Faculty of Information and Communication Technology

DEVELOPMENT OF LUNG CANCER PREDICTION SYSTEM USING META-HEURISTIC OPTIMIZED DEEP LEARNING MODEL

Pethuraj Mohamed Shakeel

2023

DEVELOPMENT OF LUNG CANCER PREDICTION SYSTEM USING META-HEURISTIC OPTIMIZED DEEP LEARNING MODEL

PETHURAJ MOHAMED SHAKEEL

A thesis submitted
in fulfilment of the requirements for the degree of Doctor of Philosophy

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2023

DEDICATION

The name of Allah SWT, my Creator and my Master. My great teacher and messenger, Prophet Muhammad SAW (May Allah SWT bless and grant him), who taught us the purpose of life. Universiti Teknikal Malaysia Melaka; my second magnificent home. My great parents, who never stop giving of themselves in countless ways. My dear wife leads me through the valley of darkness with the light of hope and support. My beloved brothers and sisters and my beloved kid Shahin, whom I can't force myself to stop loving. To all my family, the symbol of love and giving. My friends who encourage and support me. I dedicate this research.

ABSTRACT

Lung cancer is a serious disease that completely affects the human respiratory system for both men and women. The lung cancer symptoms create complexity in detecting lung cancer in the earlier stage. For this purpose, an automatic computer-aided lung cancer detection system is required to minimize the mortality rate by recognizing it in an earlier stage. The traditional system fails to predict the accurate affected cancer region with minimum computation complexity and error rate. For overcoming the difficulties, effective and optimized meta-heuristic machine learning techniques indicate lung cancer in the earlier stage. The proposed optimized automatic lung cancer prediction process minimizes the entire miss-classification and improves lung cancer recognition accuracy. The machine learning techniques are used to predict lung cancer in this study. First, the Computed Tomography (CT) images obtained from the Cancer Imaging Archive (CIA) dataset processed with the help of a weighted mean histogram equalization approach that used to eliminate the noise information from CT image. After that cancer-affected region in the lung is segmented with the help of the proposed Butterfly Optimization Algorithm-based K-Means Clustering (BOAKMC) algorithm. The algorithm detects the affected region depending on pixel similarity computation process. Then different features are derived from the segmented region using Gray Intensity Co-Occurrence Distribution Matrix (GICDM) which is processed by applying a proposed Supervised Jaya Optimized Rough Set based Feature Selection (SJORSFS) algorithm. These algorithms select the best features according to the fitness value, and its redundancy is to be reduced. Finally, the classification is implemented using an ensemble classifier, deep learning instantaneously trained a neural network and an Autoencoder-based Recurrent Neural Network (ARNN) classification algorithm. The proposed lung cancer prediction model recognizes the lung cancer up to 96.39% of accuracy, 0.981% of precision value, 0.9839% of F1-score, 6.438% of false positive rate and 448.607ms of classification time.

PEMBANGUNAN SISTEM RAMALAN KANSER PARU-PARU MENGGUNAKAN MODEL PEMBELAJARAN MENDALAM DIOPTIMUMKAN MELALUI META-HEURISTIK

ABSTRAK

Kanser paru-paru adalah penyakit serius yang menjejaskan sepenuhnya sistem pernafasan manusia untuk lelaki dan wanita. Gejala kanser paru-paru mewujudkan kerumitan dalam mengesan kanser paru-paru pada peringkat awal. Untuk tujuan ini, sistem pengesanan kanser paru-paru bantuan komputer automatik diperlukan untuk meminimumkan kadar kematian dengan mengenalinya pada peringkat awal. Sistem tradisional gagal meramalkan kawasan kanser terjejas dengan tepat dengan kerumitan pengiraan minimum dan kadar ralat. Untuk mengatasi kesukaran, teknik pembelajaran mesin meta-heuristik yang berkesan dan dioptimumkan menunjukkan kanser paru-paru pada peringkat awal. Proses ramalan kanser paru-paru automatik yang dioptimumkan meminimumkan keseluruhan klasifikasi kesilapan dan meningkatkan ketepatan pengecaman kanser paru-paru. Teknik pembelajaran mesin digunakan untuk meramalkan kanser paru-paru dalam kajian ini. Pertama, imej Tomografi Berkomputer (CT) yang diperoleh daripada dataset Arkib Pengimejan Kanser (CIA) diproses dengan bantuan pendekatan penyamaan histogram min berwajaran yang digunakan untuk menghapuskan maklumat hingar daripada imej CT. Selepas itu kawasan yang terkena kanser di dalam paru-paru dibahagikan dengan bantuan algoritma Pengoptimuman Rama-Rama berasaskan Algoritma Kelompok K-Means (BOAKMC) yang dicadangkan. Algoritma mengesan kawasan yang terjejas bergantung pada proses pengiraan persamaan piksel. Kemudian ciri-ciri yang berbeza diperolehi daripada wilayah bersegmen menggunakan Matrik Taburan Kejadian Bersama Intensiti Kelabu yang diproses dengan menggunakan algoritma Pemilihan Ciri berasaskan Set Kasar Dioptimumkan Jaya Diselia yang dicadangkan. Algoritma ini memilih ciri terbaik mengikut nilai kecergasan, dan pertindihannya akan dikurangkan. Akhir sekali, pengelasan dilaksanakan menggunakan pengelas keseluruhan, pembelajaran mendalam serta-merta melatih rangkaian saraf dan algoritma pengelasan Rangkaian Neural Berulang (ARNN) berasaskan pengekod automatik. Model ramalan kanser paru-paru yang dicadangkan mengenal kanser paru-paru sehingga 96.39% ketepatan, 0.981% nilai ketepatan, 0.9839% daripada skor F1, 6.438% daripada kadar positif Palsu dan 448.607ms masa pengelasan.

ACKNOWLEDGMENTS

In the Name of Allah SWT, the Most Merciful, the Most Compassionate, all praise be to Allah SWT, the Lord of the worlds; prayers and peace be upon Prophet Muhammad SAW, His servant and messenger. First and foremost, I must acknowledge my limitless thanks to Allah SWT, the Ever Magnificent, the Ever Thankful, for all their help and blessing.

I owe a deep debt of gratitude to Universiti Teknikal Malaysia Melaka (UTeM), for allowing me to complete this study. I am grateful to some people who worked hard with me from the beginning till the completion of the present research, especially my supervisor Professor Dr. Ts. Burhanuddin Mohd. Aboobaider from the Faculty of Information and Communication Technology, UTeM, for his essential supervision, support, and encouragement to complete this thesis.

Moreover, I would also like to thank my co-supervisor, Dr. Lizawati Binti Salahuddin, for her valuable guidance, scholarly input and consistent encouragement I received throughout the research work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ARNN	-	Autoencoder Recurrent Neural Networks
BOA	-	Butterfly Optimization Algorithm
CIA	-	Cancer imaging Archive
CNN	-	Convolution Neural Networks
CT	-	Computed Tomography
DNA	-	Deoxyribonucleic Acid
DNN	-	Deep Neural Networks
EFST	-	Ensemble Feature Selection Technique
JO	-	Jaya Optimization
LDCT	-	Low-dose C.T scan
MBPA	-	Multilevel Brightness Preserving Approach
NAACCR	-	North American Association of Central Cancer Registries
NN	-	Neural Networks
PET	-	Positron Emission Tomography
RS	-	Rough set
SAE	-	Stack Encoder
SSCF	-	Semantic Segmentation and Classification Framework
SVM	-	Support Vector Machine
TT-CNN	-	Transferable Texture Convolutional Neural Network
WONN-MLB	-	Weight Optimized Neural Network with Maximum Likelihood Boosting

LIST OF PUBLICATIONS

1. Shakeel, P.M., Burhanuddin, M.A. and Desa, M.I., 2019. Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks. *Measurement*, 145, pp. 702-712. (Index ISI- Q1)
2. Mohamed Shakeel, P., Desa, M. I., and Burhanuddin, M. A., 2020. Improved watershed histogram thresholding with probabilistic neural networks for lung cancer diagnosis for CBMIR systems. *Multimedia tools and applications*, 79(23), pp. 17115-17133. (Index ISI- Q2)
3. Shakeel, P.M., Burhanuddin, M.A. and Desa, M.I., 2020. Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier. *Neural Computing and Applications,* 34, pp. 9579–9592 (2022). *(Index* ISI-Q1)
4. Shakeel, P. M., bin Mohd Aboobaider, B., and Salahuddin, L. B., 2022. Detecting Lung Cancer Region from CT Image Using Meta-Heuristic Optimized Segmentation Approach. *International Journal of Pattern Recognition and Artificial Intelligence*, 2240001. (Index ISI- Q3)
5. Pethuraj, M. S., bin Mohd Aboobaider, B., and Salahuddin, L. B., 2023. Developing lung cancer post-diagnosis system using pervasive data analytic framework. *Computers and Electrical Engineering*, 105, p.108528. (Index ISI- Q1)

6. Pethuraj, M. S., Aboobaider, B. B. M., and Salahuddin, L. B., 2022. Analyzing CT images for detecting lung cancer by applying the computational intelligence-based optimization techniques. *Computational Intelligence*. (Index ISI- Q2)

CHAPTER 1

INTRODUCTION

1.1 Introduction

The study intends to create automatic computer systems for detecting lung cancer. Lung cancer (Kroeze et al., 2023) is one of the most affected cancers and damaging the human respiratory system. According to the survey, lung cancer's mortality rate increases daily because the survival rate (Sharma, Soni and Agarwal, 2018) of lung cancer is shallow. Once people are affected by lung cancer, it is identified by several symptoms (Bradley et al., 2019), such as shortage of breath, blood coughing, bone fracture, chest pain, weight loss, facial swelling, neurological issue, voice change, sputum colour modification, bleeding, memory loss, headache, and joint pain. These symptoms are often linked with other diseases, so lung cancer should need more attention to predict precisely (van Os et al., 2022). For this purpose, several automatic systems are created with the help of machine learning techniques and intelligent approaches (Dev et al., 2019). The established policies challenge predicting the exact lung cancer information with minimum deviation. The higher difference reduces lung cancer prediction accuracy and maximizes the computation complexity. To overcome the above issues, effective and optimized techniques (Sharma et al., 2018) are applied to get the affected region because the exact segmented region aids in improving the overall prediction accuracy. Analyzing the segmentation techniques, optimized feature selection, and classifiers needed to enhance the whole cancer recognition process. Thus, the primary

intention of the study is to maximize the lung cancer prediction rate with minimum deviations and low computation complexity.

1.2 Motivation of the study

This section discusses the purpose of the earlier lung cancer prediction system. Traditionally, doctors examine patients manually (Zhang et al., 2019) once the patient is affected by any disease. The doctors analyse the patients by asking questions and reading their signs, symptoms, previous medical histories, and health care information. From the collected details, medical records provide treatment and medical advice. During the lung cancer identification process, a lung screen is utilized, which includes a set of questions to predict the symptoms of lung cancer that helps to identify lung cancer around the next 10 to 20 years. In addition, several self-identification applications are created to diagnose their lung cancer status using their health behaviour. The self-examination does not provide the maximum accuracy because the lung cancer symptoms are sometimes relevant to other diseases. The lung cancer is predicted with the help of the CT and X-ray images that are used to increase the recognition rate. For this purpose, an earlier prediction system is needed to forecast lung cancer at the starting stage. The created earlier lung cancer detection system (Yang et al., 2019) helps understand the cancer stages, symptoms, and learning process to predict cancer effectively. The earlier prediction system uses the predictive analytics concept while identifying lung cancer. The predictive analysis process (Wang et al., 2019) uses the previous medical data that successfully guess the future impacts and patterns of lung cancer. Several predictive tools (Dranitsaris et al., 2017) are used during this predictive analytic process to get knowledge about lung cancer, personalized behaviour, and other healthy activities. From the detected activities, the status of lung cancer is recognized successfully.

The predictive analysis in earlier lung cancer detection systems has several significances. The created system helps decide on cancer, visualize the patient's health data, recognize the cancer pattern, and utilize machine learning techniques (Jiang et al., 2017) to learn the system, patient behaviour identification process, and effective models to eliminate the future medical issue. In addition to this, the predictive analytics in earlier cancer detection systems makes prediction fast because of the effective collection of lung data and predict the future impacts using specific learning concepts due to this importance, the predictive analysis algorithm used in this study earlier cancer detection system.

This study aims to create an optimized earlier lung cancer detection system to predict the cancer in the treatable stage, minimize the cancer mortality rate, increase the recognition accuracy, it helps to train the system the ability to provide the lung cancer pattern for getting an idea in the future prediction process, maximize the people's survival rate, creating earlier screening awareness for people about lung cancer, maximize the successful treatment ratio and understand the impact of disease. Due to the above reasons, this study concentrates on developing the lung cancer prediction system in the earlier stage. As discussed earlier, the screening process includes machine learning methods, in which the thesis's purpose is achieved by applying effective artificial intelligent approaches. The successful utilization of the intelligent techniques and earlier lung cancer prediction system is deliberated in the next chapters.

1.3 Significance of the study

This study's main intention is to assess patients and investigate their health using a few critical factors to predict lung cancer. It affects both genders, and it is the second most dangerous disease. The mortality rate of lung cancer increases day by day due to people's unwanted behaviour. The higher smoking behaviour leads to an increase in the mortality rate. Therefore, it is necessary to provide awareness and treatment before they are affected by lung cancer.

For this purpose, the automatic system is created because the successful identification of cancer-affected regions and cancer cells helps provide the proper treatment to the patient. If the procedure is adequately presented, the survival rate of people is increased more. The generated system can collect cancer data and store it in different patterns. The patterns are continuously trained to identify lung cancer that will affect the patient in the future. This process reduces the complexity in the future direction and maximizes the cancer prediction rate with minimum time. Due to this significance, the earlier lung cancer prediction system was created using optimized techniques.

1.4 Problem statement

According to the report of World Health Organization (WHO), International Agency for Research on Cancer states that 10.6% of people suffered due to the lung cancer. Around 23052 cases of men, 25587 cases of women were examined as lung cancer in 2020. This serious lung cancer patient has been estimated by analyzing several symptoms (MacLean et al., 2017) which are diagnosed through various screening processes (Gregg, Li and Yoneda, 2019). Among the various screening biopsy diagnosis (Tam et al., 2016), methodologies

such as MRI, CT, and PET-related process ensures effective results while examining lung cancer (McGregor et al., 2017) because the biopsy effectively examines each cell in the lung. In hospital, imaging test, biopsies, blood test and pulmonary function tests are utilized to identify the tumor. However, imperfect boundary detection and pixel misclassification result in uncertainties. It degrades the implication of computer-aided image analysis, wherein a more sophisticated method is required. The method designed by Liu, Zhao, and Pang (2019) does not achieve the expected swiftness in training the different classes due to multiple and varying resource blocks. Contrarily, using datasets, the method proposed by Chen et al. (2019) is limited. A differential dataset with random input would help to optimize the segmentation process. The R-CNN method (Jia et al., 2020) requires more time to segment the input. The traditional recognition procedure consumes more time and requires experts in the field, increasing computation complexity and cost. Hence the earlier lung cancer expert system predicts abnormal lung features from the feature list. Unwanted distortions in the image data are led to reduce the prediction of segmentation accuracy (Chen, et al., 2018). The overlapping images are decreasing the segmentation process. The primary objective of segmentation is to represent an image that is more expressive and more easily analyzed in approximation. The goal of the lung segmentation in the CT image is a too good direction in the processing. Another important issue is the irrelevant features are extracted and it affects the classification accuracy (Chen et al., 2019). Currently various classifications methods are utilized for prediction. However, its need to improve prediction accuracy (Raweh et al., 2018). Therefore, several researchers develop CAD system but still it requires to predict lung cancer in early stages with less time (Wang et al., 2018).

1.5 Research questions about lung cancer

This section discusses the few research questions regarding lung cancer because they created an automatic lung cancer prediction system. The cancer detection system is to be worked according to the research question to prove the system's efficiency.

i. What way to eliminate the distortions and how the image quality is enhanced?
ii. How the overlapping images are eliminated and improve the ROI recognition accuracy?
iii. What are the irrelevant features are involved and how those features are removed?
iv. How the system improves the prediction accuracy?
v. How to improve the diagnosis of lung cancer?

The above-listed questions are a few sample questions regarding the lung cancer prediction process. The listed items can interact with the earlier detected system process to enhance the total lung cancer prediction process. Even though the patient does not have intense symptoms, the developed system successfully analyses their CT scan images, predicting cancer with minimum complexity. The fastest computation of cancer increases people's survival rate because it improves treatment procedures. Likewise, the created automatic cancer detection system successfully resolved the above study issues.

1.6 Research objectives

This study focuses on developing a diagnostic methodology for lung cancer detection utilizing image processing and soft computing methods. This study aims to use image