CHAPTER I
INTRODUCTION

This chapter provides a brief overview of the research. Consist of six sections, the explanation starts from background, problem identification and objective, scope of work, research methodology, structure of this thesis, and time schedule. More detailed explanation later in the next chapter.

1.1 Background

Lung cancer is one of the most common cancers in the world. Based on data from the World Health Organization (WHO), there are 2.21 million lung cancer cases in the world, with a death rate of 1.80 million deaths in 2020 [1]. Lung cancer occupies the first position with the highest mortality rate among other types of cancer [2]

Cancer patients can be saved with proper treatment. The choice of therapy is based on the type of lung cancer experienced. For example, Non-small Cell Lung Cancer (NSCLC) can be treated with surgery, chemotherapy, radiation, and Small Cell Lung Cancer (SCLC), an aggressive type of cancer, can be treated with chemotherapy treatments sometimes in combination with radiation therapy [3]. Most lung cancer is diagnosed at the next stage where cancer has spread [4].

Based on this, detecting lung cancer at an early stage is essential to knowing the type of cancer and its treatment. Detection of lung cancer can be done by computed tomography (CT-scan), sputum cytology, and biopsy [4]. Lung cancer classification analysis based on conventional medical images (ie. CT-scan, histopathological images) is subjective and less accurate. The use of a medical images, which is generally done manually, is complicated to diagnose with high accuracy, especially in images with unclear quality and a lot of noise. An automatic detection system is needed to classify lung cancer.

Current technological advances have created many health platforms, one of which is telemedicine services [5]. Telemedicine is defined as the remote medical practice of providing health services to underserved communities using information and communication technology. It covers a wide range of medical activities, including diagnosis, treatment, disease prevention, and education [6]. One example of telemedicine is teleradiology. Tele-radiology technology to detect cancer uses a

radiological image (MRI or CT-scan) to analyze the internal structure of the human body [7]. CT-scan lung cancer image has a high size capacity and requires large data storage while data storage is limited.

Hospital-generated clinical imaging records yield terabytes of information per year [7]. For example, one of the best hospital in United Stated that has experience in lung cancer is Mayo Clinic. Mayo Clinic is the largest integrated, not-for-profit medical group practice in the world. Based on webiste , The Mayo Clinic performs more than 900 minimally invasive lung operations every year [8], which means that approximately 900 patients that produce medical image for detect diseases. The histopathological images used in this thesis have dimensions of 768 x 768 pixels and are in JPEG file format. Each image has a size of 58.9 KB. Thus, Mayo clinic has to produce 900 medical image x 58.9 KB/image = 53.010 KB or 51,76 MB per year. Thus, Mayo Clinic has requires more data storage 51,76 MB minimally per year for lung cancer diseases. There is a method to compress image data, assuming the storage requirement for lung cancer per year is 51.76 MB. If 90% of the image data is compressed, it will result in only 5.176 MB, which in turn can reduce storage usage by 46.584 MB per year and this is related to resource efficiency.

Based on these problems, developed a method based on deep learning, Convolutional Neural Network (CNN) to perform the classification using python and, application of Compressive Sensing (CS) as an alternative solution to overcome the large size medical images. Research on CNN and CS has been carried out by several researchers to detect cancer through medical images. Table 1.1 presents a summary of some cancer research purpose, method, result and differences.

Table 1.1 Summary of cancer classification research using CNN and CS

Author	Research purposes	Method	Result and Differences
[9]	Disease classification lung cancer with histopathological images	CNN	**Result** : The CNN model training and validation accuracy of 96.11 % and 97.2 % are obtained **Differences** : Research is limited to the CNN algorithm and does not use CS. and doesn't include portal classification.
[10]	The developed system was tested on various CNN architectures to get the best performance.	7 transfer learning models such as: VGG16, VGG19, ResNet101, ResNet152, MobileNetV2, DenseNet201 and InceptionV3, were pre-trained on the ImageNet.	**Result**: The best performance are obtained in 3 architectures, namely VGG19, ResNet101, and ResNet152. These architectures can identify and classify both types of colon cancer with 100% accuracy. **Differences**: This research is limited to the detection of colon cancer, still not using CS, so the dataset used still has a large size, and doesn't include portal classification.

Author	Research purposes	Method	Result and Differences
[11]	A combination method consisting of a CS algorithm, feature extraction, and KNN classification can work effectively and efficiently in telemedicine applications.	KNN and CS, with sparse technique such as: FFT, IFFT, DWT, without sparsing.	**Result** : CS worked effectively for compression with large compression ratios without having an influence on the accuracy results. The classification using KNN shows that the N image has uniquely extracted characteristics and gives accuracy up to 100%, whereas the image of ACA and SCC provided accuracy of 70% **Differences** : This research uses the KNN algorithm to classify lung cancer. but this research limited KNN. CNN produces high accuracy than KNN [12], and doesn't include portal classification.
[13]	lun cancer classification using texture extraction-based CS.	CS and KNN	**Result** : The simulation results show that two-stage texture extraction can improve accuracy by an average of 84%. **Differences** : Research is limited to analysis accuracy, does not include analyst different compression rasio, analysis using OMP recontruction Using small dataset quantity and doesn't include portal classification.

Author	Research purposes	Method	Result and Differences
Proposed	Classification lung cancer based CS image and CNN using different sparse technique, CR, and OMP reconstruction to find best model classification	CNN and CS, with sparse technique such as: DCT DWT, and FFT ,Change CR 70%, 80%, and 90% and OMP reconstruction	**Result** : CS worked work perfectly for compressed, DCT give best performance result than others sparse technique. Using higher CR lead to trade-off between image quality, best CR at 70%, The use OMP shows better performance metrics compared to recontruction without OMP **Differences** : This research is CS-based CNN for classification lung cancer to find best model classification, analysis different sparse technique, analysis CR effect for model classification, and analysis effect of using OMP and without OMP .

CS is also used for several applications such as radar [14], clinical images [15], signals [16], and so on. CS can compress data without reducing the quality of the resulting data and CS can also make the classification model better and reduce the training time used. Although the research presented in Table 1.1 has succeeded in classifying several types of cancer using deep learning. CS and CNN have been used for several study, but studies on CNN and CS for lung cancer classification have not been carried out.

This research combines methods CS and CNN to detect lung cancer. Use of CS to compress histopathological images data before training data on the CNN algorithm. The sparse method use for this research are DCT, DWT, FFT. This thesis will compare the use of the sparse technique, the effect of changing CR, the effect of using a OMP reconstruction and without using OMP reconstruction algorithm. This research will validate the accuracy of the estimation of the combination of the CS and CNN methods for the classification of lung cancer.

1.2 Problem Identification and Objective

The main goal of this thesis is to **classify lung cancer based on CS images on CNN and get a high performance system to classify lung cancer**. The whole contributions of this thesis are described as follows:

1. How is the effect of CS on cancer classification using CNN,
2. How is the effect of compression ratio on CS using CNN,
3. How is the effect different sparse techniques (DCT, DWT, and FFT) are used in CS,
4. How is the effect of using an OMP reconstruction algorithm on the model and not using an OMP reconstruction algorithm,

1.3 Scope of Work

Based on the explanation above, this thesis is limited to:

1. Using CNN and CS,
2. This thesis is limited to the classification of lung cancer,
3. Algorithm using python programming,
4. This thesis is limited to the classification of three types of cancer, namely: adenocarcinoma (ACA), squamous cell carcinoma (SCC), and benign lung cancer (N),
5. This thesis is limited to the sparse technique : DCT, DWT, and FFT.
6. This thesis limited to OMP reconstruction algorithm,
7. The results of this thesis include the evaluation metrics, incorporating PSNR, MSE, accuracy, F1 score, precision, and recall values. These metrics are applied to both the CNN algorithm with CS and an OMP reconstruction algorithm. Additionally, comparisons are made with the performance of the CNN algorithm without the OMP reconstruction algorithm.

1.4 Research Methodology

This thesis is divided into 6 work packages (WP).

- WP1: Study of literature
  This thesis studies from previous studies related to classification using deep learning CNN and CS to find problems that might be used as material in this study. This stage also explores and identifies problems from various classifications and looks for solutions that can be given.

- WP2: Requirement identification
  The identification of needs consists of research needs and system requirements. Identification of needs related to materials and methods needed to classify lung cancer. Research needs to include image datasets in the form of 3 types of cancer, ACA, SCC, and N, and identify what software specifications are needed to support research.

- WP3: System design
  Designing the program using the CS method and combining it with the CNN algorithm. The system was developed to classify lung cancer into three classes ACA, SCC, and N.

- WP4: System implementation
  The previously designed system will go through several steps: preprocessing data such as image RGB to gray, Sparse technique, CS, OMP reconstruction, split dataset, features extraction using CNN, and system classification.

- WP5: System testing & system evaluation
  The system that has been made is tested with data testing to determine the performance of the model that has been made.

- WP6: Result analysis
  the results of the testing and evaluation system will be analyzed in relation to the performance value of the CNN model, and the conclusions of the research carried out

1.5 Stucture of The Thesis

This thesis is organized as follows:

CHAPTER 1: INTRODUCTION
This chapter provides basic a brief overview of the research. Start from background, problem identification, objective, scope of work, research methodology, strucure of this thesis, and Time schedule.

CHAPTER 2: BASIC CONCEPT, PROPOSED CONVOLUTIONAL NEURAL NETWORK, AND COMPRESSIVE SENSING
This chapter provides basic concepts used in this thesis. The explanation focuses on CNN, CS, dataset.

CHAPTER 3: SYSTEM MODEL AND RESEARCH DESIGN
This chapter describes the system model including parameters and variables used in thesis, research methodology, scenario simulation and research design.

CHAPTER 4: PERFORMANCE EVALUATIONS
This chapter discusses the result of this thesis, start from the validation and observation to the performance of proposed CS based on clasification using CNN. PSNR, MSE, accuracy, f1 score, precision, and recall performance is used to analyze the performance of the proposed codes.

CHAPTER 5: CONCLUSIONS AND FUTURE WORKS
This chapter provides the conclusion of this thesis and notifies the future works.

1.6 Time Schedule

The schedule and Milestones are used as a reference as the schedule for the data Thesis:

Table 1.2 Time schedule thesis

No.	Description	Duration	Date of completion	*Milestone*
1	Study literature	2 Weeks	3-4-2022	Find refenrences about CNN dan CS
2	Requirement	3 Weeks	6-4-2022	Chapter I,II Done
3	System design	3 Weeks	6-4-2022	Chapter III Done
4	Implementation	4 Weeks	22-12-2023	Chapter IV Done
5	Testing and evaluation	1 Weeks	27-12-2023	Chapter IV Done
6	Result analysis	1 Weeks	2-1-2024	Chapter V Done