FRAMEWORK FOR LUNG IMAGES CLASSIFICATION BASED ON WEIGHTED AVERAGING ENSEMBLE AND ENHANCED EDGE DETECTION TECHNIQUES

STEFANUS KIEU TAO HWA

FACULTY OF SCIENCE AND NATURAL RESOURCES
UNIVERSITI MALAYSIA SABAH
2023

FRAMEWORK FOR LUNG IMAGES CLASSIFICATION BASED ON WEIGHTED AVERAGING ENSEMBLE AND ENHANCED EDGE DETECTION TECHNIQUES

STEFANUS KIEU TAO HWA

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF SCIENCE AND NATURAL RESOURCES
UNIVERSITI MALAYSIA SABAH
2023

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS

JUDUL : **FRAMEWORK FOR LUNG IMAGES CLASSIFICATION BASED ON WEIGHTED AVERAGING ENSEMBLE AND ENHANCED EDGE DETECTION TECHNIQUES**

IJAZAH : **DOKTOR FALSAFAH MATEMATIK**

BIDANG : **MATEMATIK**

Saya **STEFANUS KIEU TAO HWA**, Sesi **2018-2023**, mengaku membenarkan tesis Doktoral ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:-

1. Tesis ini adalah hak milik Universiti Malaysia Sabah
2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan (/):

☐	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA 1972)
☐	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
☐	TIDAK TERHAD	

Disahkan Oleh,

STEFANUS KIEU TAO HWA
MS1821021T

(Tandatangan Pustakawan)

Tarikh : 28 July 2023

(Professor Madya Dr. Abdullah Bin Bade)
Penyelia Utama

DECLARATION

I hereby declare that the material in this thesis is my own except for quotations, equations, summaries and references, which have been duly acknowledged.

21 October 2022

Stefanus Kieu Tao Hwa
MS1821021T

CERTIFICATION

NAME : **STEFANUS KIEU TAO HWA**

MATRIC NUM. : **MS1821021T**

TITLE : **FRAMEWORK FOR LUNG IMAGES CLASSIFICATION BASED ON WEIGHTED AVERAGING ENSEMBLE AND ENHANCED EDGE DETECTION TECHNIQUES**

DEGREE : **DOCTOR OF PHILOSOPHY IN MATHEMATICS**

FIELD : **MATHEMATICS**

VIVA DATE : **21 OCTOBER 2022**

CERTIFIED BY;

Signature

1. **MAIN SUPERVISOR**
   Assoc. Prof Dr. Abdullah Bade

   ____

2. **CO - SUPERVISOR**
   Assoc. Prof Dr. Mohd Hanafi Ahmad Hijazi

   ____

ACKNOWLEDGEMENT

First, I would like to thank both my supervisors, Assoc. Prof Dr Abdullah Bade and Assoc. Prof Dr Mohd Hanafi Ahmad Hijazi. They have supported me greatly, guided me and showed me directions to take. With their help, I was able to finish this project on time. Next, I would like to thank my family for giving me their support. Without them, I would not be able to complete this project. With their help and advices, I was able to continue my studies. Also, I want to thank all my friends that helped me along the way. It is good to have companions along for the journey. I thank them for their support and words of encouragement throughout my study. Also, I would like to acknowledge that the work presented is supported by Universiti Malaysia Sabah through grant SDK0191-2020 and grant SGI0135-2020. I would also like to thank Jabatan Kesihatan Negeri Sabah for providing the tuberculosis lung x-rays needed in this research. Finally, I would like to give thanks to Lord Jesus Christ for giving me this chance, guiding me along the way and helping me to overcome numerous obstacles.

Stefanus Kieu Tao Hwa
21 October 2022

ABSTRACT

Lung diseases impose a financial burden on society. Early detection of lung diseases may result in lifesaving treatments. In view of the need for an efficient treatment, scientists contend that deep learning has a great potential for diverse applications in aiding the diagnosis of lung diseases in medical imaging. In previous research, it was shown that deep learning has been utilized to classify lung diseases in a variety of publications. However, the majority of researchers employed features extracted automatically using convolutional neural networks (CNN) in their published studies. To the best of our knowledge, the number of ensemble-based works is likewise restricted. Thus, this research aims to produce a lung diseases classification framework by ensembling classifiers trained from features extracted from x-ray images and edge images. This research employs a modified edge detection technique to produce a new type of feature, uses image augmentation to increase the number of training images, and uses a modified weighted averaging ensemble to increase classification accuracy. The methods applied in this research is suitable to tackle the various problems in the field of computer vision, including limited available dataset, data imbalance and the lack of diverse features during ensemble. This research is significant because the production of a deep learning aided lung disease classification system can assist medical officers to detect lung diseases. There are three reasons to develop a computer-aided lung disease classification system. Reasons to develop this system also include reducing human workload, overcoming human exhaustion, and help health services in areas with a lack of medical expertise. In this research, classifiers were developed to classify chest x-rays into four conditions: COVID-19, pneumonia, tuberculosis, and normal (healthy). In this respect, the deep learning methods employed in this work include CNN, transfer learning, data augmentation, and ensemble. VGG16 and InceptionV3 were the CNN architectures used to extract features in this research. This is due to the fact that these two CNNs had been applied in other works of literature and have produced high accuracy classification models. Also, an enhanced Canny edge detection technique was introduced. This enhanced approach addresses many shortcomings of the conventional Canny technique and has been shown to be more accurate. This enhanced Canny approach was then used to generate an alternative edge image training dataset. With this alternative dataset available, a novel ensemble approach called accuracy-based weighted averaging was presented to combine classification result from classifiers trained from different features. This ensemble approach was utilized to increase the classification accuracy, sensitivity, and specificity of the individual classifiers by combining their probability scores. Accordingly, a closer analysis of the results reveals that the best performing ensemble combination achieved an accuracy of 92 %, a sensitivity of 98%, 86.9%, 95.6%, 87.5% for COVID-19, normal, pneumonia, and tuberculosis, respectively, and a specificity of 97.4%, 96.17%, 98.61%, 96.61% for COVID-19, normal, pneumonia, and tuberculosis. Moreover, the findings provide consistent accuracies ranging from 82 % to 96 %, indicating that this ensemble method has better classification results than single classifiers. We believe that this paradigm may be applicable to various diseases and image types, such as computed tomography scans or sputum smear microscopy images.

ABSTRAK

RANGKA PENGELASAN GAMBAR-GAMBAR PARU-PARU BERDASARKAN PENYATUAN PEMURATAAN BERPEMBERAT DAN TEKNIK PENGESANAN SISI DIUBAHSUAI

Penyakit paru-paru memberikan beban kewangan kepada masyarakat. Pengesanan awal penyakit paru-paru boleh mendatatangkan rawatan yang menyelamatkan nyawa. Berdasarkan keperluan untuk rawatan yang cekap, para saintis berpendapat bahawa pembelajaran terdalam mempunyai potensi yang besar untuk pelbagai aplikasi dalam membantu mengesankan penyakit paru-paru dalam gambar perubatan. Dalam penyelidikan dahulu, didapati bahawa pembelajaran terdalam telah digunakan untuk mengelaskan penyakit paru-paru dalam pelbagai penerbitan. Walau bagaimanapun, kebanyakan penyelidik menggunakan ciri-ciri yang diekstrak secara automatik menggunakan convolutional neural network (CNN) dalam kajian yang mereka diterbitkan. Pada pengetahuan kami, jumlah kerja berdasarkan penyatuan juga terhad. Oleh itu, penyelidikan ini bertujuan untuk menghasilkan sistem klasifikasi penyakit paru-paru dengan mengumpulkan pengelas-pengelas yang dilatih daripada ciri-ciri yang diekstrak daripada gambar sinar-x dan gambar sisi. Penyelidikan ini menggunakan teknik pengesanan sisi yang diubahsuai untuk menghasilkan jenis ciri baru, menggunakan pengubahsuaian gambar untuk meningkatkan bilangan gambar untuk latihan, dan menggunakan teknik penyatuan yang bernama pemurataan berpemberat berdasarkan ketepatan untuk meningkatkan ketepatan klasifikasi. Kaedah yang digunakan dalam penyelidikan ini sesuai untuk menangani pelbagai masalah dalam bidang penglihatan komputer, termasuk set data yang terhad, ketidakseimbangan data dan kekurangan ciri yang pelbagai semasa penyatuan. Penyelidikan ini penting kerana penghasilan sistem klasifikasi penyakit paru-paru yang dibantu oleh pembelajaran terdalam dapat membantu pegawai perubatan untuk mengesan penyakit paru-paru. Terdapat tiga sebab untuk mengembangkan sistem klasifikasi penyakit paru-paru yang dibantu oleh komputer. Sebab-sebab untuk membangunkan sistem ini termasuk mengurangkan beban kerja manusia, mengatasi keletihan manusia, dan membantu perkhidmatan kesihatan di kawasan-kawasan yang kurang berkemahiran perubatan. Dalam kajian ini, pengelas-pengelas telah dihasilkan untuk mengklasifikasikan sinar-x kepada empat keadaan: COVID-19, pneumonia, tuberkulosis, dan normal (sihat). Dalam hal ini, kaedah pembelajaran terdalam yang digunakan dalam kajian ini termasuk CNN, pembelajaran pemindahan, pengubahsuaian data, dan penyatuan. VGG16 dan InceptionV3 adalah senibina CNN yang digunakan untuk mengekstrak ciri-ciri dalam kajian ini. Ini adalah kerana kedua-dua CNN ini telah digunakan oleh kajian lain dan menghasilkan model klasifikasi yang tinggi ketepatannya. Selain itu, teknik pengesanan sisi Canny yang diperbaiki diperkenalkan. Pendekatan Canny yang diperbaiki ini menangani banyak kelemahan teknik Canny klasik dan telah terbukti lebih tepat. Pendekatan Canny yang diperbaiki ini kemudian digunakan untuk menghasilkan dataset latihan gambar sisi alternatif. Dengan adanya dataset alternatif ini, pendekatan penyatuan baharu yang digelar sebagai pemurataan berpemberat berdasarkan ketepatan telah diperkenalkan untuk menggabungkan hasil klasifikasi daripada pengelas yang dilatih dari ciri yang berbeza. Pendekatan penyatuan ini digunakan untuk meningkatkan ketepatan, kepekaan, dan spesifisiti pengelas *individu dengan menggabungkan skor kebarangkalian mereka. Oleh itu, analisis yang lebih terperinci menunjukkan bahawa kombinasi penyatuan yang paling bagus mencapai ketepatan 92%, kepekaan 98%, 86.9%, 95.6%, 87.5% untuk COVID-19, normal, pneumonia, dan tuberkulosis, masing-masing, dan spesifisiti 97.4%, 96.17%, 98.61%, 96.61% untuk COVID-19, normal, pneumonia, dan tuberkulosis. Selain itu, hasil kajian memberikan ketepatan yang konsisten dalam julat 82% hingga 96%, menunjukkan bahawa kaedah penyatuan ini mempunyai hasil klasifikasi yang lebih baik daripada pengelas individu. Kami percaya bahawa cara ini boleh digunakan dalam pelbagai penyakit dan jenis gambar, seperti imbasan tomografi berkomputer atau gambar mikroskopi sapuan kahak.*

LIST OF CONTENTS

CHAPTER 3 : METHODOLOGY

CHAPTER 4 : ENHANCED CANNY EDGE DETECTION

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

CNN	-	Convolutional Neural Network
COVID-19	-	Coronavirus Disease 2019
CT	-	Computed Tomography
DBN	-	Deep Belief Network
DDA	-	Digital Differential Analyzer
FN	-	False Negative
FP	-	False Positive
MSE	-	Mean Square Error
RBM	-	Restricted Boltzmann Machine
RNN	-	Recurrent Neural Network
SSIM	-	Structural Similarity Index
TB	-	Tuberculosis
TN	-	True Negative
TP	-	True Positive

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 The Problem of Lung Diseases

Lung diseases are deadly, killing millions of people annually, and they cause suffering and unease to human beings (Forum of International Respiratory Societies, 2017). Lung diseases also impose a financial burden on the society. Early detection of lung diseases is critical to decreasing morbidity and mortality (Yahiaoui *et al.*, 2017). To avoid deaths caused by lung diseases, there is a need for early lung disease detection. The earlier the disease was detected, the sooner the appropriate treatment can be administered, thus resulting in a higher chance of recovery and survivability.

Traditionally, most lung diseases are detectable via chest x-ray inspection and computed tomography (CT) scan inspection (Setio *et al.*, 2017), blood analysis, skin analysis, and sputum sample analysis (American Thoracic Society, 2000). However, many of the procedures for diagnosing lung diseases are costly and time consuming, especially for early-stage detection. Many cases were detected at the advanced stage where the patients have a very low survival chance.

Human errors may occur because of factors such as similarity of veins, tissues and small nodules present in the x-ray during the initial stage of the disease. X-rays can be consulted to detect lung diseases to some extent, but they cannot guarantee an accurate diagnosis of which infection was afflicted. For centuries, radiologists encountered the problem of distinguishing different lung diseases, because they mimicked each other (Hammen, 2015). Furthermore, detecting lung diseases from medical images requires the presence of medical experts to give an opinion on the images. Images with poor quality also hinder human performance.