Chapter 3

Data Acquisition and Processing

This chapter presents the data acquisition procedure of the CMT histopathological images with the experimental setup. The datasets used in this study and histopathological image pre-processing steps are also detailed here.

3.1 Introduction

Researchers in computer vision, machine learning, and the medical community benefit significantly from the availability of biomedical imaging datasets. Digital pathology supports the analysis of histopathological images using AI-based algorithms. The advances in AI and computer image analysis raise the possibility of implementing CAD systems to improve histopathological interpretation and clinical care. Combining the sensitivity of AI algorithms along with the specificity and expertise of pathologists has the exciting potential to improve the efficiency, accuracy, and consistency of manual reads, particularly in a time-limited clinical setting. 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]. However, due to the complexity of the disease, it is a challenging task

to develop a CAD system for cancer classification using histopathological images.

Mammary tumours in canines are similar to HBC and are considered excellent HBC models. Diagnosis of CMTs by routine cytology of biopsy or by the extirpated gland is difficult 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.

Though CMT presents an important neoplastic disease of dogs, with mortality rates three times higher than HBC, no studies have focused on developing a system for CAD of CMTs. The main reason for this is the lack of a publicly available dataset of CMT histopathological images. Thus, we generated histopathological images from clinical cases of CMTs in this study and developed a dataset of CMT histopathological images. The images were utilized for developing programs and algorithms for CAD of CMTs. Since CMTs are an excellent model for human cancer studies, the same data set can also be utilized for HBC studies. Availability of more number of datasets of cancer histopathology images allows researchers to develop improved programs and algorithms for CAD of breast cancers. The CAD systems developed in this study were first evaluated on a standard and challenging BreakHis dataset comprising 7909 images from 82 HBC patients. Once their efficacy was proved, the model was tested on the CMTHis dataset introduced in this study. This chapter presents a brief overview of different datasets used in this study, acquisition set up for collection of CMTHis data, and data pre-processing steps used in this study.

3.2 CMT Tissue Processing and Image Acquisition Setup

This chapter introduce a dataset for CMTs, called CMTHis. The dataset was acquired from 44 clinical cases of CMTs that were presented to Referral Veterinary Polyclinics at ICAR–Indian Veterinary Research Institute (IVRI), Izatnagar. Tissue samples were fixed in 10% neutral buffered formalin, paraffin-embedded and, after cutting into 5 μm sections, were mounted on 3-aminopropyl-triethoxy-silane (APTES) coated slides. Sections were stained using H&E stain, covered with a glass coverslip, and visualized microscopically for histopathological analysis. The histopathological classification of CMT tissues was done as per Goldschmidt et al. [154], and the tissues were classified as malignant or benign as described in [155]. Histopathological analysis of H&E stained CMT tissue sections was done by experienced veterinary pathologists, and, wherever required, confirmation was done using complementary tests. Images were visualized on an Olympus BX-53 system and captured using an Olympus DP-73 Peltier cooled digital colour camera with 17.28-megapixel resolution. The images were captured using

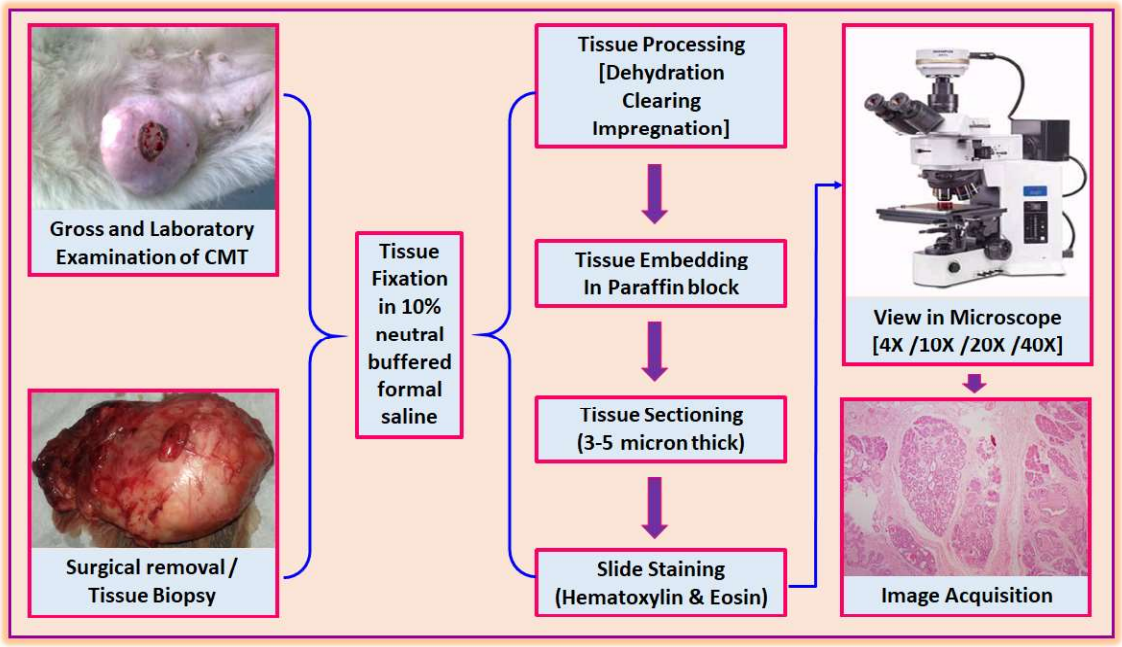

Figure 3.1: Image acquisition framework.

objective lenses of 4×, 10×, 20×, and 40× corresponding to magnifying factors of 40×, 100×, 200×, and 400×, respectively. The captured images were of high quality and clarity with reduced noise because of advanced algorithms and fine detail processing provided by the DP73 CCD camera. Thus, 1600 × 1200 pixel high-resolution RGB images with 24–bit colour depth were captured from 20 benign and 24 malignant CMT cases. Details are given in Table 3.1, and CMT tissue processing and image acquisition setup are depicted in Figure 3.1.

3.3 Datasets Description

3.3.1 CMTHis dataset

CMTHis is a dataset of CMT histopathological images introduced in this study. The dataset presently comprises 352 images collected from 44 clinical cases of CMTs presented for surgery to Referral Veterinary Polyclinics at ICAR–Indian Veterinary Research Institute (IVRI), Izatnagar. For each case, the images are captured at four different magnification (40x, 100x, 200x and 400x). Generally, for image analysis, histopathologists begin by recognizing regions of interest (ROI) at the lowest magnification level (40X), then go further for detailed analysis using increasing magnification levels (100X, 200X) until they have a profound insight (400X). Figure 3.2 shows a CMTHis slide sample captured at four different magnification settings to demonstrate this process. This dataset currently contains four histopathologically distinct types of benign tumours, namely, adenoma, ductal adenoma, fibroadenoma, and fibroma, and four malignant tumours, namely, adenocarcinoma, solid carcinoma, tubular carcinoma, and papillary carcinoma. The representative H&E stained images from the CMTHis dataset showing typical benign and malignant CMTs are illustrated in figure 3.3. CMTHis dataset image distribution in terms of class and magnification factor is displayed in Table 3.1.

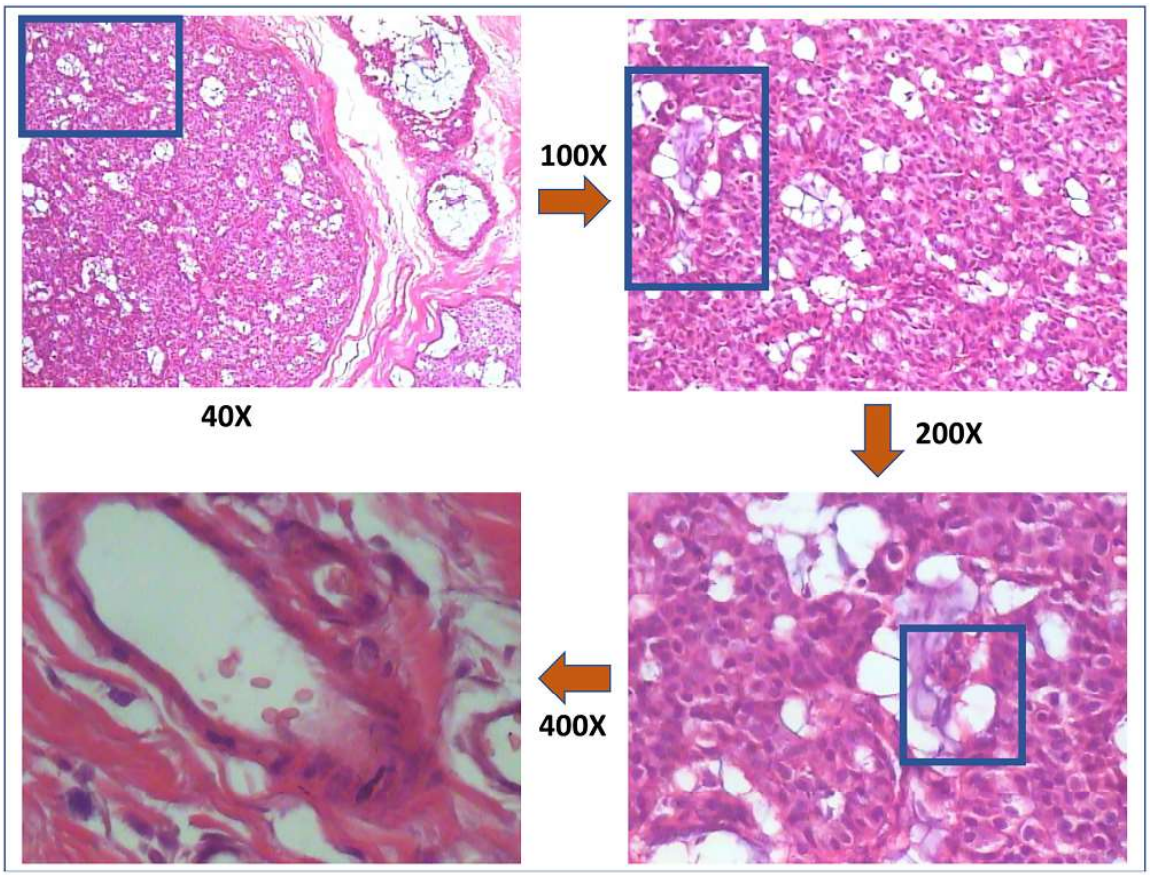

Figure 3.2: Magnifications of same tissue slide in the CMTHis dataset.

Table 3.1: CMTHis dataset image distribution in terms of class and magnification factor.

Magnification	**Benign (n=20)**	**Malignant (n=24)**	**Total (n=44)**
40×	40	48	**88**
100×	40	48	**88**
200×	40	48	**88**
400×	40	48	**88**
Total number of images	**160**	**192**	**352**

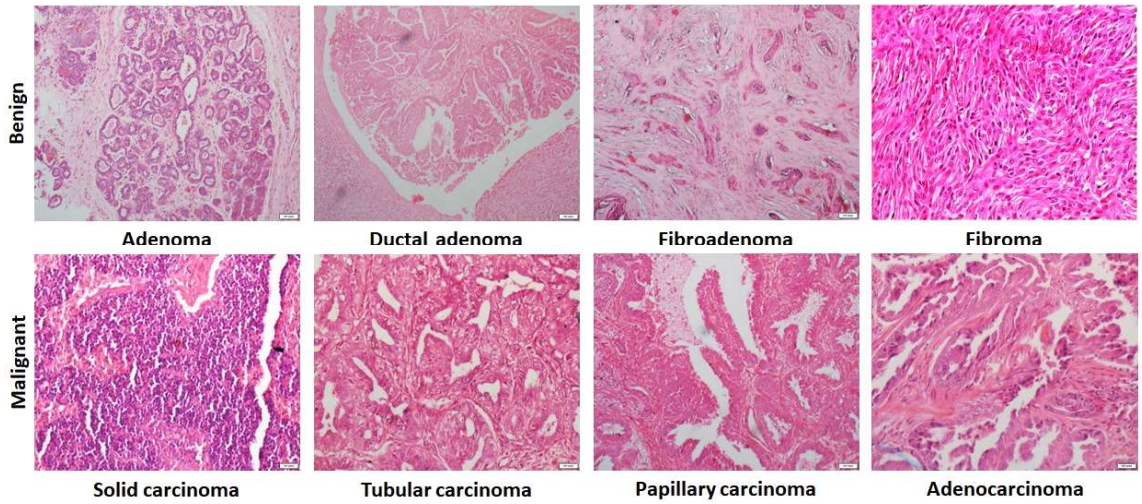

Figure 3.3: Representative H&E stained images of different classes present in CMTHis dataset.

3.3.2 BreakHis dataset

There are a number of datasets available for cancer histopathology imaging. As per our knowledge, BreakHis [14] is probably the most important breast cancer histopathology dataset, with a larger clinical relevance than the previously listed ones. BreakHis dataset contains 7909 histopathology biopsy images from 82 breast cancer patients. The P&D Laboratory – Pathological Anatomy and Cytopathology, Parana, Brazil, collaborated on the development of this dataset. P&D Laboratory gathered these images between January and December of 2014. The images are 700×460 pixels in size, with 3-channel RGB, 8-bit depth in each channel, and are saved in PNG format. This dataset has a unique ID for each subject.

BreakHis dataset is broadly categorized into two main classes: benign and malignant, with 2480 and 5429 images of each class, respectively. The images were captured at four different magnification settings (40X, 100X, 200X, and 400X). BreakHis distribution into four magnification levels for each tumour category and sub-category is presented in Table 3.2. Different numbers of images are present in each magnification subgroup, ranging from 1794 images in the 400X subset to 2051 images in the 100X sub-

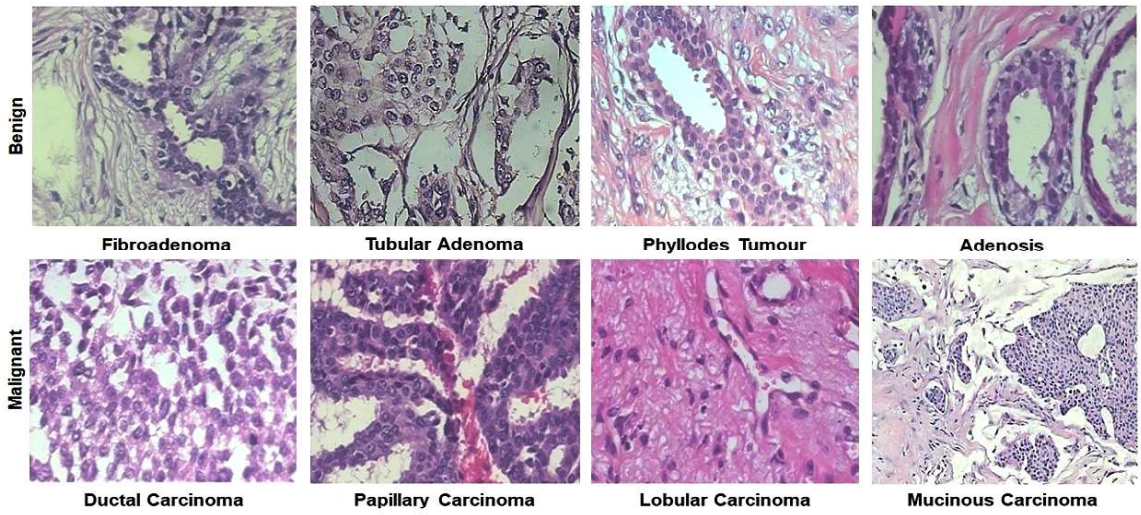

Figure 3.4: Representative H&E stained images of different classes present in BreakHis dataset

Table 3.2: BreakHis dataset image distribution in terms of class and magnification factor.

Magnification	Benign (n=24)	Malignant (n=58)	Total (n=82)
40×	625	1370	**1995**
100×	644	1437	**2081**
200×	623	1390	**2013**
400×	588	1232	**1820**
Total number of images	**2480**	**5429**	**7909**

set. Because for each patient, histopathological images are taken at all magnifications, each magnification factor subset comprises exactly 82 patients for patient distribution. In 40X, 100X, 200X, and 400X subgroups, on an average, 24, 25, 24, and 22 images per patient are available. The benign category is further subclassified into four subcategories: Adenosis, Fibroadenoma, Tubular Adenoma, and Phyllodes Tumor. The Malignant class has images of the following subclasses: Lobular Carcinoma, Ductal Carcinoma, Papillary Carcinoma, and Mucinous Carcinoma for malignant ones. The availability of well-annotated images of different cancer subtypes allows researchers to

undertake classification problems for predicting input images as benign or malignant. For each classification task, the adopted model use either a magnification-specific or a magnification-independent training scenario. In the magnification- specific approach, one model is trained on each magnification level subset, resulting in four specific models. While in the magnification-independent approach, a unique model is trained using all magnification levels combined. Since its release, a number of studies have evaluated the potential of the BreakHis dataset.

3.4 Data Pre-Processing Techniques

For developing CAD systems based upon either deep learning or other traditional approaches, the raw images need to undergo various pre-processing steps to recompensate for differences in images, which can be variations in colour, staining, and other problems, such as noise, owing to the scanning procedure. For histopathological analysis, the paraffin-embedded tissue sections are cut and stained with one or more stains to analyze the tissue's architecture and components under the microscope. The staining is used to visualize cellular components for the diagnosis of structural as well as architectural tissue analysis. H&E is the most common stain used in different laboratories, and it separates the connective tissue, cytoplasm, and nuclei. Nuclei are stained blue by Hematoxylin, while connective tissue and cytoplasm are stained pink by Eosin. The classification performance of the CAD systems depends upon the consistency of the features extracted from the images. Therefore, it is important to define the appropriate conditions under which the image pre-processing techniques will work. As noise and other illumination fluctuations severely affect the image processing techniques, eliminating these factors leads to improved performance. Various image pre-processing methods attempt to control changes in the colour, brightness and contrast of the image and reduce noise.

3.4.1 Colour and stain normalization

Histopathological images have substantial colour variations due to differences in scanning and training techniques and sample age. As it is difficult to adopt efficient colour calibration between samples [156], thus colour normalization is necessary for most circumstances. Some examples of colour normalization strategies used by researchers are deconvolution-based and histogram-based methods [11]. In [157], Anghel et al. suggested improving stain normalization in low-quality whole slide images for increasing the classification accuracy. To deal with the difficulties of stain variability, Yang and Foran [37] introduced a robust colour-based segmentation technique for histological structures that used image gradients predicted in the LUV colour space. Authors in [158] have addressed the stain variations in the image of the BreakHis dataset by learning the variation in colour and texture of these images, rather than trying to reduce the colour-variations between them. Various types of colour-texture descriptors were tried in different combinations along with various classifiers. Once the best combinations of features-classifier were identified for each magnification subset, they were combined to generate an integrated model. In [159], they tried to explore the possibility for a model to learn the colour-texture variability instead of normalizing it. Their results showed that stain normalization could be substituted by the joint colour- texture features learning to achieve superior results, and grey-scale transformation is not a good stain normalization method as it can decrease the classification accuracy. Recently, authors in [160] started from the conviction that conventional normalization techniques amplify the existing noise in images when applied directly and propose a normalization strategy that includes a noise amplification control step.

In this study, the images were pre-processed and normalized for variations in staining procedures based upon the methods given by Macenko et al.[161]. This method takes into account the staining technique used to prepare slides in histology. First, the image colours are converted to an optical density (OD) using the logarithmic transformation

Algorithm 3.1: SVD-geodesic method

Input: RGB Slide $\{(X_i, y_i)\}_{i=1}^{N}$

Output: Optimal stain vectors

$X \leftarrow$ Normalized input to [0, 1]

$A \leftarrow$ Optical_Density(X)

$A' \leftarrow$ Update_Operation (A)

$A'' \leftarrow$ SVD(A')

$(\alpha_1, \alpha_2) \leftarrow$ Two maximum values of A''

$P \leftarrow$ Construct_Plane (α_1, α_2)

Project A'' on P and normalize to unit length

$\delta \leftarrow$ Compute angle of A'' using first SVD direction

Identify the robust extremes from δ

Transform extreme value back to OD space

return: Optimal stain vectors

Algorithm 3.2: Update Operation

Input: A, β

Output: A

```
for ∀a_ij ∈ A do
    if a_ij < β then
        a_ij ← 0
    end
end
return: A
```

shown in the equation 3.1.

$$OD = -log_{10}(I), \quad where \quad I = (i_r, i_g, i_b) \tag{3.1}$$

Here, I is an RGB color vector with each component normalized to [0,1] and representing the OD converted RGB color space by following matrix A of size [**m**,**n**], where **m**

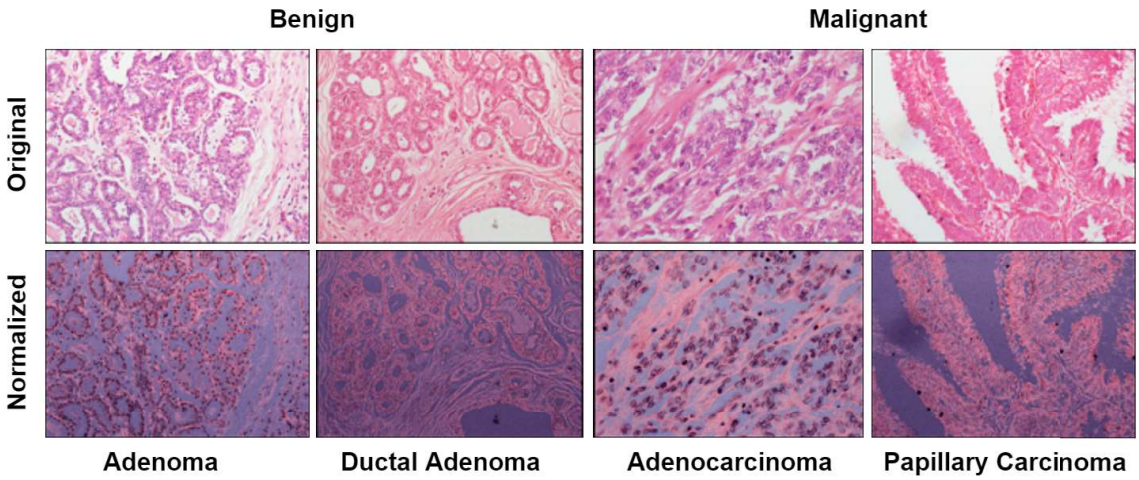

Figure 3.5: Representative H&E stained histopathological images with and without stain normalization from the CMTHis dataset.

represents number of stains and **n** is number of color channel. In our case n is 3.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Here, rows represent specific stains, and columns represent the optical density detected for each stain by the red, green, and blue channels.

Further, to acquire independent information for each stain, colour deconvolution as described in [161] was used. Here the colour values were transformed using the ortho-normal transformation of the RGB information using the equation:

$$A = VS \quad \Rightarrow S = V^{-1}A \tag{3.2}$$

In this equation, A is the observed optical density, V and S are the matrices of the stain vectors and the saturation of each of the stains, respectively.

To find 2D projections with higher variance, the singular value decomposition (SVD) algorithm is applied to the OD tuples. The resulting transformation of colour space

applies to the original image. Finally, the image histogram is extended to cover the dynamic range of the lower 90% of the data. This process is described in algorithm 3.1 and 3.2.

3.4.2 Data augmentation

One of the important concerns with deep learning models, particularly CNNs, is the requirement of huge volumes of training data, especially when they are to be fine-tuned. Features extracted from pre-trained CNN are not guaranteed to be invariant in terms of the position or orientation of the tissue in an image or image patch. Also, CNN needs enough data to achieve impressive performance. Data augmentation technique allows researchers to increase the size of data that is available for training models without collecting new data. Data augmentation makes the model more robust for feature transformation by increasing the chances of a subsequent classifier to rely on invariant features mostly or adjust to the variation within one feature. In a number of studies potential of data augmentation is explored for deep learning frameworks [162], as this approach helps in the building of more generalized models by allowing analysis on a higher volume of data. Researchers have tried various approaches for data augmentation like geometric transformations, such as rotation by 90∘, 180∘ or 270∘; positive scaling; and mirror projections, such as left-right-top-bottom. Other data augmentation techniques like cropping, padding, and horizontal flipping, are also popularly used to train neural networks systems. Authors in [163] compared different data augmentation techniques for generating a sufficient number of data samples for fine-tuning a pre-trained inception v3.

Thus, to make our model robust for feature transformation, data augmentation was performed in this study to increase the data size. Various geometric transformations, like rotation by 90^o, 180^o, 270^o, positive scaling, and mirror projection such as left-right-top-bottom, have been applied to the original image. Besides, gaussian blur was

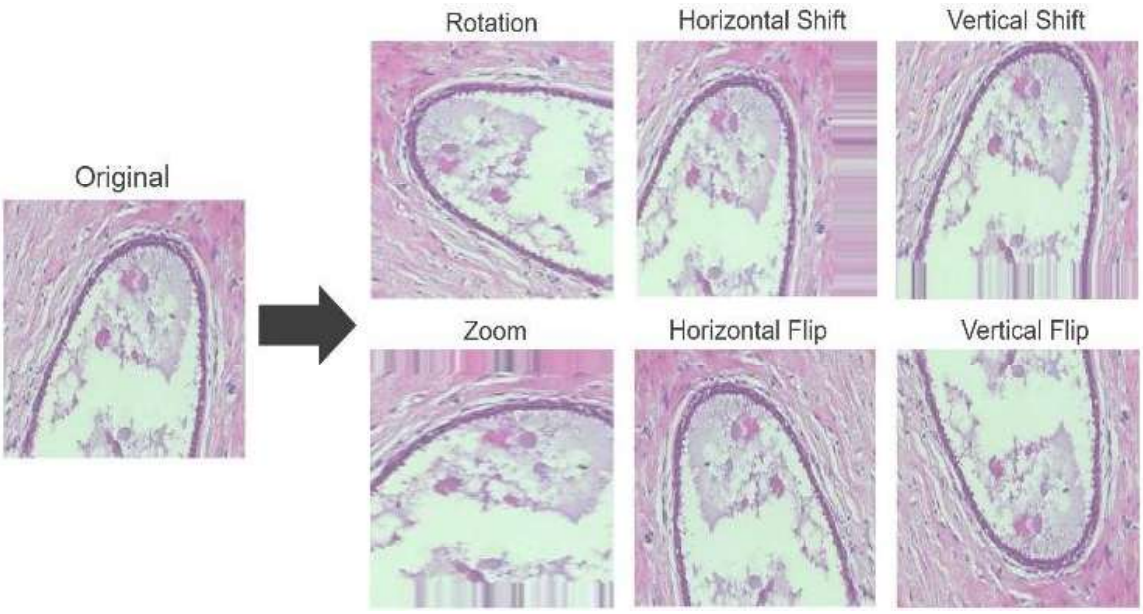

Figure 3.6: Sample images after applying data augmentation.

also used to increase data on the original image. Thus, for all magnification factors, the total number of sample images is increased by around eight times. The sample images after applying data augmentation on the BreakHis dataset are shown in Figure 3.6.

3.4.3 Data splitting protocol

One of the first decisions to make before feeding the data to the model is which samples will be used to evaluate performance. To ensure that the results are unbiased, the model should be tested on samples that were not used to create or fine-tune the model. When working with a large amount of data, a subset of samples may be saved to test the final model. The samples used to build the model are known as the "training" data set, while the "test" data set is used to validate performance. Split protocol refers to the method of dividing a whole dataset into separate chunks. To make a fair comparison, we used a split protocol similar to the one used for BreakHis dataset [14] to generate 5-folds for CMTHis data, and results were presented by taking an average of five folds. To ensure that the classifier was generalized for unseen patients, it was ensured that

the test set did not include patients used to create the training set. This was done to ensure that the efficiency of the test is assessed on a dataset that is not used to train the classifier. Since the CMTHis dataset is introduced for the first time in this study, the accuracy and performances of the proposed algorithms were first validated on a BreakHis dataset to ensure that the results were tested on the standard dataset with a large number of images. In our experiments with CMTHis, we have randomly chosen 31 patients (70%) for training and the remaining 13 for testing (30%). The results of each test are given in terms of accuracy (percentage of correctly classified instances).

3.5 Summary

The main contribution of this work is the introduction of the CMTHis dataset, which allowed the development of programs and algorithms for CAD of CMTs. The dataset was initially introduced with 352 histopathological images from 44 clinical cases of CMTs. The dataset is being expanded on a regular basis so that this dataset with a substantially large number of well-classified CMT histopathological images present opportunities for various researchers to work in the area of CAD of CMTs.