. *Comput Med Imaging Graph*. 2011;35(7-8):629-645. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med7&NEWS=N&AN=21269807.

31. Orlov N V, Weeraratna AT, Hewitt SM, et al. Automatic detection of melanoma progression by histological analysis of secondary sites. *Cytometry A*. 2012;81(5):364-373. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med7&NEWS=N&AN=22467531.

32. Angel Arul Jothi J, Mary Anita Rajam V. Effective segmentation of orphan annie-eye nuclei from papillary thyroid carcinoma histopathology images using a probabilistic model and region-based active contour. *Biomed Signal Process Control*. 2016;30:149-161.

doi:10.1016/j.bspc.2016.06.008

33. F.A. S, L.S. O, C. P, et al. A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Trans Biomed Eng*. 2016;63(7):1455-1462. doi:10.1109/TBME.2015.2496264

34. Ninos K, Kostopoulos S, Kalatzis I, et al. Microscopy image analysis of p63 immunohistochemically stained laryngeal cancer lesions for predicting patient 5-year survival. *Eur Arch Oto-Rhino-Laryngology*. 2016;273(1):159-168. doi:10.1007/s00405-015-3747-x

35. K.-H. Y, C. Z, G.J. B, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun*. 2016;7:no pagination. http://www.nature.com/ncomms/index.html.

36. L. Y, W. C, P. M, et al. *High Throughput Analysis of Breast Cancer Specimens on the Grid*. Vol 10. L. Yang, Dept. of Electrical and Computer Eng., Rutgers Univ., Piscataway, NJ 08544, USA.; 2007:617-625. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed11&NEWS=N&AN=350325804.

37. Fukuma K, Prasath VBS, Kawanaka H, Aronow BJ, Takase H. A Study on Nuclei Segmentation, Feature Extraction and Disease Stage Classification for Human Brain Histopathological Images. In: *Procedia Computer Science*. Vol 96. ; 2016. doi:10.1016/j.procs.2016.08.164

38. Bayramoglu N, Kannala J, Heikkilä J. Deep learning for magnification independent breast cancer histopathology image classification. *2016 23rd Int Conf Pattern Recognit*.

2016:2440-2445. doi:10.1109/ICPR.2016.7900002

39. Bentaieb A, Nosrati MS, Li-Chang H, Huntsman D, Hamarneh G. Clinically-inspired automatic classification of ovarian carcinoma subtypes. *J Pathol Inform*. 2016;7(1). doi:10.4103/2153-3539.186899

40. Waheed S, Moffitt RA, Chaudry Q, Young AN, Wang MD. Computer Aided Histopathological Classification of Cancer Subtypes. *2007 IEEE 7th Int Symp Bioinforma Bioeng*. 2007:503-508. doi:10.1109/BIBE.2007.4375608

41. Doyle S, Rodriguez C, Madabhushi A, Tomaszeweski J, Feldman M. Detecting Prostatic Adenocarcinoma From Digitized Histology Using a Multi-Scale Hierarchical Classification Approach. *2006 Int Conf IEEE Eng Med Biol Soc*. 2006:4759-4762. doi:10.1109/IEMBS.2006.260188

42. Mete M, Xu X, Fan C y., Shafirstein G. Head and Neck Cancer Detection in Histopathological Slides. *Sixth IEEE Int Conf Data Min - Work*. 2006:223-230. doi:10.1109/ICDMW.2006.90

43. Y. X, J.-Y. Z, E.I.C. C, et al. Weakly supervised histopathology cancer image segmentation and classification. *Med Image Anal*. 2014;18(3):591-604. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=24637156.

44. Budinská E, Bosman F, Popovici V. Experiments in molecular subtype recognition based on histopathology images. *2016 IEEE 13th Int Symp Biomed Imaging*. 2016:1168-1172. doi:10.1109/ISBI.2016.7493474

45. K. N, S. K, K. S, et al. Computer-based image analysis system designed to differentiate

between low-grade and high-grade laryngeal cancer cases. *Anal Quant Cytopathol Histopathol*. 2013;35(5):261-272. http://www.aqch.com/toc/auto_article_process.php?year=2013&page=261&id=23248&sn=0.

46. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45(10):1113-1120. doi:10.1038/ng.2764

47. Fukuma K, Surya Prasath VB, Kawanaka H, Aronow BJ, Takase H. A study on feature extraction and disease stage classification for glioma pathology images. In: *2016 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2016*. ; 2016. doi:10.1109/FUZZ-IEEE.2016.7737958

48. W. W, J.A. O, Wang W, Ozolek JA, Rohde GK. *Detection and Classification of Thyroid Follicular Lesions Based on Nuclear Structure from Histopathology Images*. Vol 77. G. K. Rohde, Department of Biomedical Engineering, HHC 122, 5000 Forbes Avenue, Pittsburgh, PA 15232, United States. E-mail: gustavor@cmu.edu: Wiley-Liss Inc. (111 River Street, Hoboken NJ 07030-5774, United States); 2010:485-494. doi:10.1002/cyto.a.20853

49. Angel Arul Jothi J, Mary Anita Rajam V. Effective segmentation and classification of thyroid histopathology images. *Appl Soft Comput J*. 2015. doi:10.1016/j.asoc.2016.02.030

50. Xu Y, Zhu J-Y, Chang E, Tu Z. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. ; 2012. doi:10.1109/CVPR.2012.6247772

51. Slaoui M, Fiette L. Histopathology Procedures: From Tissue Sampling to Histopathological Evaluation. *Methods Mol Biol*. 2011;691:69-82. doi:10.1007/978-1-60761-849-2_4

52. Ameisen D, Deroulers C, Perrier V, et al. Towards better digital pathology workflows: programming libraries for high-speed sharpness assessment of Whole Slide Images. *Diagn Pathol*. 2014;9(1):S3. doi:10.1186/1746-1596-9-S1-S3

53. Aggarwal 1st Lt. Pushkar. Data augmentation in dermatology image recognition using machine learning. *Ski Res Technol*. 2019;25(6):815-820. doi:10.1111/srt.12726

54. Bang J, Hur T, Kim D, et al. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. *Sensors*. 2018;18(11):3744. doi:10.3390/s18113744

55. Solomon C. *Fundamentals of Digital Image Processing : A Practical Approach with Examples in Matlab*. (Breckon T, ed.). Chichester: Chichester : Wiley-Blackwell, ©2011.; 2011.

56. CIE Recommendations on Uniform Color Spaces, Color-Difference Equations, and Metric Color Terms. *Color Res Appl*. 1977;2(1):5-6. doi:10.1002/j.1520-6378.1977.tb00102.x

57. Lyon HO, De Leenheer AP, Horobin RW, et al. Standardization of reagents and methods used in cytological and histological practice with emphasis on dyes, stains and chromogenic reagents. *Histochem J*. 1994;26(7):533-544. doi:10.1007/BF00158587

58. Khan AM, Rajpoot N, Treanor D, Magee D. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Trans Biomed Eng*. 2014;61(6). doi:10.1109/TBME.2014.2303294

59. Roy S, kumar Jain A, Lal S, Kini J. A study about color normalization methods for

histopathology images. *Micron*. 2018;114:42-61. doi:https://doi.org/10.1016/j.micron.2018.07.005

60. Jayalakshmi T, Santhakumaran A. Statistical Normalization and Back Propagationfor Classification. *Int J Comput Theory Eng*. 2011;3(1):89-93. doi:10.7763/IJCTE.2011.V3.288

61. He BS, Zhu F, Shi YG. Medical Image Segmentation. *Adv Mater Res*. 2013;760-762:1590-1593. doi:10.4028/www.scientific.net/AMR.760-762.1590

62. Carpenter AE, Jones TR, Lamprecht MR, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol*. 2006;7(10):R100. doi:10.1186/gb-2006-7-10-r100

63. Blake 1956- A, Kohli P, Rother C. *Markov Random Fields for Vision and Image Processing*. Cambridge, Mass.: Cambridge, Mass. : MIT Press, ©2011.; 2011.

64. Sakaridis C, Drakopoulos K, Maragos P. Theoretical Analysis of Active Contours on Graphs. *SIAM J Imaging Sci*. 2017;10(3):1475-1510. doi:10.1137/16M1100101

65. Peizhuang W. Pattern Recognition with Fuzzy Objective Function Algorithms (James C. Bezdek). *SIAM Rev*. 1983;25(3):442. doi:10.1137/1025116

66. Poli R, Kennedy J, Blackwell T. Particle swarm optimization. *Swarm Intell*. 2007;1(1):33-57. doi:10.1007/s11721-007-0002-0

67. 13. Density-Based Clustering Algorithms. In: *Data Clustering: Theory, Algorithms, and Applications*. ASA-SIAM Series on Statistics and Applied Mathematics. Society for Industrial and Applied Mathematics; 2007:219-226. doi:doi:10.1137/1.9780898718348.ch13

68. Broeke author J. *Image Processing with ImageJ - Second Edition*. 2nd editio. (Pérez

author J, Pascau author J, Safari an OMC, eds.). Packt Publishing, 2015.; 2015.

69. Huang D, Shan C, Ardabilian M, Wang Y, Chen L. Local binary patterns and its application to facial image analysis: A survey. *IEEE Trans Syst Man Cybern Part C Appl Rev*. 2011;41(6):765-781. doi:10.1109/TSMCC.2011.2118750

70. Nanni L, Brahnam S, Lumini A. Local phase quantization descriptor for improving shape retrieval/classification. *Pattern Recognit Lett*. 2012;33(16):2254-2260. doi:10.1016/j.patrec.2012.07.007

71. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Trans Syst man Cybern*. 1973;smc-3(No. 6):610-621. doi:10.1109/TSMC.1973.4309314

72. Pathak RS. *The Wavelet Transform*. Vol v.4. Amsterdam : Amsterdam ; 2009.

73. Schreier 1961- R. *Understanding Delta-Sigma Data Converters*. (Temes 1929- GC, John Wiley & Sons publisher, eds.). Piscataway, NJ : Hoboken, N.J. : Piscataway, NJ : IEEE Press ; 2005.

74. Ünay D, Stanciu SG. An Evaluation on the Robustness of Five Popular Keypoint Descriptors to Image Modifications Specific to Laser Scanning Microscopy. *IEEE Access*. 2018;6:40154-40164. doi:10.1109/ACCESS.2018.2855264

75. A V, Hebbar D, Shekhar VS, Murthy KNB, Natarajan S. Two Novel Detector-Descriptor Based Approaches for Face Recognition Using SIFT and SURF. *Procedia Comput Sci*. 2015;70(C):185-197. doi:10.1016/j.procs.2015.10.070

76. Lowe D. Distinctive Image Features from Scale-Invariant Keypoints. *Int J Comput Vis*. 2004;60(2):91-110. doi:10.1023/B:VISI.0000029664.99615.94

77. Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-Up Robust Features (SURF). *Comput Vis*

Image Underst. 2008;110(3):346-359. doi:10.1016/j.cviu.2007.09.014

78. Rublee E, Rabaud V, Konolige K, Bradski G. ORB: An efficient alternative to SIFT or SURF. In: *Proceedings of the IEEE International Conference on Computer Vision*. ; 2011:2564-2571. doi:10.1109/ICCV.2011.6126544

79. Mair E, Hager GD, Burschka D, Suppa M, Hirzinger G. Adaptive and Generic Corner Detection Based on the Accelerated Segment Test BT - Computer Vision – ECCV 2010. In: Daniilidis K, Maragos P, Paragios N, eds. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010:183-196.

80. Calonder M, Lepetit V, Ozuysal M, Trzcinski T, Strecha C, Fua P. BRIEF: Computing a Local Binary Descriptor Very Fast. *IEEE Trans Pattern Anal Mach Intell*. 2012;34(7):1281-1298. doi:10.1109/TPAMI.2011.222

81. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: a review. *IEEE Rev Biomed Eng*. 2009;2:147-171. doi:10.1109/RBME.2009.2034865

82. Sarangi PK, Ravulakollu KK. Feature Extraction and Dimensionality Reduction in Pattern Recognition Using Handwriten Odia Numerals. 2014;65(3).

83. Michalewicz Z. *How to Solve It : Modern Heuristics*. 2nd ed., r. (Fogel DB, ed.). Berlin : Berlin ; 2004.

84. Pearson K. Mathematical Contributions to the Theory of Evolution. X. Supplement to a Memoir on Skew Variation. *Philos Trans R Soc London Ser A, Contain Pap a Math or Phys Character*. 1901;197(287-299):443-459. doi:10.1098/rsta.1901.0023

85. Papailiopoulos DS, Dimakis AG. Interference Alignment as a Rank Constrained Rank

Minimization. *IEEE Trans Signal Process*. 2012;60(8):4278-4288. doi:10.1109/TSP.2012.2197393

86. Fisher RA. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Ann Hum Genet*. 1936;7(2):179-188. doi:10.1111/j.1469-1809.1936.tb02137.x

87. Cacciatore S, Luchinat C, Tenori L. Knowledge discovery by accuracy maximization. *Proc Natl Acad Sci U S A*. 2014;111(14):5117. doi:10.1073/pnas.1220873111

88. Duda RO. *Pattern Classification*. 2nd ed. (Hart PE, Stork DG, eds.). New York: New York : Wiley, 2001.; 2001.

89. Litchfield JT, Wilcoxon F. The Rank Correlation Method. *Anal Chem*. 1955;27(2):299-300. doi:10.1021/ac60098a038

90. Ziarko W. Variable precision rough set model. *J Comput Syst Sci*. 1993;46(1):39-59. doi:10.1016/0022-0000(93)90048-2

91. Iman RL, Davenport JM. Rank correlation plots for use with correlated input variables. *Commun Stat - Simul Comput*. 1982;11(3):335-360. doi:10.1080/03610918208812266

92. Hess AS, Hess JR. Principal component analysis. *Transfusion*. 2018;58(7):1580-1582. doi:10.1111/trf.14639

93. Kupper LL, Hafner KB. On assessing interrater agreement for multiple attribute responses. *Biometrics*. 1989;45(3):957. doi:10.2307/2531695

94. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Roweis ST, ed. *Science*. 2000;290(5500):2323-2326. doi:10.1126/science.290.5500.2323

95. Wang C, Chen J, Sun Y, Shen X. Wireless Sensor Networks Localization with Isomap. In: *2009 IEEE International Conference on Communications*. ; 2009:1-5.

doi:10.1109/ICC.2009.5199576

96. Teng L, Li H, Fu X, Chen W, Shen I-F. Dimension reduction of microarray data based on local tangent space alignment. In: *Fourth IEEE Conference on Cognitive Informatics 2005, ICCI 2005*. ; 2005:154-159. doi:10.1109/COGINF.2005.1532627

97. Anderson JA. Point biserial correlation. *Stata Tech Bull*. 1994;3(17).

98. Kent JT. Information gain and a general measure of correlation. *Biometrika*. 1983;70(1):163-173. doi:10.1093/biomet/70.1.163

99. Angel Arul Jothi J, Mary Anita Rajam V. *Segmentation of Nuclei from Breast Histopathology Images Using PSO-Based Otsu's Multilevel Thresholding*. Vol 325.; 2015. doi:10.1007/978-81-322-2135-7_88

100. Inuiguchi M, Hirano S, Tsumoto 1963- S. *Rough Set Theory and Granular Computing*. Vol v. 125. Berlin : Berlin ; 2003.

101. Guenther N, Schonlau M. Support Vector Machines. *Stata J*. 2016;16(4):917-937. doi:10.1177/1536867X1601600407

102. Darrell T, Indyk P, Shakhnarovich G. *Nearest-Neighbor Methods in Learning and Vision : Theory and Practice*. Cambridge, Mass.: Cambridge, Mass. : MIT Press, 2005; 2005.

103. Despontin M. Decision theory: An introduction to the mathematics of rationality: Ellis Horwood, Chichester, 1986, 448 + pages, £55.00. *Eur J Oper Res*. 1987;29(2):212-213. doi:10.1016/0377-2217(87)90113-5

104. Pavlov YL. *Random Forests*. Utrecht : Utrecht ; 2000.

105. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. *Mach Learn*. 1997;29(2):131-163. doi:10.1023/A:1007465528199

106. Mayr, Binder, Gefeller, Schmid. The evolution of boosting algorithms. From machine learning to statistical modelling. *Methods Inf Med*. 2014;53(6):419-41927. doi:10.3414/ME13-01-0122

107. De Wilde 1958- P. *Neural Network Models : Theory and Projects*. Second edi. London ; 1997.

108. Herrera F, Ventura S, Bello R, et al. *Multiple Instance Learning : Foundations and Algorithms*. Cham: Cham : Springer, 2016.; 2016.

109. Lin J-Y, Ke H-R, Chien B-C, Yang W-P. Classifier design with feature selection and feature extraction using layered genetic programming. *Expert Syst Appl*. 2008;34(2):1384-1393. doi:10.1016/j.eswa.2007.01.006

110. Fushiki T. Estimation of prediction error by using K -fold cross-validation. *Stat Comput*. 2011;21(2):137-146. doi:10.1007/s11222-009-9153-8

111. Wong T-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit*. 2015;48(9):2839-2846. doi:https://doi.org/10.1016/j.patcog.2015.03.009

112. Ruan F, Qi J, Yan C, Tang H, Zhang T, Li H. Quantitative detection of harmful elements in alloy steel by LIBS technique and sequential backward selection-random forest (SBS-RF). *J Anal At Spectrom*. 2017;32(11):2194-2199. doi:10.1039/C7JA00231A

113. Efron B. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Vol 38. (Mathematics S for I and A, ed.). Philadelphia, Pa.: Philadelphia, Pa. : Society for Industrial and Applied Mathematics SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104, 1982.; 1982.

114. Cui Z, Gong G. The effect of machine learning regression algorithms and sample size on

individualized behavioral prediction with functional connectivity features. *Neuroimage*. 2018;178:622-637. doi:10.1016/j.neuroimage.2018.06.001

115. Vapnik V. *The Nature of Statistical Learning Theory*. Springer science & business media; 2013.

116. Byrd RH, Chin GM, Nocedal J, Wu Y. Sample size selection in optimization methods for machine learning. *Math Program*. 2012;134(1):127-155.

117. Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a classifier for microarray data? *Clin Cancer Res*. 2008;14(1):108-114. doi:10.1158/1078-0432.CCR-07-0443

118. Mukherjee S, Tamayo P, Rogers S, et al. Estimating dataset size requirements for classifying DNA microarray data. *J Comput Biol*. 2003;10(2):119-142. doi:10.1089/106652703321825928

119. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak*. 2012;12(1):8. doi:10.1186/1472-6947-12-8

120. Tran T, Pham T, Carneiro G, Palmer L, Reid I. A Bayesian Data Augmentation Approach for Learning Deep Models. In: Guyon I, Luxburg U V, Bengio S, et al., eds. *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.; 2017:2797-2806. http://papers.nips.cc/paper/6872-a-bayesian-data-augmentation-approach-for-learning-deep-models.pdf.

121. Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10,000 classes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. ;

2014:1891-1898.

122. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*. Vol 2. ; 2012:1097-1105. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84876231242&partnerID=40&md5=621b2cd1757ccc77341281f8a2f2ecaf.

123. Farabet C, Couprie C, Najman L, LeCun Y. Learning hierarchical features for scene labeling. *IEEE Trans Pattern Anal Mach Intell*. 2012;35(8):1915-1929.

124. Burlina P, Billings S, Joshi N, Albayda J, Miller FW. Automated diagnosis of myositis from muscle ultrasound: Exploring the use of machine learning and deep learning methods. *PLoS One*. 2017;12(8). doi:10.1371/journal.pone.0184059

125. Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2012:3642-3649.

126. Choo J, Liu S. Visual Analytics for Explainable Deep Learning. *IEEE Comput Graph Appl*. 2018;38(4):84-92. doi:10.1109/MCG.2018.042731661

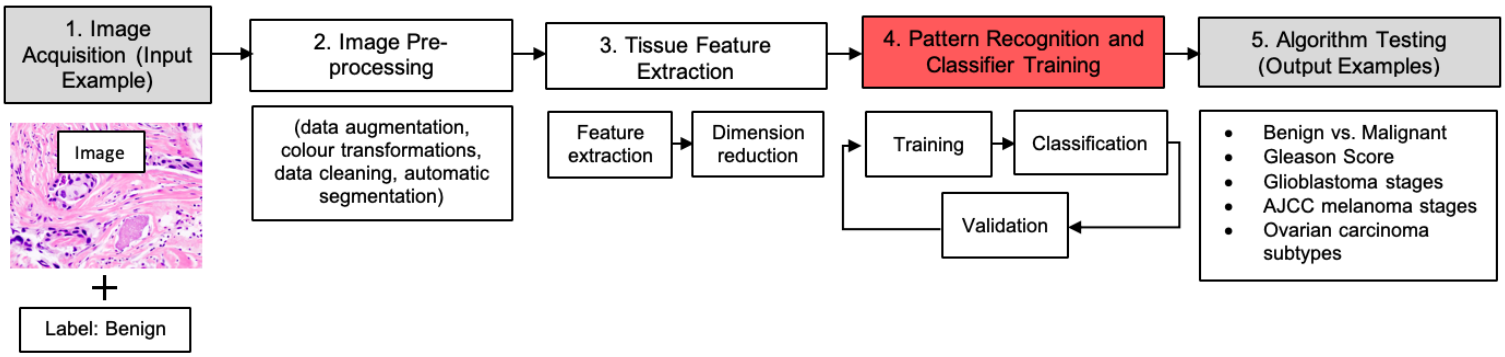
127. Došilović FK, Brčić M, Hlupić N. Explainable artificial intelligence: A survey. In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. ; 2018:210-215. doi:10.23919/MIPRO.2018.8400040

128. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7(1):91.

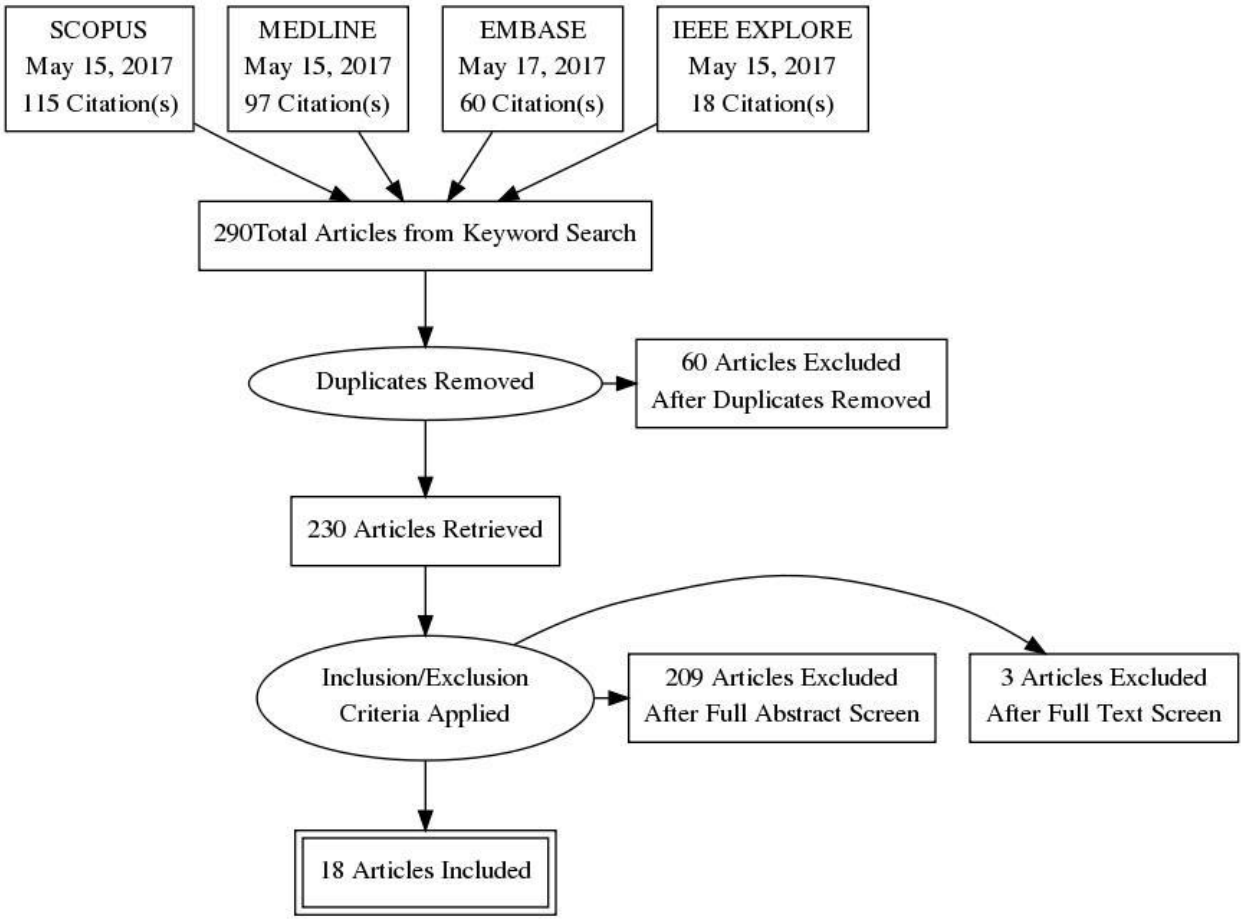
129. Chicco D. Ten quick tips for machine learning in computational biology. *BioData Min*. 2017;10(1):35. doi:10.1186/s13040-017-0155-3

Figures

1.


1. Image Acquisition (Input Example)
2. Image Pre-processing
3. Tissue Feature Extraction
4. Pattern Recognition and Classifier Training
5. Algorithm Testing (Output Examples)
Image
+
Label: Benign
(data augmentation, colour transformations, data cleaning, automatic segmentation)
Feature extraction
Dimension reduction
Training
Classification
Validation
Benign vs. Malignant
Gleason Score
Glioblastoma stages
AJCC melanoma stages
Ovarian carcinoma subtypes

2.


SCOPUS
May 15, 2017
115 Citation(s)
MEDLINE
May 15, 2017
97 Citation(s)
EMBASE
May 17, 2017
60 Citation(s)
IEEE EXPLORE
May 15, 2017
18 Citation(s)
290Total Articles from Keyword Search
Duplicates Removed
60 Articles Excluded
After Duplicates Removed
230 Articles Retrieved
Inclusion/Exclusion
Criteria Applied
209 Articles Excluded
After Full Abstract Screen
3 Articles Excluded
After Full Text Screen
18 Articles Included

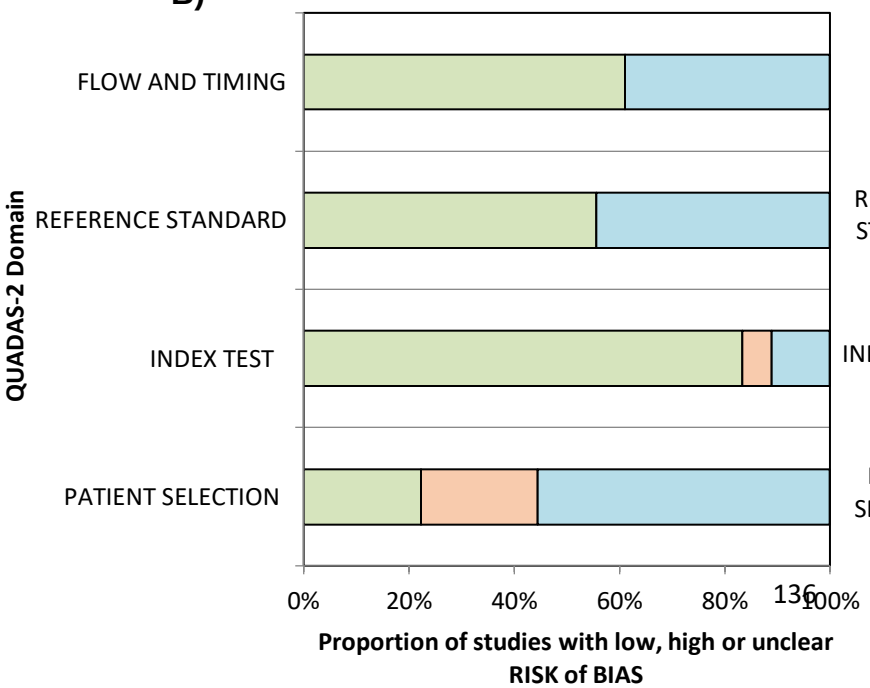
3.

A)

Study	RISK OF BIAS				APPLICABILITY CONCERNS		
	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD
Spanhol et. al.	○	○	?	○	○	○	○
Fukuma et. al.	?	○	?	?	○	○	○
Orlov et. al.	?	○	?	?	○	○	○
BenTaieb et. al.	○	○	○	○	○	○	○
Waheed et. al.	○	?	?	?	○	?	○
Ninos et. al.	?	○	○	○	○	○	○
Lin et. al.	?	○	○	?	○	○	○
Bayramoglu et. al.	?	○	?	?	○	○	?
Doyle et. al.	?	○	?	?	○	○	?
Wang et. al.	○	○	○	○	○	○	○
Jothi et. al.	○	○	○	○	○	○	○
DiFranco et. al.	?	○	○	○	○	○	○
Budinska et. al.	○	○	?	?	○	○	○
Mete et. al.	○	○	?	○	○	○	○
Yang et. al.	?	○	○	○	○	○	○
Ninos et. al.	?	?	○	○	○	○	○
Xu et. al.	○	○	○	○	○	○	○
Yu et. al.	?	○	○	○	○	○	○

○ Low Risk ○ High Risk ? Unclear Risk

B)



C)

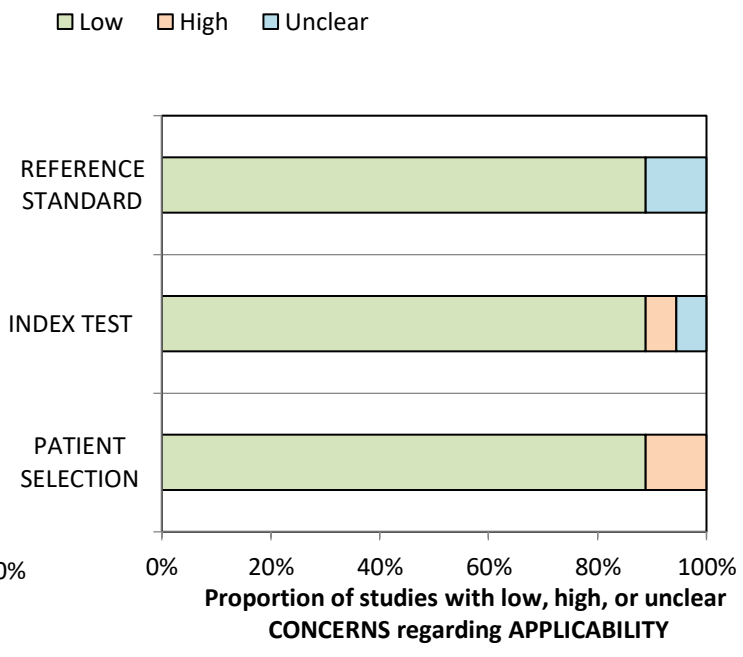


Figure Legends

Figure 1. 5-Step computer vision workflow for CAD development on digital histopathology images. Upper panels show the workflow elements common to each reviewed article and beneath each panel is an example of the inputs (step 1), outputs (step 5), or tasks (steps 2-4) involved at each step.

Figure 2. Flow chart of article selection process. Medline, Embase, Scopus, and IEEE Explore were searched for articles relating to CAD on histopathology images. After abstract and full text screening, 18 articles remained for review.

Figure 3. Risk of bias summary. Risk of bias and applicability were assessed along 7 criteria outlined in the QUADAS-II tool. Low risk items are marked in green, high risk in red, and unclear in blue. **A.** Summary of risk score for each article along 4 Risk of Bias dimensions and 3 Applicability Concerns dimensions. **B.** Proportion of risk ratings along 4 Risk of Bias dimensions (Patient Selection, Index Test, Reference Standard, Flow and Timing) across all articles. **C.** Proportion of risk ratings along 3 Applicability Concerns dimensions (Patient Selection, Index Test, Reference Standard) across all articles.

Supplementary Tables

Supplementary Table 1. MeSH terms and word variants searched in Medline

1. Visual pattern mining in histology image collections using bag of features.m_titl.
2. histological techniques/ or histocytochemistry/ or immunohistochemistry/
3. (histopathology adj2 imag*).tw,kf.
4. exp Diagnostic Imaging/
5. 2 and 4
6. histology ajd2 imag*.tw,kf.
7. 3 or 5
8. exp pattern recognition, automated/ or "neural networks (computer)"/ or support vector machine/
9. (pattern adj2 (recognition? or classification?)).tw,kf.
10. image segmentation.tw,kf.
11. 8 or 9 or 10
12. 7 and 11
13. limit 12 to english language

Supplementary Table 2. MeSH terms and word variants searched in Embase

1. Visual pattern mining in histology image collections using bag of features.m_titl.
2. histology/ or immunohistology/ or immunohistochemistry/
3. (histopathology adj2 imag*).tw.
4. exp diagnostic imaging/
5. 2 and 4
6. 3 or 5
7. exp automated pattern recognition/ or exp artificial neural network/ or exp support vector machine/

8. (pattern adj2 (recognition? or classification?)).tw.
9. image segmentation.tw.
10. 7 or 8 or 9
11. 6 and 10

Supplementary Table 3. MeSH terms and word variants searched in Scopus

1. TITLE-ABS-KEY ((histopathology W/2 imag*) AND ((pattern W/2 (recognition? OR classification?)) OR image AND segmentation))

Supplementary Table 4. MeSH terms and word variants searched in IEEE Explore

1. histopathology image* AND (pattern recognition* or pattern classification*) in Basic Search