LUNG AND COLON CANCER HISTOPATHOLOGICAL IMAGE ANALYSIS

SURYANSH PATEL

Final Thesis Report

OCTOBER 2022

TABLE OF CONTENTS

DEDICATION

I dedicate this work to my family who have been my biggest cheerleaders and encouraged me at every phase of this research. A special mention to all my family members, who were impacted by lung cancer and were the reason for this research. It hasn't been easy to see friends and family suffer from this disease, which stirred me to research how Machine Learning could understand and advance the fight against lung and colon cancer.

ACKNOWLEDGEMENTS

The completion of this thesis has been one the most exciting and challenging experiences in my academic life so far. It would not have been possible without the help of many people. I would like to express my deepest gratitude to all my professors, who not only gave me the opportunity to pursue a master's program at such a great university like LJMU but also extended their support by providing an enjoyable research environment and access to the best resources. I would like to extend my sincere thanks to my co-supervisor, Professor Srinath Murugan, who was an incredible guide. Finally, I would like to acknowledge the unlimited love and support of my parents and family members.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Expansion
AI	Artificial Intelligence
AJCC	American Joint Committee on Cancer
AUC	Area under the Curve
BRIEF	Robust Independent Elementary Features
CFCMC	Cumulative Fuzzy Class Membership Criterion
CNN	Convolutional Neural Network
CT	Computed Tomography
HE	Hematoxylin and Eosin
MRI	Magnetic Resonance Imaging
NEP	Neighbor Ensemble Predictor
ORB	Oriented Fast and rotated Brief
PET	Positron Emission Tomography
SIFT	Scale Invariant Feature Transform
SURF	Speeded-Up Robust Features
TNM	TumorNode-Metastasis
UM	Unsharp Masking
WHO	World Health Organization
WSIs	Whole Slide Images
XAI	Explainable Artificial Intelligence

CHAPTER 1. INTRODUCTION

1.1. Background of the Study

Based on the data collected by the World Health Organization (WHO) between 2015 and 2016, 2 of every 5 American citizen will be identified with cancer in their lifetime. Furthermore, according to American Cancer Society, far more than 1.8 million new cancer cases are predicted to be reported in 2022. Cancer is not an apparent disease and there are many mysteries when diagnosing cancer through symptoms alone, hence Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET) scan, ultrasound, or Biopsy techniques are critical tools used to evaluate a patient [7]. In this paper, histopathology images will be used to diagnose these cancers and specifically the use of biopsy and surgical tissue specimens.

Hardware technology has significantly advanced in high-resolution image processing, coupled with advanced artificial intelligence (AI) algorithms and large amounts of data, have contributed to a production of AI applications in therapeutic images. Convolutional neural network (CNN) has ensured considerable gains in the ability to classify images and detect objects from illustrations.

If cancer can be detected early in these procedures, it can be treated and ultimately suppressed. For example, for lung cancer patients, the chance of survival for 93% of patients who are between 18 and 65 years of age is high if the cancer is detected in stage 0. The majority of cancers have four stages, according to the TumorNode-Metastasis (TNM) classification these are the phases: 0, Stage I, Stage II, Stage III, and Stage IV [2]. The four stages of colon cancer are shown in Figure 1 and lung cancer are shown in Figure 2.

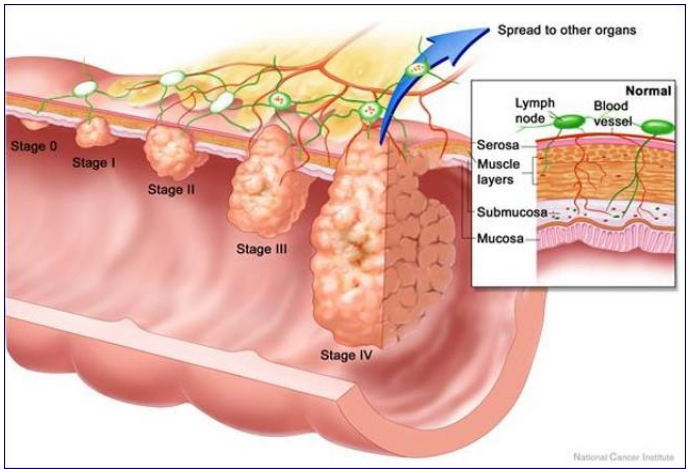


Figure 1: Colon Cancer Stages

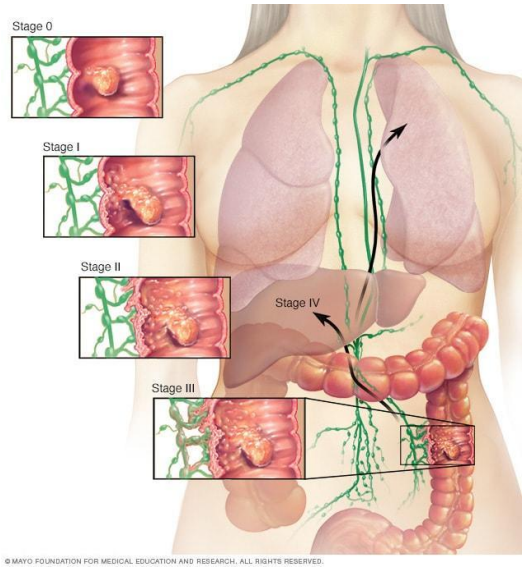


Figure 2: Lung Cancer Stages

The treatment approach considers several constraints, including the main tumour's size and location, the amount of its diffusion to lymph nodes and other organs, and the existence of any biomarkers that impact cancer spread. At certain phases, the probability of survival fluctuate dramatically. In the case of colon cancer, for example, more than 93% of patients between 18 and 65 years of age may survive with effective treatment if they are discovered at stage 0; however, survival rates linearly decrease for stage 1 at 87% and stage 2 at 74% and significantly decrease to 18% in stage 3. Cancer does not have a 100% treatment available thus the sooner a person is detected, the more time physicians must design a treatment plan for the patients, and the chance of surviving is greater. Early detection and early treatment are currently the only ways to prevent cancer-related fatalities. However, most of the population lacks access to competent diagnostic facilities, making the fight against this deadly illness even more difficult. The most common type of colon cancer is Colon Adenocarcinoma and it makes up more than 95% of all colon cancer cases. When a particular type of polyp (tissue growth) called Adenoma is developed inside the large intestine, and when it later turns into cancer it is called the Adenocarcinoma. Lung Adenocarcinoma accounts for about 40% of all lung cancers and is found more in women than men. Lung cancer cells usually develop in the glandular cells, and then spread near the alveoli inside the lungs. Cancer tumors that develop in the lungs and colon are not cancerous,they do not spread to other parts of the body. These types of tumors are called benign tumors, which are usually not life-threatening. However, they still need to be surgically removed and checked for the presence of cancer through biopsy. And finally, small-cell cancer

that develops in the lungs' air passages or bronchi. It is the second most common type of lung cancer and accounts for about 30% of all cases is called a Lung Squamous Cell Carcinoma.

1.2. Problem Statement

The purpose of this research paper is to determine the effectiveness of the CNN proposed algorithm. The attributes of the CNN architecture and algorithm work in a pattern similar to the function of the neurons inside the human brain and requires little pre-processing compared to other classification algorithms. The goal is to select the most optimal model that gives higher accuracy than 95% for the provided 25000 image datasets.

1.3. Aim and Objectives

To classify the image dataset provided into the accurate lung and colon categories:

- Clean up the imaging files before modeling
- Compare various predictive models to identify the most accurate model to classify lung and colon cancers
- Performance of the proposed CNN models on image classification is evaluated

1.4. Scope of the Study

To compare our model with the existing models in the literature, the process from the existing models was replicated. Images of colon and lung cancer from the LC25000 database were used and as the name of the dataset represents nearly 25000 images were used. Among them, 70% were used for training and the remaining for testing. Table 1 [1] compares the results achieved based on the classification of colon and lung cancer subtypes using the same dataset. Table 2 compares the results achieved based on the XGBoost model for colon cancer classification and Table 3 [1] compares the results achieved based on the lung cancer classification using the same dataset and using 10% for testing and 90% for training. Figure 3 illustrates the confusion matrix of the XGBoost model for the three types of classification. Precision, recall and f1-score of the XGBoost model for the different classes of colon and lung cancer with 70% for training and 30% for testing are shown in Table 1 [1]. Table 2 and Table 3 present the precision, recall and f1-score of the XGBoost model for the diverse classes of colon cancer and lung cancer respectively with 10% for testing.

Table 1: Classification Models using the same 25000 Image Dataset

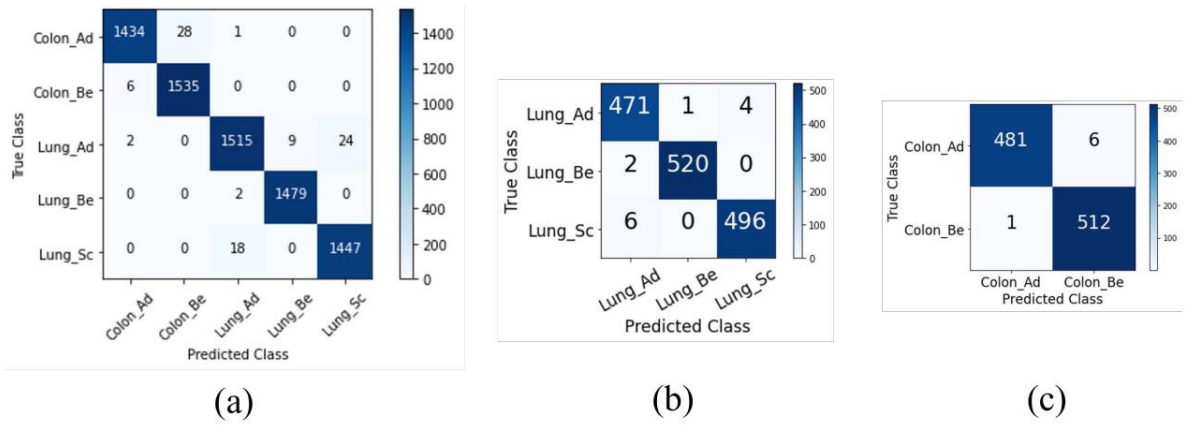
Classifier	Accuracy	Precision	Recall	F1-score
XGBoost	95.6%	95.8%	96%	95.9%
SVM	95%	95%	95.2%	95.1%
RF	94.6%	94.8%	95%	94.9%
LDA	91%	91.2%	91%	91%
MLP	92.2%	92.6%	92.4%	92.5%

Table 2: XGBoost Model for Colon Cancer

Class	Precision	Recall	F1-score	Accuracy
Colon_Ad	100%	99%	99%	99%
Colon_Be	99%	100%	99%	

Table 3: XGBoost Model for Different Classes of Lung Cancer

Class	Precision	Recall	F1-score	Accuracy
Lung_Ad	99%	99%	99%	
Lung_Be	100%	100%	100%	100%
Lung_Sc	99%	99%	99%	



(a) (b) (c)

Figure 3: Predicted Class

There are standard supervision losses, there were 11 losses and it made it hard to balance all of those objectives. Labelled depth data is required in the training phase and too much supervision from the label information makes transferring of the learned model to other datasets using different label information a difficult task.

However, 11 losses were relied upon, five of which are standard supervision losses, making it hard to balance all those objectives [3]. Labelled depth data is also required in the training phase and too much supervision from label information may make the learned model less transferable to other datasets using different label material [4]. Xu et al. [173] produced a model to learn a non-linear feature mapping from an RGB image to its thermal counterpart. They employed an architecture that reworked thermal images from the associated RGB data. However, one deconvolutional layer signifies their decoder part. This structure is too simple to achieve good rebuilding results [9]. Furthermore, it can provide limited supervision to learn the better encoder if the modernized modality object is more complex than the thermal data even though the reconstruction quality is not the main purpose.

1.5. Significance of the Study

Accurately categorizing histopathology pictures is critical in clinical practice and getting a trustworthy diagnosis of illnesses. Using ML algorithms things can be automated, and more specifically Deep Learning to replace the consuming and expensive manual labor of human specialists while also meeting the requirements for high exactness, large data sets, and other factors. Transfer learning (TL) is used in visual labeling to solve cross-domain learning issues by transferring useful knowledge from the given dataset to the task domain [3]. Through TL, knowledge from different yet related sources is transferred to target learners for improved performance on target domains. This can decrease the need for enormous data to construct target learners. TL is becoming an increasingly prevalent and auspicious area within ML as it can be applied to a broad range of applications. Different types of TL approaches include feature illustration transfer, which includes cross-domain and cross-view knowledge trade, and classifier-based knowledge exchange, which includes TrAdaboost, Support Vector Machine (SVM)-based, and generative models.

1.6. Structure of the Study

The overall structure of this study was to collect the data, prepare the data for machine learning modelling, and finally analyse the results. To collect the data, it is important to do research and find data that is high-quality and meets the requirements of the data model [8]. The requirements are images identified without any errors, no high resolution, and created with collection standards. Then prepare the data for CNN modelling, define the training, test, and validation sets, and optimize the image sizes. There are multiple CNN algorithms that could be possible solutions and used for parameter tuning. Finally, a complete analysis will be done to

gauge the results of the algorithm and provide a suggestion on the model that is most optimal to use for this dataset.

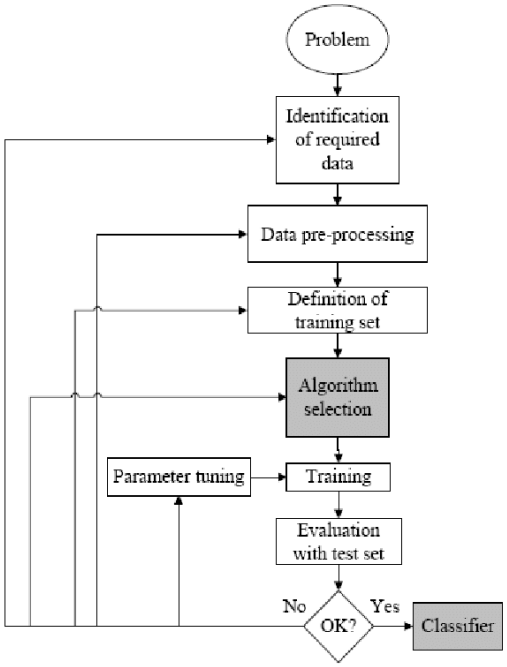


Figure 4: Structure of the Study

CHAPTER 2. LITERATURE REVIEW

2.1. Introduction

In the effort to understand how other researchers have analyzed and derived solutions on these 25000 image datasets, over 40 different literature reviews from professionals in the medical and artificial intelligence domain were analyzed. Different CNNs were used for the classification of performance and accuracy. A Hybrid Deep Learning Network used for the research of Colon Cancer enhanced accuracy by up to 90% in most cases.

2.2. Literature Review

To distinguish the human behaviors in videos, Earnest Paul Ijjina, Chalavadi Krishna Mohan [4] used a crossbreed deep learning system. They utilized CNN classifiers that provided accuracy as high as 99.68 % [6]. Machine Learning and Deep Learning algorithms were also employed to accurately categorize colon cancer. Along with this, there were other models developed to accurately detect colon cancer early. Shapcott et al. (2019) [5] strategy to classify colon cancer was different from the other models. It located the cells first and then classified them using the double CNN model. In his method, the first CNN processes the Whole Slide Images (WSIs) and then the output of this process is used as the input for the second CNN, where the classification is done. 900 patches of 500 pixels wide with a large amount of tissue were used for this sampling. The systems were based on Sirinukunwattana et al. [18], where histology pictures were used to identify and categorize the nuclei in colon cancer cells.

To ensure accurate labeling, the distance from the nucleus was measured using a deep learning approach. So the position of the nucleus is critical. This also consists of a boundary layers and a spatially limited level for spatial regression [10]. SoftMax CNN and NEP techniques were used for classification. Among the 29,756 nuclei identified, 22,444belonged to cell types such as fibroblast, epithelial, inflammatory, and others. The average weighted score of this technique was 0.784 and the multiclass AUC was 0.917. Shapcott et al.(2019) [24] also used a systematic random cell sampling to develop a cell identification algorithm. 5 TCGA diagnostic pictures and 1,500 cells were used to create a cell profile. The accuracy in this case was 65%. This alogotrithm can be used when the regions of interest are spread out.

Mangal's et al. (2020) [28] deep learning technique was based on Borkowski et al. [23]. He used Microscopic LC25000 dataset of whole slide lung and colon images. He segregated the dataset into the following five categories – lung adenocarcinomas, lung squamous cell carcinomas, lung benign, colon adenocarcinomas, and colon benign. Each category had 5000 images. The LC25000 dataset had 500 colon pictures with 1024x768 pixels. The pixels of this dataset were changed 768x768 pixels [12]. The dataset was augmented and enhanced to 25000

frames. 4500 data points were collected in each class training set and 500 data points were collected in each class test set. The photos were resized to 150x150 pixels and then transformed with randomized shear, zoom transformation, and picture normalization. One of the things that make deep learning so powerful is its ability to identify and then extract the essential semantic features, as Mangal et al. (2020) [28] explain. CNNs are a powerful tool for image classification and feature extraction. By stacking trainable layers on top of a trained classifier, CNNs can generate feature maps from particular data input. This makes them ideal for a variety of applications. The CNN architecture has the following three layers –the max pooling layer, the convolution layer, and the dense layers. The authors used a CNN architecture with the following input and training technique: they fed 150x150 pixel images with 3 color channels (RGB) into the first convolution layer [11]. The convolution layer helps the model understand the geospatial structure of the data by using adaptable filters. This method uses three convolution layers, each a 3x3 filter size, a stride of 2, and constant padding. Hence, the first layer consists of 32 filters, which are preceded by two tiers of 64 filters each. The filters are initialized using a Gaussian distribution.

ReLU activation is used in nonlinear procedures to improve performance. The max pooling layer is used to reduce the size of the convolution layer's output images. This is also known as pooling process. The pooling size is set to 2 and padding is set to valid. The max pooling approach is utilized by all pooling levels [9]. The flatten layer converts the convolution layer's output into a 1D tensor and connects it to the dense layer. The dense 5 layer takes a simple vector and provides the result in a vector. This model has two dense layers. The first layer has 512 neurons. The second layer, depending on the input class, contains either 3 for neurons for lung cancer or 2 neurons for colon cancer. The last layer's output, which is the fully connected layer's output, is activated when Softmax is activated. To prevent overfitting, a dropout layer is added across fully connected layers. This periodically eliminates neurons both from visible and hidden layers. The gradients were calculated using the RMSprop approach, which uses backpropagation. A batch size of 32 was used to adjust network weights. and the learning rate was set at 10-4, ρ was set at 0.9 and ε was set at $1e-7$. The categorical cross-entropy loss function 2 is used to keep the model's performance consistent while it is being trained. The CNN model was trained on data that was divided into three sets: training, testing, and validation. Each set comprised 80-10-10 of the total data. The model was trained for 100 iterations. Deep learning models outperformed traditional machine learning techniques for classification, with anaccuracy of approximately 96.9503% on the training and 96.6110% on the validation.

In a 2020 study, researchers from Hiroshima University Hospital in Japan detected gastric and colonic epithelial tumors by histological labelling. They developed 4,128 stomach WSIs and 4,036 colon WSIs in their study.% on the training set and 96.6110% on the validation set. The remaining 500 stomach WSIs and 500 colon WSIs (Fukuoka, Japan) were donated by Haradoi Hospital. The Hiroshima University Hospital subjects were randomly split into two groups: 500 WSIs for each organ's test set, and the rest for validation and testing (5%). The Haradoi Hospital cohorts were used as autonomous test sets, rather than for training. Only a few surgical deletion cases were included in the colon training set (5%).A publicly available TCGA dataset was used to get test sets. The Cancer Genome Atlas (TCGA) initiative provides a storage area of stomach and colon surgical excision patients[2]. The inception-v3 network was used as the core model architecture. The models were trained on 512 by 512 pixel tiles at a magnification of 20X. Tiles in specific the training sets were chosen at random and labeled as Adenocarcinoma, adenoma, or non-neoplastic. Other data augmentation like tile rotations and color changes, were used during training to improve the network's robustness and regulation. During inference, a sliding glass technique is used with input tiles of 512x512 pixels, and a fixed rate less than the input tile size[1]. A sliding window technique was employed to categorize all the tiles in the WSI. Heatmaps with smaller strides take longer to process, while heatmaps with larger strides take less time. The WSI classifications were created by using max pooling to find the tile with the maximum probability. An RNN model was trained and used to include data from all the tiles utilizing deep CNN features as input. The colon model was tested on its own test using two aggression methods: max pooling (MP-aggr) and RNN (RNN-aggr). Performance Area under the curve (AUC) scores were collected and documented. The authors also evaluated the model on a second medical institution's test set, which resulted in a unique test set. This trained inception-v3 6 network can be used as a feature extractor by omitting the last fully connected classification layer. A depth multiple of 0.35 is generator with an extracted. inception-v3 feature extractor. A sliding window with a width of 256 by 256 pixels was used to extract tiles from a WSI. To validate a WSI classification, every tile from the WSI must be checked during inference. This can be done by training an RNN model to produce a single output from any length sequence. To do this, the number of tiles required from the tissue regions of each slide must be determined and used as input to the RNN model. To ensure the model was not relying on the order of the input tiles, the order of the attributes of the tiles at each level of training was randomly changed. The RNN used LSTM layers, about 2 layers were used, and each with 30 levels in 128 hidden state approximations. The model has been trained

using stochastic gradient descent with a batch size of one. The classifier was tested for 50 epochs at 0.001 learning rate. The best performing model from the 5% validation subgroup was chosen as the final model . TensorFlow31 was used to generate and train the models. The AUCs were in Python using the scikit-learn package32 and visualized using matplotlib3 [14]. The bootstrap method34 with 1000 iterations was used to obtain the 95 per cent CIs of the AUCs. While the two-tailed DeLong's test35 was used to compare the AUCs of two correlated ROC curves, the partnered two-sided student t-test was used to compare log loss pairs [21]. This compative study helped in establishing the relationship between diagnosticians' years of expertise and accuracy. A student t-test was used to examine pathologists' and medical students' diagnosis truthfulness.

Sabol et al.(2020) [29] studied histopathological images and evaluated them with Hematoxylin and Eosin (HE) staining. The cumulative Fuzzy Class Membership Criterion (CFCMC) uses three processes to make decisions: representational explanation of the probability of misclassification, showing the training samples for the results and training samples for class conflicts. The dataset used for training consisted of 5000 small tiles with 150× 150 pixels labeled with one of eight tissue classes. CFCMC was used to classify the training set. This study used 10 non-annotated WSIs or slides of tissue, including tumor epithelium, simple stroma, complex stroma, immune cells, debris, normal mucosal glands, and adipose tissue. A total of 625 tiles were used, consisting of data for each of the tissue classes. The CNN produces the unexplainable results and the X-CFCMC produces the explainable results. The explainable results include a semantic explanation and visualization of the results. CNN results are the probability distribution and prediction and the explainable AI results are the explanations for the CFCMC. This study employed CNN to improve the accuracy of the CFCMC classifier. Though the goal was achieved, the data compression made it difficult to explain or backtrack. Furthermore, the CFCMC classifier visually represented the end result extracted from the classification of the WSIs of colorectal cancer. The results of 14 diagnosticians of Explainable Artificial Intelligence (XAI) were reviewed to assess their results. The resnet50 and CFCMC models had the highest accuracies among the 11 architectures they tested, at 93.80% and 92.78% respectively.

2.3. CNN Architecture and Training Strategy

To classify the dataset, the deep CNN was constructed with the following layers and parameters.

- Input Layer: This is the layer that loads the data, connects to feed it to the first convolution layer. The input is an image of size 150x150 pixels with color channels, which are 3 for RGB.
- Convolution Layer: this layer is used to the input image with trainable filters to learn the geospatial form of images. This model contains three convolution layers with stride set to 2 and no changes to padding and filter size 3x3,. The first layer contains 32 filters, followed by two layers with 64 filters each and they are initialized with Gaussian distribution otherwise known as a normal distribution curve. In addition, ReLU activation is applied for the non-linear operation to improve the performance. Behnke, Pooling Layer Pooling operation is used for down sampling the output images received from the convolution layer. There is a pooling layer after each convolution layer with a padding set to valid and pooling size of 2. All the pooling layers use the most common max pooling operation.
- Flatten Layer : This layer is used to convert the output from the convolution layer into a 1D tensor to connect a dense layer or fully connected layer. Dense layer or Fully connected layer - These layers treat the input as a simple vector and produce output in the form of a vector. Two dense layers are used in this model, the first layer contains 512 neurons and the last layer contains 3 and 2 neurons for lung and colon cancer respectively depending on the input class. The full connected layer has an output and its connected layer could be activated when softmax is activated.
- Dropout Layer: This layer is used between connected layers which randomly drops neurons from both visible and hidden layers to prevent overfitting of the model layers. Srivastava et al.. Table 1 illustrates the parameters of the layers, where CONV+POOL stands for convolution layer followed by a pooling layer and FC by a fully connected layer or dense layer.

2.4. Summary

For the most part, CNN includes three types of layers that are pooling, convolutional, and fully connected layers. In convolutional layers, a set of learnable filters is convolved over the image to detect specific features and patterns in the image. Then, pooling layers reduce the number of parameters learned by convolutional layers by reducing the size of feature maps. The feature maps obtained from the pooling layers are transformed into feature vectors for classification using fully connected layers. The hyperspectral image is a cube consisting of multiple channels. For example, X is an input cube with a dimension of $w \times h \times s$, where $w \times h$ is the spatial size

of the image and s is the number of channels. Output is attained by convolving filters over the entire image, and bias terms are added to the output. Finally, an activation function is applied to generate activations as shown in Equations (1) and (2). These activations from layer 1 act as the input for layer 2, and so on.

$Z[l] = W[l] * a[l-1] + b[l]$ (1) $a[l] = g(z[l])$ (2)

In this equation:

- b ([l]) is the bias of the current layer
- a [l−1] is the activation of the previous layer
- Z [l] is the output of the current convolution layer
- W([l]) are weights of the current layer
- ([l]) is the activation of the current layer
- g is the activation function

Grateful to have the opportunity to have many literatures from the most talented researchers and machine learning enthusiasts available to be able to analyse. It was because of their papers that one can grow a desire to learn and pursue the improvement of the research that is currently out there for machine learning.

CHAPTER 3. RESEARCH METHODOLOGY

3.1. Introduction

This section introduces the methodology researched in this dissertation, that is, image augmentation through CNNs. It contains details of the data that was selected for this dissertation and how it was pre-processed and wrangled. It also includes how the data transformation was conducted to finalize the data before setting it up for ingestion into the machine learning models. It also documents the results, evaluation, and interpretation conducted to showcase the best models.

3.2. Data Selection

The images of lung and colon cancer in this healthcare dataset must be classified into 5 distinct classes, such as, lung benign tissue, lung adenocarcinoma, lung squamous cell carcinoma, colon adenocarcinoma, and colon benign tissue.

There are 25000 varieties of lung and colon cancer and based on a new image observation and the historical data, the type of cancer of cancer can be predicted with an overall accuracy of 95%. 70% of the data is used for training and 30% is used for validation. The images will be classified into 5 distinct classes with an overall test accuracy of at least 95%.

The subfolder colon_image_sets contains two secondary subfolders: colon_aca subfolder with 5,000 images of colon adenocarcinomas and colon_n subfolder with 5,000 images of benign colonic tissues. The subfolder lung_image_sets contains three secondary subfolders lung_scc subfolder with 5,000 images of lung squamous cell carcinomas, lung_aca subfolder with 5,000 images of lung adenocarcinomas, and lung_n subfolder with 5,000 images of benign lung tissues.

Table 4: Contents of the LC25000 Dataset and the Assigned Class Labels

The Type of Cancer	Class Name	Class ID	Number of Samples
Colon Adenocarcinoma	Col_Ad	0	5000
Colon Benign Tissue	Col_Be	1	5000
Lung Adenocarcinoma	Lun_Ad	2	5000
Lung Benign Tissue	Lun_Be	3	5000
Lung Squamous Cell Carcinoma	Lun_SC	4	5000

Research Methodology

Through the literature review that was discussed earlier, the most optimal way to get the results required from an imaging dataset are through CNN modeling and image augmentation techniques. The following process was followed to classify the dataset:

1. Select the imaging dataset
2. Extract the features using two separata algorithms
3. Combine the features to form a feature set
4. Classify the features using the multi-channel CNN

The details of each of these processes is provided in the following sections.

3.2.1. Data Selection

This data was selected as it is the most recent data collected in 2020. Also because there could be a good number of research papers on this topic to study, analyze, compare, and provide backing. As this dataset consists of over 25000 images, it would ensure the research and model are accurate as well as timesaving. There were other datasets candidates such as the CT Chest scan dataset [6]. However, this dataset did not have a massive photo repository and the data usability reported was low. Additionally, putting data models through histological images is the biggest test.

The images of lung and colon cancer in this healthcare dataset must be classified into 5 distinct classes, such as, lung benign tissue, lung adenocarcinoma, lung squamous cell carcinoma, colon adenocarcinoma, and colon benign tissue.

There are 25000 varieties of lung and colon cancer and based on a new image observation and the historical data, the type of cancer of cancer can be predicted with an overall accuracy of 95%. 70% of the data is used for training and 30% is used for validation. The images will be classified into 5 distinct classes with an overall test accuracy of at least 95%.

The subfolder colon_image_sets contains two secondary subfolders: colon_aca subfolder with 5,000 images of colon adenocarcinomas and colon_n subfolder with 5,000 images of benign colonic tissues. The subfolder lung_image_sets contains three secondary subfolders: lung_aca subfolder with 5,000 images of lung adenocarcinomas, lung_scc subfolder with 5,000 images of lung squamous cell carcinomas, and lung_n subfolder with 5,000 images of benign lung tissues.

3.2.2. Data Pre-processing

The entire process from choosing the CT images to their corresponding ROI and predicting the disease-free survival for each patient is represented in Figure 4. The size of the original CT image is 512 * 512 with slices from 23 to 682 (mean = 162), and the valid slices which have tumor lesions of ROI for each CT image range between 3 and 77. Only slides with valid areas were retained and resized to 256 * 256 * 12 with SciPy ndimage python submodule so that it aids in the deep learning model and reduces the computational parameters. The tumor ROI area of colorectal cancer is only 1–5% of the whole CT image, hence it is challenging to extract detailed information about the tumor from the deep learning model. So, the tumor area is cropped and magnified at different magnifications. Meanwhile, the cropped CT images were also augmented by rotating at random angles and flipping with a certain probability. Finally, all the images for each patient were stacked together to feed into the neural network.

Additionally, the images had to split into their independent folders such as test, validation, and training. There were several options for the ratio between test, validation, and training.

The next step in the process is to normalize the images, which is called data re-scaling. In this process, the image data pixels (intensity) are projected to a predefined range (usually (0,1) or (-1, 1)). This ensures that the images in different data formats are converted to a common data format so that the same algorithm can be applied.

Normalization not only converts an image's pixel values to a common value but also provides the following benefits.

- Fairness across all images - High-pixel images result in strong loss while low-pixel images result in weak loss, rescaling all images to an equal range ensures the images contribute equally to the total loss.
- Standard learning rate - High-pixel images require a low learning rate and low-pixel images require a high learning rate, re-scaling images ensure standard learning rate for all images.

To enlarge the dataset without collecting new data, minor alterations are done to the existing data to increase its diversity. This is called *Data augmentation*. Some augmentation techniques are *horizontal & vertical flipping*, *rotation*, *cropping*, *shearing*, etc. Data augmentation is also done so that the neural network does not learn from irrelevant features, thereby resulting in better model performance. Data augmentation was already applied to the imaging dataset. However, the image augmentation tool was researched further and will be discussed further in the analysis.

3.2.3. Data Transformation

The image augmentation tool was used before the images were included in the CNN models. The image augmentation tool as shown in Figure 6 uses the ImageDataGenerator library. The ImageDataGenerator does random rotations, random shifts, random shifts, random brightness, and random zoom.

```
SIZE_X = SIZE_Y = 224
datagen = tf.keras.preprocessing.image.ImageDataGenerator(validation_split = 0.2)
train_set = datagen.flow_from_directory(image_set,
                                        class_mode = "categorical",
                                        target_size = (SIZE_X,SIZE_Y),
                                        color_mode="rgb",
                                        batch_size = 128,
                                        shuffle = False,
                                        subset='training',
                                        seed = 42)

validate_set = datagen.flow_from_directory(image_set,
                                        class_mode = "categorical",
                                        target_size = (SIZE_X, SIZE_Y),
                                        color_mode="rgb",
                                        batch_size = 128,
                                        shuffle = False,
                                        subset='validation',
                                        seed = 42)
```

Figure 5: ImageDataGenerator Code

Figure 6 shows overfitting, a divergence in the accuracy models and disparity in the trend of validation and loss. While the training loss is exponentially decreasing, the validation loss is increasing.

Figure 7 that both training and validation loss are decreasing with little divergence compared to the results shown in Figure 6. Furthermore, the accuracy of training and validation are increasing together. This is the advantage of using imagedatagenerator augmentation to generalize the images.

CHAPTER 4. ANALYSIS

4.1. Introduction

The images are first resized to a length of 224 and a width of 244 using a Resnet-50 pre-trained model. However, when a larger image size (400,400) was used, too many parameters were impacted that affected the model performance and it failed. With a smaller image size (150,150), the model's accuracy dropped below 35%. It required the image size to be exactly as it is used in ResNet50. Additionally, the number of training samples in one epoch or batch size had to be reduced. Figure 5 provides an overview of the image sizes and the batch size of the images.

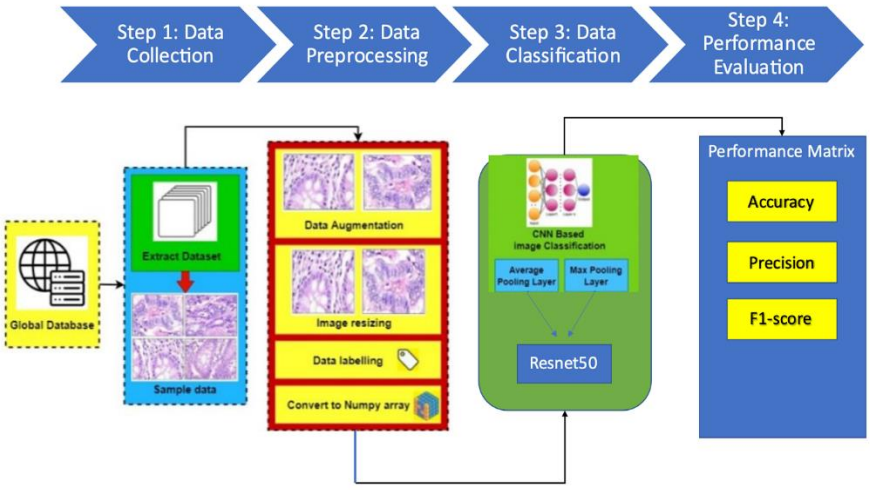


Figure 5: Overview of the Image Sizes and Batch Size

4.2. Results

The following figure shows the overall accuracy of the five-layer Resnet50 model as 99% and a loss of 8.99%.

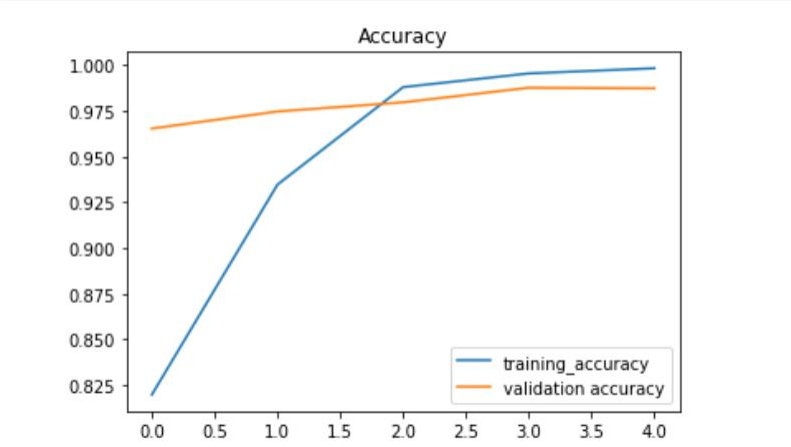


Figure 6: Accuracy of the Five-layer Resnet50 Model

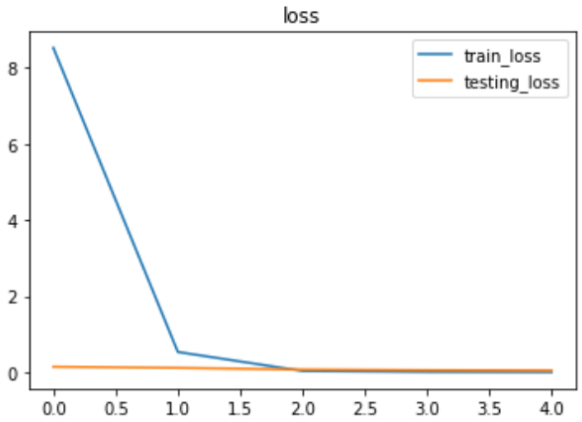


Figure 7: Five-layer Resnet50 Model Loss

The confusion matrix report shows three different classifications for Lung images – Lung_scc, Lung_n, and Lung_aca.

```
Confusion Matrix
[[988   1  13]
 [  2 998   0]
 [ 22   0 978]]
Classification Report
              precision    recall  f1-score   support

         aca       0.98      0.99      0.98      1002
           n       1.00      1.00      1.00      1000
         scc       0.99      0.98      0.98      1000

    accuracy                           0.99      3002
   macro avg       0.99      0.99      0.99      3002
weighted avg       0.99      0.99      0.99      3002
```

Figure 8: Confusion Matrix

4.3. Discussion and Conclusion

Machine learning and deep learning have helped to make significant impacts on image processing and when speaking about the medical industry specifically, there have been incredible leaps and bounds made. The proposed approach takes around 45 minutes to identify lung cancer and the specific type of lung cancer in the 15000-image dataset. Eventually, the goal is to cut down this processing time and still keep a high level of accuracy. About the Paper Deep Learning Predictive Model for Colon Cancer Patients using the VGG19 model, the processing time was 5 minutes for a 5000-image dataset with 99.67% accuracy. This time is still 67% in comparison if the ResNet50 model. When trying to use the VGG19 dataset against this model the epoch would error out at the second iteration or the second epoch. In future work, the VGG19 model should be utilized and developed to be able to handle a dataset 3 times its size to provide higher accuracy with a faster processing time than ResNet50. Medical researchers can use this model to predict lung cancers with confidence as the results of this model on the dataset show it can provide accurate results with confidence and speed.

4.4. Further Research

After image acquisition, the images must be pre-processed for image sharpening. Preprocessing images is the key to improve the quality of the image and extract important information that makes them more suitable for the learning algorithm. Unsharp Masking (UM) is a technique

used to make images sharper. It does this by enhancing the contrast of each image, which makes texture and details stand out. It works by subtracting the original image from a blurred version of the image. The unsharp masking formula is Sharpened = original + (original − blurred) × amount (1). The radius and amount parameters determine the outcome of the Unsharp Masking. The blurring step traditionally uses a Gaussian filter. The radius parameter in the unsharp masking filter refers to the sigma parameter of the Gaussian filter. The radius determines how much blurring occurs on the original image. This, in turn, affects the size of of the area around the edges that will be sharpened. To find the best values for the unsharp masking method, different radius and amount values were tested. The best results came from radius values from 1 to 5, and amount values from 1 to 20. the models performed best when the radius and amount were set to 2 and 5 respectively. Therefore, these values are used in the rest of our study. Figure 3 represents the result of enhancing a histopathological image using the Unsharp Masking method under the indicated conditions. Figure 3 A sample of colon cancer image: (a) original image and (b) sharpened image using Unsharp Masking. (https://assets.researchsquare.com/files/rs-1211832/v1_covered.pdf?c=1641239335)

CHAPTER 5. RESULTS AND CONCLUSIONS

5.1. Introduction

This section lists the results and the conclusion. The results will be shown with the accuracy and loss graphs, confusion matrix, the number of parameters analyzed, and the epochs results. The confusion matrix is one of the key tools that help to determine accuracy, prediction, recall, and specificity. Accuracy and prediction are self-explanatory; however, it is difficult to compare two models with low precision and high recall or vice versa. Hence, F-score is used to measure recall and precision at the same time. It uses harmonic mean in place of arithmetic mean by punishing the extreme values more.

5.2. Interpretation of Visualization

The following figure shows the overall accuracy of the five-layer Resnet50 model accuracy to be 99% and loss 8.99%.

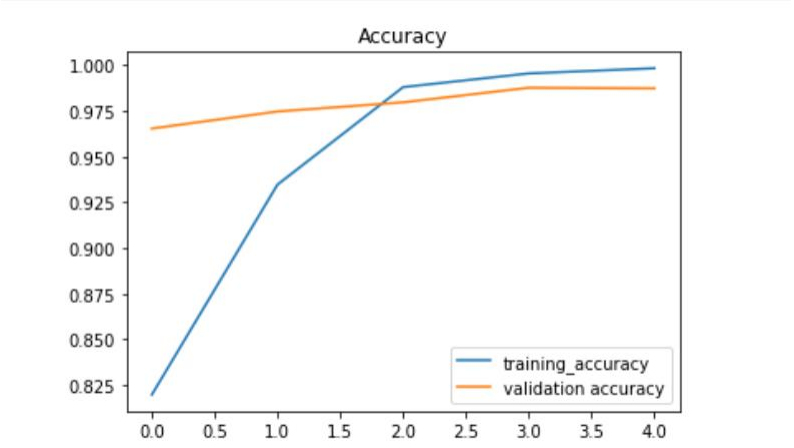


Figure 9: Accuracy of the Five-layer Resnet50 Model

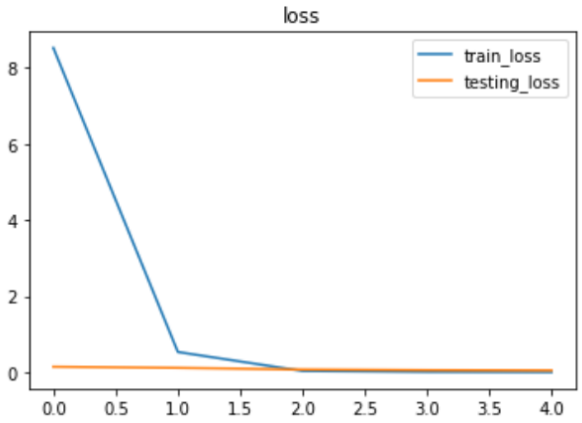


Figure 10: Five-layer Resnet50 Model Loss

The confusion matrix report shows three different classifications for Lung images – Lung_scc, Lung_n, and Lung_aca.

```
Confusion Matrix
[[988   1  13]
 [  2 998   0]
 [ 22   0 978]]
Classification Report
              precision    recall  f1-score   support

         aca       0.98      0.99      0.98      1002
           n       1.00      1.00      1.00      1000
         scc       0.99      0.98      0.98      1000

    accuracy                           0.99      3002
   macro avg       0.99      0.99      0.99      3002
weighted avg       0.99      0.99      0.99      3002
```

Figure 11: Confusion Matrix

The report shows the main category metrics precision, recall and f1-score on a per class basis. The calculated metrics use true and false positive, and true and false negatives. Positive and negative in this case are generic names for the predicted classes. The precision, recall, and f1-score are between 98-100%, so the model's performance was very good.

```
=================================================================================================
Total params: 49,311,363
Trainable params: 25,723,651
Non-trainable params: 23,587,712
_________________________________________________________________________________________________
```

Figure 12:

```
24/24 [==============================] - 494s 21s/step - loss: 0.0571 - accuracy: 0.9843
test_loss, test accuracy [0.05705161765217781, 0.9843437671661377]
```

Figure 13:

```
Epoch 1/5
94/94 [==============================] - 5669s 61s/step - loss: 14.4877 - accuracy: 0.8135 - val_loss: 2.3085 - val_accuracy: 0.8508
Epoch 2/5
94/94 [==============================] - 2510s 27s/step - loss: 2.2294 - accuracy: 0.9023 - val_loss: 0.8127 - val_accuracy: 0.9390
Epoch 3/5
94/94 [==============================] - 2588s 28s/step - loss: 0.2394 - accuracy: 0.9694 - val_loss: 0.1128 - val_accuracy: 0.9737
Epoch 4/5
94/94 [==============================] - 2515s 27s/step - loss: 0.1376 - accuracy: 0.9717 - val_loss: 0.0757 - val_accuracy: 0.9807
Epoch 5/5
94/94 [==============================] - 2545s 27s/step - loss: 0.0121 - accuracy: 0.9953 - val_loss: 0.0571 - val_accuracy: 0.9843
```

Figure 14:

Machine learning and deep learning have helped to make significant impacts on image Processing, especially in the medical industry, there have been incredible leaps and bounds. The proposed approach takes around 45 minutes to identify lung cancer as well as the type of lung cancer using the 15000 image dataset. Eventually, the goal is to cut down this processing time but retain a high level of accuracy. In the paper Deep Learning Predictive Model for Colon Cancer Patients using the VGG19 model, the processing time was 5 minutes for a 5000-image dataset with 99.67% accuracy. This time is still 67% in comparison to the ResNet50 model. When the VGG19 dataset was used against this model, the epoch would error out at the second iteration or the second epoch. In future, the VGG19 model should be utilized and developed to be able to handle a dataset 3 times its size to provide higher accuracy with a faster processing time than ResNet50. Medical researchers can use this model to predict lung cancers with confidence as the results of this model on the dataset show it can provide accurate results with confidence and speed.

After performing the comparisons, the handcrafted outcomes were better than the baseline of the convolution network model. We used a very deep network, using a highly increased dataset by the patching technique, doing a robust model with a patch voting process of several images ranked by the deep learning model. Handcrafted features are a great alternative to deep learning techniques in digital pathology, and can achieve great outcomes with fewer resources. Deep learning techniques, convolution, and recurrent networks performed the best when used together.This model was quite close to the state of the art as shown in Table 6.1. Efficient deep learning makes this is a very useful combination of technique to use in this analysis. It is designed to avoid overfitting and only use the necessary resources at the beginning. The model uses the best aspects of each technique while discarding the unsuccessful ones, like patching. It also adds an extra layer for implementation. Although it takes up considerable resources, it

is used effectively. The accuracy and area under the curve both had good outcomes. In fact, the difference between the state of art for 40X and 100X magnification accuracy was only 2%.

5.3. Conclusion

In this analysis, over 50 million parameters were analyzed, which is far more than any other dataset available. The dataset was preprocessed using an image boost method known as unsharp masking. Three feature sets were extracted for the classification of images. Machine-learning algorithms were concatenated to create a combined feature set that was fed into the algorithm. The Resnet-50 model has the best categorization performance in terms of accuracy, precision, recall and F1-score for distinguishing lung and colon cancer subtypes. Resnet-50 achieved an accuracy of 99% and a F1-score of 98.8%. The output of a ML model was used for SHAP and to evaluate how the contribution of each feature affects the model. Using this method, specialists can then understand which features of the histopathological image contributed to its classification as cancer. This paper presents a new approach to using DL in the medical field. Unlike previous papers, which used black box networks that are difficult to interpret, this new model is straightforward and can be interpreted without a machine learning background. With the right guidance, specialists can understand what is happening inside the model and use it to improve patient care.

CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

Digital pathology has not been utilized to its full potential for image processing. Several of its handcrafted features and architecture could help in enhancing image processing and addressing open issues. Along with better hardware facilities, this process can be extended to evaluate models for other cancer datasets too. With more opportunities for hardware access, a whole slide image experiment. It can also be used for spoofing, human, and object recognition. Subsequently, an original handcrafted model can be created by combining the outcomes of the features and the features implemented in the model discussed in this paper. In the interim, image classification features are being globalized and standardized.

6.1. Feature Selection for Machine Learning

This research implemented 10 strategies of machine learning techniques as training models for feature classification. These techniques ranged from the simplest techniques like KNN and Decision Trees to techniques implying boosting algorithms and kernels like Adaboost and Support Vector Machines (SVM). The techniques are selected based on the features they offer. SVM is used to find the best hyperplane to classify elements of different classes. Each side of this plane separates each class. The closest inputs to the plane are the Support Vectors. KNN and SVM separate the vector and input elements. Though SVM and KNN techniques are similar and share the same workflow, the results were contrasting. SVM performed better than KNN, in fact better than all other machine learning techniques. But the performance can be fine-tuned to improve its efficiency. Once the hyperplane is calculated, it is possible to use the model's coefficients. These coefficients are the orthogonal directs of the hyperplane. The direction of these coordinates represents the forecast class. The main features used for

classification can be identified by comparing the size of these coefficients. Though several combinations of features and different output sizes have been used and they have been fine-tuned in the histogram of oriented gradients for computer vision techniques, the comparison of the coefficients could improve several outcomes in these experiments, where the features did not represent a real alternative for image classification.

6.2. Image Features for Medical Images

Generally, medical images are used for classification, segmentation, prediction, and annotation. It is not an easy task to obtain these directly using deep learning as they require specific calculations. In this research, 5 computer vision techniques with different functions and density levels were used. The texture detection feature implemented here is different from the object detection and dissection features, which include Gaussian and Fourier transformations, pixel magnification, and inclinations. Though there are other features for object and key points detection like it would not have been able to match the outcome of the texture feature. Scientists are exploring continuously how best to leverage computer vision features to process specific medical images. Zwanenburg et al. [50] have initiated a project to standardize image biomarkers by developing features that can predict, classify, and segment. Zwanenbunrguses 11 different features inspired by features like Haralick textures and Local Binary Patterns, which have been used in the research of this paper. These features classify non-small-cell lung carcinoma images. Zwanenburg also includes a feature based on the intensity-based coefficient of variation Haralick textures, which calculates the covariance factor of the pixels in the image. The intensity ranges from minimum intensity, maximum intensity, and 10 and 9 intensity percentiles, which are based on the mean and variance intensity calculation. This intensity-based coefficient variance helps in:

- Computing features from a 2D indicator matrix and average over 2D directions and slices.
- Computing features from a single matrix after merging 2D directional matrices per slice, and then averaging over slices.
- Computing features from a single matrix after merging 2D directional matrices per direction and then averaging over directions.
- Computing features from a single matrix after merging all 2D directional matrices.
- Computing features from a 3D directional matrix and average over the 3D directions.

Based on the processes selected, the calculation varies and complements the features. Thereby enhancing the configuration to provide a more accurate result. Like intensity-based coefficient of variation used in this research paper, there are similar calculations that can be selected from

other processes like correlation, inverse variance, dissimilarity, and contrast. There is a feature available for every problem, for example, the SIFT features like SURF, BRIEF, FAST and ORB detects the key points at different conditions. The features are selected based on the new techniques implemented. They also alter the configurations, which helps in evaluating the selected features. The features and configuration affect in two ways – it aids in improving the resulting metrics and it aims at reducing the problem diversity that other features were not able to handle. To sum up, this research paper asserts that texture features can be effectively used for classification of histopathological images. In future, this can be extended in different ways by including new features or modifying the current features. Texture features can also be standardized for the classification of histopathological images. The standardization of computer vision techniques for medical images can define a precise technique for the detection of one of the principal causes of death in the world.

6.3. Future Recommendations

In this research, the fully programmed deep learning-based method for discovery of lung cancer in whole slide histopathology images was proposed. VGG16 and ResNet50 CNN architectures were assessed and the first one shows higher AUC and patch classification accuracy. Presented results show that convolutional neural networks have the potential to perform lung and colon cancer diagnosis from whole slide images, but more effort is needed to increase group accuracy. In future, the training set size will be increased and image augmentation and stain normalization will be added. Also, the focus will be on teaching from the scratch instead of using weights pretrained on ImageNet. Additionally, there are many optimizers for the Keras' Imagedatagenerator that were not accessed in this review, only the Adam optimizer was used. There are many optimizers in the literature review that were mentioned that could have worked better for the image dataset provided. This system can aid radiologists and doctors in early diagnosis, and this implementation can be further improved with higher accuracy in future studies.

REFERENCES

1. Dan XU, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe, 'Learning Cross-Modal Deep Representations for Robust Pedestrian Detection', University of Trento, The Chinese University of Hong Kong, The University of Sydney, Fondazione Bruno Kessler, University of Perugia, accessed 12 September 2022, https://openaccess.thecvf.com/content_cvpr_2017/papers/Xu_Learning_Cross-Modal_Deep_CVPR_2017_paper.pdf
2. Lian Xu, 'Deep Learning for Image Classification and Segmentation with Scarce Labelled Data', The University of Western Australia School of Physics, Mathematics and Computing, 2021, accessed 24 August 2022, https://api.research-repository.uwa.edu.au/ws/portalfiles/portal/114982246/THESIS_DOCTOR_OF_PHILOSOPHY_XU_Lian_2021.pdf
3. Monirul Islam, Belal Hossain, Nasim Akhtar, Mohammad Ali Moni, Khonodokar FIda Hasan, 'CNN Based on Transfer Learning Models Using Data Augmentation and Transformation for Detection of Concrete Crack', 15 August 2022, accessed 6 May 2022, https://www.mdpi.com/1999-4893/15/8/287/htm
4. Shahid Mehmood, Taher M. Ghazal, Muhammad Adnan Khan, Muhammad Zubair, Muhammad Tahir Naseem, Tauqeer Faiz, And Munir Ahmad, 'Malignancy Detection in Lung and Colon Histopathology Images Using Transfer Learning With Class Selective Image Processing', 10 March 2022, accessed 8 June 2022, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9709814

5. E. P. Ijjina and C. K. Mohan, 'Hybrid Deep Neural Network Model For Human Action Recognition', August 2015, *Applied Soft Computing*, Vol. 46, pp. 936–952, accessed 29 July 2022, https://www.researchgate.net/publication/282397608_Hybrid_deep_neural_network_model_for_human_action_recognition#:~:text=In%20this%20paper%2C%20we%20propose%20a%20hybrid%20deep,fusion%20of%20homogeneous%20convolutional%20neural%20network%20%28CNN%29%20classifiers.
6. M. Shapcott, K. J. Hewitt, and N. Rajpoot, 'Deep Learning With Sampling In Colon Cancer Histology', Frontiers in Bioengineering and Biotechnology, vol. 7, p. 52, 27 March 2019, accessed 24 August 2021, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6445856/
7. Mohamed Hany, 'Chest CT- Scan Image Dataset', accessed 4 August 2021, https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images.
8. 'The Confusion Matrix', accessed 5 September 2022, https://datatron.com/the-confusion-matrix/#:~:text=The%20confusion%20matrix%20is%20especially,for%20a%20binary%20classification%20problem
9. James Mygre and Dennis Sifris, 'Colon Cancer Survival Rates by Stage and Types', 20 February 2022, accessed 21 September 2022, https://www.verywellhealth.com/colon-cancer-prognosis-survival-rate-life-expectancy-5093897
10. Andrew Borkowski, Marilyn M Bui, Branoon Thomas, Catherine Wilson, 'Lung and Colon Cancer Histopatholical Image Dataset', accessed 25 September 2022, https://arxiv.org/ftp/arxiv/papers/1912/1912.12142.pdf
11. Aya Hage Chehade, Nassib Abdallah, Jean Marie Marion, Mohamad Oueidat, and Pierre Chauvet, 'Lung and Colon Cancer Classification Using Medical Imaging: A Feature Engineering Approach', 3 January 2022, accessed 25 September 2022, https://assets.researchsquare.com/files/rs-1211832/v1_covered.pdf?c=1641239335
12. Gang Yu, Kai Sun, Chao Xu, Xing-Hua SHi, Chong Wu, 'Accurate Recognition of Colorectal Cancer with Semi Supervised Deep Learning on Pathological Images', 2 November 2021, accessed 10 October 2022, https://www.nature.com/articles/s41467-021-26643-8

13. Priya Dwivedi, ‘Understanding and Coding a ResNet in Keras’, 5 January 2019, accessed 1 October 2022, https://towardsdatascience.com/understanding-and-coding-a-resnet-in-keras-446d7ff84d33
14. Nicholas Vaicaitis, ‘Histological Image’, accessed 29 September 2022, https://www.sciencedirect.com/topics/computer-science/histological-image
15. Mumtaz Ali, Riaz Ali, ‘Multi-Input Dual-Stream Capsule Network for Improved Lung and Colon Cancer Classification’, 16 August 2021, accessed 27 September 2022, https://pubmed.ncbi.nlm.nih.gov/34441419/
16. Md. AlaminTalukder, Md. ManowarulIslam, Md AshrafUddin, ArnishaAkhter, Khondokar FidaHasan, Mohammad AliMoni, ‘Machine Learning-Based Lung and Colon Cancer Detection Using Deep Feature Extraction and Ensemble Learning’, accessed 30 September 2021, https://www.sciencedirect.com/science/article/abs/pii/S0957417422009873
17. Jiatai Lin, Guoqiang Han, Xipeng Pan, Zaiyi Liu, Hao Chen, Danyi Li, Xiping Jia, Zhenwei Shi, Zhizhen Wang, Yanfen Cui, Haiming Li, Changhong Liang, Li Liang, Ying Wang, and Chu Han, ‘PDBL: Improving Histopathological Tissue Classification With Plug-and-Play Pyramidal Deep-Broad Learning’, accessed 27 September 2021, https://chuhan89.com/publication/lin-2022-pdbl/lin2022PDBL.pdf
18. Imran Hasan, Shahin Ali, Habibur Rahman, and Khairul Islam, ‘Automated Detection and Characterization of Colon Cancer with Deep Convolutional Neural Networks’, 24 August 2022, accessed 15 September 2021, https://www.hindawi.com/journals/jhe/2022/5269913/
19. Mohammad Ali Abbas, ‘The Histopathological Diagnosis of Adenocarcinoma & Squamous Cells Carcinoma of Lungs by Artificial Intelligence: A Comparative Study of Convolutional Neural Networks’, accessed 17 June 2021
20. Satvik Garg, ‘Prediction of Lung and Colon Cancer through Analysis of Histopathological Images by Utilizing Pre-Trained CNN Models with Visualization of Class Activation and Saliency Maps’, accessed 3 December 2021
21. Deiva Nayagam R, Aarthi K, Mirra S, ‘Colon Cancer Classification on Histopathological Images using Deep Learning Techniques’, accessed 3 December 2021, https://www.nature.com/articles/s41467-021-26643-8
22. Rajesh Rangan and G Srimugambigai, Prediction of Lung and Colon Cancer Using Image Processing in Machine Learning, accessed 3 December 2021, file:///C:/Users/surya/Downloads/EasyChair-Preprint-8257.pdf

23. Syed Usama Khalid Bukhari, Asmara Syed, Syed Khuzaima Arsalan Bokhari, Syed Shahzad Hussain, Syed Umar Armaghan, Syed Sajid Hussain Shah, 'The Histological Diagnosis of Colonic Adenocarcinoma by Applying Partial Self Supervised Learning', 17 August 2020, accessed 3 December 2021, https://www.medrxiv.org/content/10.1101/2020.08.15.20175760v1.full