Chapter 1

Introduction

Artificial intelligence (AI) is becoming increasingly powerful, reliable, and capable of outperforming humans in a wide range of tasks. Nowadays, AI permeates many aspects of human life and assists people in their daily lives, such as virtual assistants, self-driving cars, drones, email, smart devices, web searches, etc. AI is being developed at an increasing rate every day, becoming more intelligent and being applied in various fields, with the goal of improving human life. With recent advancements in the AI and Internet of Things (IoT), huge volumes of data are being generated. Differing in form, data could be speech, text, image, etc., with image data contributing a significant share of global data creation. The field of computer vision includes a set of main problems such as image classification, localization, image segmentation, and object detection. Among these, image classification is one of the core problems in computer vision. Despite its simplicity, it has many practical applications and forms the basis for other computer vision problems. With the advent of machine learning and deep learning, combined with powerful hardware and GPUs, outstanding performance on image classification tasks is now possible. Hence, AI brought great success in the entire field of image recognition, object detection, and image classification algorithms achieving above human-level performance and real-time object detection and classifi-

cation. In the healthcare system, AI is applied for managing and recording massive amounts of medical data, data analysis, and helping doctors to make decisions. Digital histopathology has made computer-aided diagnosis (CAD) the most prominent development in the field of healthcare. The machine outperforms humans when it comes to performing repetitive tasks quickly and consistently. The machine can also pay more attention even to the smallest detail, such as a single pixel in an image. As an emerging discipline, AI-based CAD systems have recently shown considerable potential in histopathological image classification. Figure 1.1 shows different steps in manual and digital pathology workflow.

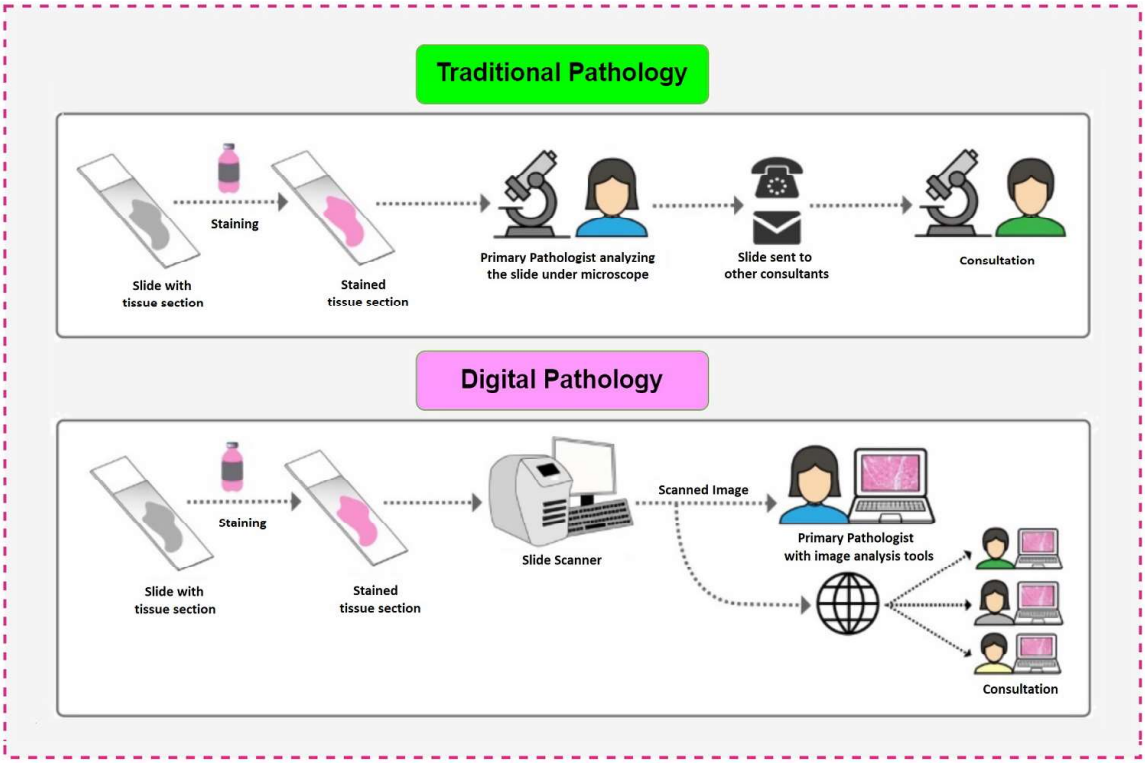

Figure 1.1: Comparison between the workflow of traditional and digital pathology.

The developments made in the past decade in the field of AI have revolutionized the field of cancer detection and treatment. Cancer is a longstanding critical disease threatening human and animal health. In 2018 alone, 18.1 million new cases of cancer and 9.6 million cancer-related deaths were reported in humans [1]. Recent global cancer

statistics have shown that breast cancer is still the most common type of cancer and the leading cause of mortality among women, accounting for 24.2% (2.1 million) new cases and 626,679 deaths per year [1]. Among animal species, pet animals, especially dogs, are more prone to cancer, often leading to poor prognosis and high mortality [2]. In unspayed female dogs, canine mammary tumour (CMT) is the most common malignancy with thrice higher mortality rates as compared to human breast cancer (HBC) [3]. CMTs are considered excellent models for HBC studies as they share many similarities with HBC [4]. Owing to the influence of several hormonal, genetic and other associated factors, CMTs present diverse histological subtypes. Thus, the correct interpretation of CMTs is a major challenge for clinicians. Diagnosis of CMTs by routine cytology of biopsy or by the extirpated gland is complex and requires interpretation by trained veterinary pathologists. In humans, due to increased awareness about the disease, early diagnosis is possible with the help of routine self-check-ups and mammography followed by a biopsy. However, it is difficult to detect cancer at an early stage in pets because they are unable to convey warning signs and symptoms. Therefore, the diagnosis is made only when the tumour becomes visibly apparent to the animal owner. Thus, accurate diagnosis and differentiation between benign and malignant neoplasms are crucial for the successful outcome of treatment modalities, especially in canines.

Both in humans and canines, histopathological analysis remains the gold standard for cancer diagnosis. However, diagnosis using hematoxylin and eosin (H&E) stained biopsies is very time-consuming, costly, and laborious, requiring the intense efforts of specialized pathologists. Furthermore, diagnosis based upon manual analysis of slides suffers from inter-observer variability, with approximately 75% diagnostic concordance between specialists [5]. Hence, computer-aided approaches could be included in digital pathology in order to achieve rapid and reproducible results. They help in improving classification accuracy and reducing variability in interpretations, as errors made by machine learning methods have been reported to be less than those made by a single

pathologist [6]. These techniques are also helpful for assisting pathologists and reducing their labour in localizing and identifying abnormalities in the cancer tissue images. Therefore, researchers are trying to exploit the morphological criteria in the usual classification approach to develop CAD systems for improving the diagnostic efficacy and increasing the level of inter-observer agreement [7]. Recently, Convolutional Neural Networks (CNN) based on deep learning architecture have been reported to be a powerful tool in the automated classification of human cancer histopathology images [8, 9]. CNN, along with multiple-instance learning, have accomplished good performance in the binary classification of human cancers and have evolved as a method of choice for analyzing histopathological images [10]. Despite the noteworthy performance of these systems for the binary classification of cancer, colour variations in histopathology images are a concern for automated analysis. For a pathologist, these colour variations may not hinder the analysis, but these variations can significantly affect image interpretation in automated image analysis. Hence, developing a framework with effective stain normalization for histopathological images is also an area to work upon [11, 12]. While results from preliminary studies in the field of HBC image classification are promising, this is a rapidly evolving field with new knowledge emerging in both cancer biology and deep learning, providing opportunities to improve upon existing approaches.

Undoubtedly CNNs have achieved remarkable success in the field of computer vision; however, a major limiting factor for deep learning-based approaches is the requirement for huge amounts of labelled data. In the field of medical imaging, such large labelled datasets are difficult to acquire. Machine learning-based algorithms, in particular, are a potential alternative to deep learning for medical imaging with smaller datasets. A major constraint of conventional machine learning techniques is the requirement of complex processing for the extraction of discriminatory features. Efficient feature engineering simplifies the model by reducing training and execution times, input requirements, and computational costs. Furthermore, it not only improves the model compactness and

transparency by removing insignificant features from the dataset but also facilitates fast interpretation. Therefore, valuable features extraction presents a critical challenge for automated cancer histopathology. Though various feature extraction modalities have been explored for cancer histopathological image analysis, to date, none have proved entirely convincing. Thus, designing an efficient feature extractor for such complex tissue prediction is still an open and challenging task.

On the other hand, the advancements in IoT and cloud services have made smart e-healthcare services available in a remote and distributed environment. However, this has raised significant privacy and efficiency concerns that must be addressed. Privacy is a major challenge when sharing clinical data across the cloud, which frequently contains sensitive patient-related information. Adequate patient privacy protection contributes to increased public trust in medical research. Besides that, deep learning-based models are complex, and efficient data processing in such models is complicated in a cloud-based approach. Although deep learning in healthcare acts as a fast diagnostic tool to support doctors in making critical decisions, it must also be secured against adversary attacks. Developing a secure and private learning model for histopathology image classification remains less explored and remains an open research problem.

Moreover, in today's era, edge computing is emerging as a widespread technique in the healthcare sector [13], and research communities are inclined towards developing models that can be easily embedded in edge devices without compromising performance. However, deep learning-based approaches are computationally expensive and have huge parameters, making them less affordable for edge devices. In order to make them affordable for edge devices, the whole classification model needs to be compressed while maintaining accuracy. Providing a low-cost solution for histopathological diagnosis in the recent edge-computing world is of utmost importance. The integration of AI, edge computing, and IoTs have the potential to revolutionize the healthcare system, particularly in the field of cancer theranostics. Thus, developing a robust histopatho-

logical image classification model for resource-constrained devices with high accuracy is a challenging and priority research area to work upon.

Considering the importance of correct diagnosis in patient management, considerable efforts have been made in the past for developing robust, precise, and automated CAD systems for humans. However, despite higher incidences and mortality rates in dogs, to date, no efforts have been made to automate the diagnosis of CMTs to relieve the burden on veterinary oncologists to focus more on the cases that are difficult to diagnose. This may be due to the lack of any publicly available dog mammary tumour image database for automated analysis. Thus, the primary focus of this thesis is to solve the classification problem for CMTs by developing a framework for automated histopathological image classification. The focus is also to improve the classification

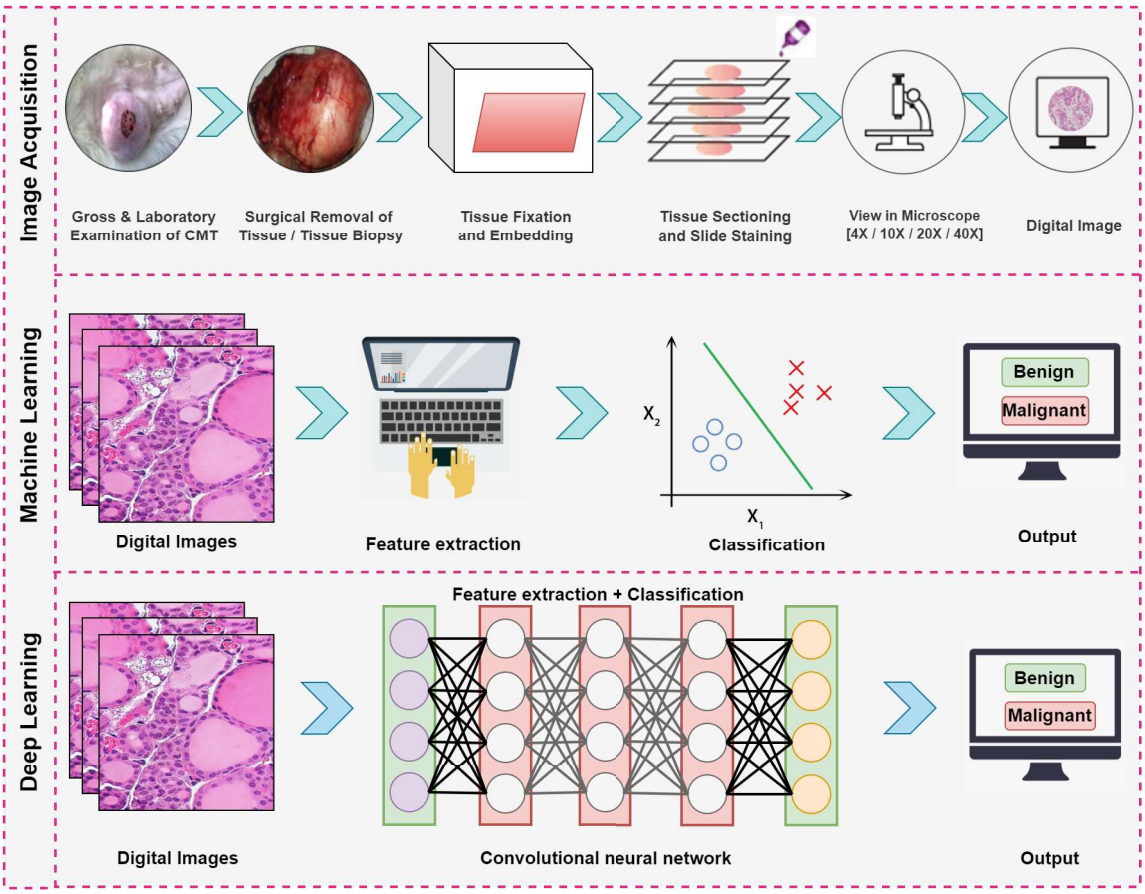

Figure 1.2: Generalize overview of the histopathological image classification.

performance by developing a framework with high accuracy for HBC classification as compared to other state-of-the-art classifiers. Since deep learning-based approaches are data expensive, the present study also attempts to develop a machine learning-based algorithm for feature extraction from histopathological images utilizing the novel concepts of center of mass and fuzzy modelling to deal with the lack of large labelled data samples and uncertainty in histopathology images. The study also attempts to develop a secure model to deal with the issues of data privacy and security. Further, the aim is also to develop a lightweight framework for mobile edge computing and other resource-constrained devices by utilizing innovative approaches. The general overview of the histopathological image classification is presented in Figure 1.2.

The rest of the chapter is organized as follows. The next section presents the motivation of the thesis, followed by the challenges and main objectives. Section 1.3 presents the contribution of this thesis, and Section 1.4 provides the thesis organization details.

1.1 Motivation

Recent global cancer statistics have shown that breast cancer is still the most common type of cancer and the leading cause of mortality among women. In the case of pet animals, mammary tumours have the highest incidences (16.8%) in females. Considering the importance of correct diagnosis in patient management, in humans, considerable efforts have been made in the past for developing robust, precise, and automated CAD systems for HBC; however, no studies have focused so far on automated classification of CMTs. There is currently a lack of any existing digital database for CMTs. Heterogeneous and diverse types of CMTs and a paucity of skilled veterinary pathologists justify the need for automated diagnosis to reduce the burden on trained veterinary pathologists. The demand for pathologists is growing due to the increasing interest in personalized cancer therapy, which demands accurate and immediate histopathological

assessments for making therapeutic decisions. However, the shortage of trained and experienced cancer histopathologists and the time involved in the manual analysis of slides lead to delays in diagnosis, which, in turn, delays these therapeutic decisions. Advantages for automating cancer histopathology tasks are manifold. Firstly, it increases diagnostic accuracy and quality by reducing inter-observer and intra-observer variability. Secondly, it speeds up the analysis process and saves pathologists time and labour, allowing them to focus on more difficult-to-diagnose cases. Enabling a software solution to perform some of the work of a pathologist can reduce the amount of pathological work, reduce costs and provide a timely diagnosis.

The recent success of most of the world's popular services is based on big data and AI. Smartphones and the IoT have accelerated the integration of technology and healthcare, allowing for a plethora of innovation that improves people's lives. Healthcare is now mobile and no longer confined to the waiting room. It inspires us to adopt AI techniques for tumour diagnosis. Histopathological data are quite complex to analyze due to multiple nuclei in a single cell, staining and different magnifications. Therefore, enhancing the prediction ability of the learning model based on limited labelled data is another challenge that boosts researchers' interest in innovating a new strategy to deal with it. Nowadays, deep learning is widely accepted due to its exceptional feature representation capability and thus gives a contemporary edge to the healthcare industry. However, new variants of deep neural networks are being developed, but their performance needs to be improved and analyzed by comparing with the performance of state-of-the-art AI techniques. Most of the work dealt with image-wise classification, while in critical disease analysis, patient-wise classification is to be considered to design an unbiased and accurate prediction model.

Furthermore, the limitation of large data requirements for training deep learning-based models can be addressed by designing an efficient feature extractor alongside the machine learning model because the quality of features heavily influences the perfor-

mance of these models. Besides, in today's world of cloud computing, IoT and mobile edge computing, developing a model that is secure, robust, and easy to run on mobile edge devices is a significant challenge that needs to be addressed. Therefore, limitations of existing approaches and the current need of the healthcare industry also act as a driving force to consider this challenging problem for our thesis.

The motivations of the present study can be summarised as follows:

1. **Need of CMT datasets:** It is already proven that CMT is an excellent model for HBC study. Severity and complexity of the disease and lack of a publicly available dataset for CMT histopathological images are the main motivation behind creating the CMTHis dataset to proceed and present in-depth analysis for research.

2. **Feature representation capability of deep learning model:** Recently, CNNs have shown their potential in computer vision due to their capability of feature representation and classification. They have been reported to be a powerful tool in the automated classification of human cancer histopathology images. This inspires us to utilize its feature representation capability to develop a novel framework based on VGGNet-16, a popular CNN variant, along with different classifiers for CAD of CMTs in this study.

3. **Data privacy and secure modelling for Cancer Histopathology Images in cloud-based services:** Adequate protection of patients' privacy helps to increase public trust in medical research. Furthermore, deep learning-based models are complex, and data processing in such models is complicated in a cloud-based approach. To address these challenges and current demand, we are also motivated to design a framework where data and learning model privacy can be incorporated.

4. **A Light Weight model for mobile edge devices:** Although CNN-based frameworks are effective in the classification of breast cancer, these algorithms are

computationally expensive and are directly not applicable on resource-constrained devices. Since providing a low-cost solution is highly valuable in the computing world, we are motivated to propose an efficient and lightweight CNN model for histopathological image classification based on MobileNet that can be utilized efficiently on edge devices.

5. **Need for a robust and efficient framework for limited and small labelled histopathological data:** A major limitation of deep learning-based frameworks is the requirement of a huge labelled dataset, and high compute resources, which limits their applicability on smaller medical datasets. On the other hand, machine learning-based approaches require less data size and computing resources; however, they have limited ability to reveal the most sophisticated features of cancer histopathological images. Therefore to enhance the performance capability of the machine learning framework, we need to design an effective illumination invariant feature extractor which is suitable to address the challenges associated with cancer histopathological images. Also, fuzzy concepts are quite popular to address the uncertainty of the model and are used widely to increase the robustness of machine learning techniques.

1.2 Challenges

The development of a classification model for cancer histopathology images is challenging due to the following reasons:

1. **Complexity of cancer histopathology images:** Histopathological image classification itself is a very challenging task because of the biological heterogeneities and rich geometrical structures. Moreover, cancer is a complex disease, and diverse histological subtypes are present for CMTs. Thus, correct interpretation of cancer histopathological images is a big challenge. Furthermore, recognition,

enumeration, and classification of mitotic figures in histopathological images are tedious tasks in many histopathological grading systems.

2. **Availability of datasets:** Limited datasets with a sufficient number of high-resolution correctly labelled histopathology images are available for HBC histopathology images. Benchmark datasets generally have hundreds of thousands of images for training. In histopathology slide analysis, we don't have that many images, and so we have to figure out a way to train a generalizable network without overfitting on the training set. No data set is currently available for CMT histopathology images. This is a major challenge for developing a model for CAD of CMTs.

3. **Data imbalance:** The majority of the datasets currently available suffer from the problem of data imbalance, which occurs at different levels. The Imbalance Ratio (IR) between malignant and benign classes in the BreakHis Dataset is 0.41 at the patient level and 0.45 at the image level. At the image and patient levels, there is also an uneven distribution between different subcategories.

4. **Colour variations in histopathology images:** Colour variations in images may occur because of several reasons, such as differences in the chemical reactivity of stains from different manufacturers, staining procedures, storage times, colour responses of slide scanners or differences in slide thickness leading to variations in transmission of light. These colour variations in histopathology images can severely affect the classification accuracies of automated systems. Resolving this issue requires stain normalization, and selecting an algorithm for stain normalization is a crucial task.

5. **Illumination:** Yet another variable in visible light images is illumination, as even slight differences in illumination conditions can affect algorithmic outcomes. Many digital microscopy images suffer from poor illumination at the peripheries (vignetting), often attributable to factors related to the light path between the

camera and the microscope. Thus, developing an illumination invariant approach for feature extraction from histopathological images is also an area of concern.

6. **Data security and privacy:** The patients expect the healthcare agencies to follow the guidelines to safeguard the confidentiality of their data and prevent its malicious use. Thus, in the modern era of health-IoTs and cloud-based clinical diagnostic models, the security and privacy of data as well as prediction models without compromising their accuracy has emerged as an open research area.

7. **Application of CAD models on resource-constrained devices:** Increased computational costs and high-resource requirements due to enormous parameters make deep learning models less affordable for resource-constrained devices. Providing a low-cost solution for histopathological diagnosis in the recent edge-computing world is another challenging area of research.

1.3 Objectives

The main objectives of this study are listed as follows:

1. To develop a dataset of histopathological images of CMTs collected from clinical cases.

2. To propose novel methods for automatic detection of cancers using histopathological images while focussing on the current demand of the healthcare industry.

3. To evaluate the effect of stain normalization and data augmentation of histopathological cancer images on the efficacy of proposed machine learning-based methods and deep learning framework.

4. To address the concern of smaller datasets, privacy of sensitive information, and efficient approach for low compute devices.

5. To validate the proposed method on the standard dataset as well as on newly collected and clinically verified datasets.

1.4 Thesis Contributions

The major contributions of the present study can be summarised as:

1. **Creation of CMTHis histopathological dataset:** In this study, for the first time, we have introduced a dataset of canine mammary tumour histopathological images.

2. **Development of VGGNet based framework for classification of CMT and HBC histopathological images**: A modified framework based on VGGNet-16, a popular CNN, has been utilized for the generation of a robust and reliable feature set. The proposed framework has presented a fused model of VGGNet-16 with Support Vector Machines (SVM) and Random Forest (RF) for binary classification of H&E stained cancer images. The model was tested on a standard HBC dataset (BreakHis) and the CMT dataset (CMTHis) introduced in this study. Besides this, the effects of data augmentation, stain normalization, and magnification on the performance of the proposed framework were also analyzed.

3. **Development of novel cloud-assisted secure deep feature classification framework for cancer histopathology images:** This work addresses the secure multi-party differentially private cancer diagnostic framework for canine mammary histopathological image classification. To the best of our knowledge, this is the first work for CMTHis classification while preserving the privacy of data as well as prediction model over cloud services. In this work, we present a novel integrated approach for histopathological image classification that combines differential privacy, secure multi-party computation, and deep learning. The pro-

posed framework is capable of reducing the risk of leakage of sensitive medical data and model under consideration by using TensorFlow privacy to facilitate training with differential privacy.

4. **Development of MobiHisNet–a lightweight CNN for mobile edge devices:** In this study, "MobiHisNet", an efficient and lightweight deep learning model was developed that uses a series of depth-wise separable convolutions to reduce computational parameters. Further, we have performed post quantization to compress the model and reduce its inference time on the edge device. The model has been quantized with Floating Point 32 (FP32), Floating Point 16 (FP16), and Integer 8 (INT8). Finally, the proposed model was tested on a Raspberry Pi and three different smartphones to demonstrate its efficiency on a lightweight processor. MobiHisNet outperforms all baseline models with the moderate model size and FLOP counts. Experiments on BreakHis datasets show the superior performance of MobiHisNet on edge devices in terms of higher accuracy, lesser complexity, and lesser memory requirements.

5. **Development of novel CoMHisP framework based on a fuzzy support vector machine with within-class density information (FSVM-WD):** This is the first work that integrates illumination invariant feature extraction and fuzzy theory for CMT classification. Thus, this work presents an interdisciplinary approach to solve one of the most challenging problems of histopathological image interpretation in the veterinary and medical sciences. It utilizes a novel feature extraction technique by optimizing the block size to extract image micro-patterns and computing center of mass (CoM) for each pixel to extract feature vectors. To handle the uncertainty in data, we also utilize the concept of fuzzy theory with SVM for classification, which presents an efficient and robust model.

1.5 Organization of the Thesis

This thesis has been organized into the following eight chapters.

Chapter 1 provides the overall introduction to this thesis consisting of histopathological cancer tissue classification and its relevance, followed by motivations, challenges, research objectives, and a summary of significant contributions.

Chapter 2 provides a detailed literature survey focusing on the importance of automated intelligent techniques for histopathological image analysis. It presents various data-driven approaches, including deep learning and machine learning for histopathological image classification. The comparative analysis of these approaches helps in identifying the research gap, presented in a summary section.

Chapter 3 describes the data acquisition process for the CMT histopathological image dataset created for the study. Furthermore, it provides brief details of a standard publicly available real-world HBC dataset, BreakHis [14], which is also used in this study for experimental purposes.

Chapter 4 introduces a deep feature extractor and classification framework for CMT and HBC. Considering the importance of correct diagnosis in patient health management, a framework is proposed to automate the diagnosis process by relieving the extra burden on pathologists. It consists of a modified VGGNet-16 model as a deep feature extractor along with potential machine learning classifiers. Finally, it discussed how augmentation, magnification, and stain normalization affected the proposal when using HBC and CMT datasets.

Chapter 5 extends the deep feature extractor module presented in Chapter 4 by addressing the feature privacy and model security in a cloud-based approach. Differential privacy and secure multi-party computation are incorporated, resulting in a novel cloud-assisted secure deep feature classification framework for cancer histopathology images. Extensive analysis is presented using HBC and CMT data, demonstrating that the pro-

posed framework can efficiently handle the trade-off between two objectives, namely model performance and privacy.

Chapter 6 examines the performance of the proposed low-cost solution "MobiHisNet" for histopathological diagnosis in edge devices. It presents an efficient and lightweight CNN model based on MobileNet that is cost-effective on low compute devices, with MobileNet acting as a feature extractor. The proposed model has been deployed successfully on a Raspberry Pi as well as three mobile devices, demonstrating its ability to run on a lightweight and portable processor. Despite its lightweight structure, it is capable of balancing the accuracy, inference time, and memory peak requirements. In-depth analysis is also presented on HBC, demonstrating that MobiHisNet is computationally faster than other state-of-the-art systems.

Chapter 7 highlights the need for an efficient illumination invariant feature extractor for a machine learning-based framework to address the histopathological classification problem with a limited dataset on low compute resources. It introduces a novel CoMHisP framework that integrates multidisciplinary concepts CoM, machine learning and fuzzy theory to design a robust and efficient classification model. A CoM based feature extractor is designed by proposing a new approach through reducing and optimizing the block size resulting in sophisticated and illumination invariant features computation. Further, uncertainty in data is addressed by a fuzzy-based classifier (FSVM-WD). Extensive results on CMT data are analyzed, demonstrating the effect of stain normalization and magnification on the proposed framework. Result shows that the proposed model outperforms in a low-cost clinical setting with a low magnification factor. It also shows superior performance as compared to deep feature and local feature descriptors.

Chapter 8 summarises the thesis work with promising future research directions in the area of histopathological image classification.