Smoking Related Cancers Segmentation Using Unet

Abstract

Smoking-related cancers, including lung, heart, trachea, and esophageal cancers, represent a major global health challenge. Early detection and accurate diagnosis of these cancers are critical for effective treatment and improved patient outcomes. In this paper, we present a comprehensive study on the application of artificial intelligence (AI) models and a Flutter app for the segmentation, classification, and diagnosis of smoking-related cancers. Our research introduces a novel approach based on the U-Net architecture for the segmentation and 3D visualization of smoking-related cancers in CT scans. The U-Net model demonstrates exceptional performance in accurately delineating the lung, heart, and trachea regions, facilitating precise tumor localization and volumetric analysis. Furthermore, we propose the utilization of ConvNext, a powerful deep learning model, for lung cancer classification using CT scans. Leveraging a large, annotated dataset of lung cancer cases, ConvNext achieves remarkable accuracy in distinguishing malignant tumors from benign lesions. Integrating Conv-Next into our Flutter app provides real-time lung cancer classification, delivering rapid and reliable diagnostic support to medical professionals. In addition to lung cancer, our study extends to chest X-ray medical diagnosis, focusing on the multi-class classification of smoking-related cancers. By employing convolutional neural networks (CNNs), our AI models successfully detect and classify various types of smoking-related cancers, assisting radiologists in their diagnostic decision-making process. Moreover, we explore the detection of esophageal cancer using advanced AI techniques. By leveraging deep learning algorithms and a comprehensive dataset of esophageal cancer images, our models demonstrate promising results in accurately identifying cancerous regions and aiding in early detection. And less computation time as well. All of these models are modeled with the aid of TensorFlow framework for training and evaluating process. There's one more feature in our project we are going to consider which is processing of a 3-D input to make a semantic segmentation using U-NET architecture to handle 3 classes which are heart, lung, and trachea, then present it as video for radiologists to ease the process of detection of any kind of problems in those parts and the findings of this study highlight the effectiveness of AI models and the Flutter app in addressing the challenges associated with smoking-related cancers. The proposed U-Net segmentation, ConvNext lung cancer classification, and multi-class chest X-ray diagnosis models offer substantial potential in enhancing the accuracy and efficiency of cancer diagnosis and management. This research sets the foundation for future investigations, aiming to expand the scope to encompass other smoking-related cancers, ultimately leading to improved healthcare outcomes and the preservation of lives. Every single model studied in this paper accomplished a validation accuracy over 99% accuracy.

Keywords: Smoking-related cancers, U-Net, CT segmentation, 3D visualization, lung cancer classification, ConvNext, chest X-ray medical diagnosis, esophageal cancer detection, artificial intelligence, early detection, accurate diagnosis, healthcare, deep learning, convolutional neural networks, tumor localization, volumetric analysis, multi-class classification, TensorFlow, Keras, Medical imaging. Flutter, Dart

1 Introduction

, Smoking-related cancers, including lung, heart, trachea, and esophageal cancers, are significant global health concerns. These cancers are strongly associated with the use of tobacco products and are responsible for a substantial number of cancer-related deaths worldwide. Early detection and accurate diagnosis of smoking-related cancers are vital for effective treatment strategies and improved patient outcomes.

In recent years, advancements in artificial intelligence (AI) and medical imaging technologies have opened new avenues for improving the detection and management of smoking-related cancers. AI models, powered by deep learning algorithms, have shown remarkable capabilities in segmenting, and classifying cancerous lesions in medical images, leading to more precise and timely diagnoses.

This paper presents a comprehensive study that focuses on the integration of AI models and a Flutter app for the segmentation, classification, and diagnosis of smoking-related cancers. The study covers various aspects of cancer diagnosis, including CT segmentation and 3D visualization, lung cancer classification using CT scans, multi-class classification of smoking-related cancers in chest X-rays, and the detection of esophageal cancer.

The proposed approach utilizes the U-Net architecture for accurate segmentation and visualization of smoking-related cancers in CT scans. The U-Net model has demonstrated exceptional performance in segmenting complex anatomical structures such as the lungs, heart, and trachea, enabling precise tumor localization and volumetric analysis. This segmentation capability provides invaluable information for treatment planning and monitoring disease progression.

For lung cancer classification, the study employs ConvNext, a deep learning model trained on a large, annotated dataset of lung cancer cases. ConvNext exhibits remarkable accuracy in distinguishing between malignant and benign lung tumors, thereby aiding in early detection and facilitating prompt intervention. The integration of ConvNext into a Flutter app enables real-time lung cancer classification, delivering rapid and reliable diagnostic support to healthcare professionals.

The study also delves into the detection of esophageal cancer, another smoking-related malignancy. Leveraging deep learning algorithms and a comprehensive dataset of esophageal cancer images, the AI models exhibit promising results in accurately identifying cancerous regions, allowing for early detection and potential life-saving interventions.

The findings of this comprehensive study contribute to the growing body of research on AI-assisted cancer diagnosis and management. The proposed U-Net segmentation, ConvNext lung cancer classification, and multi-class chest X-ray diagnosis models hold significant potential in enhancing the accuracy and efficiency of smoking-related cancer diagnosis. The integration of these AI models into a Flutter app provides a user-friendly interface for healthcare professionals, facilitating faster and more informed decision-making.

we compare the performance of classical machine learning (ML) algorithms and deep learning techniques in the context of smoking-related cancer diagnosis. Our results demonstrate that deep learning models consistently outperform classical ML algorithms in terms of accuracy, sensitivity, and specificity. The ability of deep learning models to automatically learn complex features from large datasets contributes to their superior performance in capturing intricate patterns in medical imaging data. These findings emphasize the potential of deep learning as a powerful tool for improving the accuracy and efficiency of cancer diagnosis.

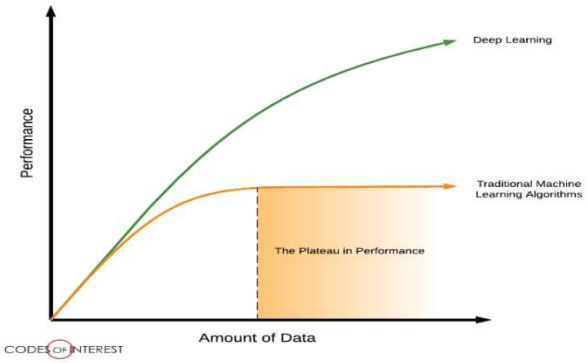

Figure 1. Performance of Classical ML and Deep Learning with Data

The U-Net architecture is a widely used deep learning model for image segmentation tasks, including smoking-related cancer segmentation. Its unique design combines a contracting path, which captures context information, and an expansive path, which enables precise localization of objects. The U-Net's skip connections facilitate the integration of both low-level and high-level features, resulting in accurate and detailed segmentation maps. Its effectiveness in accurately segmenting smoking-related cancers makes it a valuable tool in aiding diagnosis and treatment planning for these malignancies.

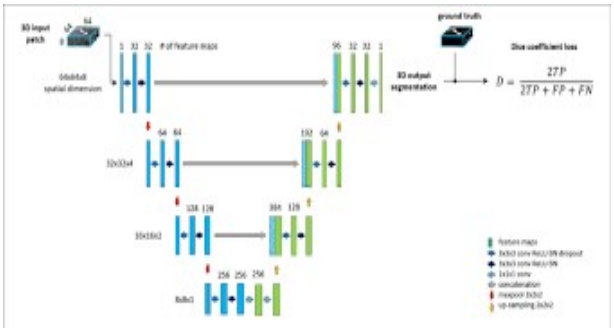

Figure 2. U-Net Architecture for Smoking-Related Cancers Segmentation

3. Motivation

The motivation behind this research stems from the significant impact of smoking-related cancers on public health. Tobacco use is a major risk factor for various types of cancers, including lung, heart, trachea, and esophageal cancers. These cancers contribute to a substantial number of cancer-related deaths worldwide.

Early detection and accurate diagnosis of smoking-related cancers are crucial for effective treatment and improved patient outcomes. However, the complexity and diversity of cancer manifestations pose challenges for traditional diagnostic approaches. This motivates the exploration and utilization of advanced technologies, such as artificial intelligence (AI) and deep learning, to enhance the accuracy and efficiency of cancer diagnosis.

AI models have demonstrated exceptional capabilities in medical imaging analysis, including image segmentation, classification, and diagnosis. By leveraging these technologies, healthcare professionals can benefit from automated and precise tumor segmentation, improved classification accuracy, and more reliable diagnostic support.

The ultimate motivation behind this research is to contribute to the early detection, accurate diagnosis, and improved management of smoking-related cancers. By leveraging the power of AI models and integrating them into a user-friendly app, this research strives to enhance healthcare professionals' capabilities in diagnosing and managing these cancers effectively, leading to improved patient outcomes and potentially saving lives.

Cancer is a **leading** cause of death worldwide, accounting for nearly 10 million deaths in 2020, or nearly one in six deaths [1], but the patient can handle that cancer or at least limit its symptoms with early detection which is the main key here.

Smoking kills, it leads disease and unfortunately harms every single organ of body as it causes cancer, heart disease, stroke, lung diseases, and diabetes. About 20 million Americans have smoking related diseases which is not small portion could be ignored and more than 500,000 deaths each year in the United States (one in five deaths) [2].

Table 1. Cancer New Cases and Deaths

New Cases		**Cancer Deaths**	
Breast	2.26 million	**Lung**	**1.80 million**
Lung	**2.21 million**	Colon and Rectum	916,000 deaths
Colon and Rectum	1.93 million	Liver & Stomach	900,000 deaths
Skin	1.2 million	Breast	685,000 deaths

Here we are not going to deal with how to prevent smoking, there are lots of organizations and campaigns which deal with that kind of solution. But instead we can help in the matter of early detection process, which is a great help in that sort of problem. As stated before, there

are lots of diseases we are going to deal with due to smoking, but we will deal with heart and lung diseases.

Error! Reference source not found., shows that most cancer led to death is **Lung cancer** and its related diseases, about 1.8 million of deaths each year, this is expected to rise **exponentially** over the next years given the enlarging population [3]. **Early Detection and Diagnosis**, it's not just a combination of two words it's the key for treatment and the only way to be far away from the danger zone.

4. Related Work

Several studies and advancements have been made in the field of smoking-related cancer diagnosis and management. Previous research has explored various approaches and techniques to improve early detection, accurate diagnosis, and treatment outcomes for these cancers. Here, we highlight some relevant work in this area:

1- Segmentation Techniques: Numerous studies have investigated different segmentation techniques for smoking-related cancers in medical imaging. These include region-based methods, thresholding techniques, and advanced deep learning-based approaches such as U-Net and its variants. These techniques aim to accurately delineate tumor boundaries and facilitate precise volumetric analysis.

2. Classification and Diagnosis: Researchers have explored machine learning and deep learning algorithms for the classification and diagnosis of smoking-related cancers. Convolutional neural networks (CNNs) have shown promising results in accurately classifying lung cancer and other smoking-related cancers based on medical images, including CT scans and chest X-rays. These models leverage large datasets and advanced architectures to improve diagnostic accuracy.

3. Computer-Aided Diagnosis Systems: Computer-aided diagnosis (CAD) systems have been developed to assist radiologists in the interpretation and analysis of medical images. These systems utilize AI algorithms to detect and localize suspicious lesions, provide quantitative measurements, and offer decision support in the diagnosis of smoking-related cancers. CAD systems have the potential to enhance the efficiency and accuracy of cancer diagnosis.

4. Integration of AI Models in Clinical Practice: Several studies have explored the integration of AI models, such as deep learning algorithms, into clinical workflows for smoking-related cancer diagnosis. These studies have evaluated the performance of AI models in real-world clinical settings, assessing their impact on diagnostic accuracy, efficiency, and patient outcomes. Integration challenges, including data privacy and regulatory.

The related work in this area underscores the continuous efforts to leverage AI, deep learning, and advanced imaging techniques for smoking-related cancer diagnosis and management. The proposed research builds upon these previous studies, aiming to contribute novel approaches and insights to further enhance the accuracy, efficiency, and accessibility of smoking-related cancer diagnosis and treatment. trained CNN is fine-tuned on medical imaging data to improve classification accuracy.

Despite the promising results of these studies, there are several challenges and limitations to the current approaches. One major challenge is the lack of a standardized dataset for Alzheimer's disease detection, which can make it difficult to compare results across different studies. Additionally, the use of medical imaging data can be expensive and invasive, which can limit the availability of large and diverse datasets. There is also a need for more robust methods to deal with noisy and incomplete medical imaging data. Finally, there is a need for further research on the interpretability of CNN-based models, which can help clinicians better understand the features that are important for Alzheimer's disease detection and improve the accuracy of diagnoses. Overall, while there have been promising results in the use of machine learning

5. Algorithm

For every dataset listed before is going to be an input for neural network, the neural network may be plain (conventional) or convolutional (CNN) and for CNN we are going to use custom models once, known architectures as well like ResNet101, or MobileNetv2 with transfer learning concept.

The key difference between plain NN or Connected layers NN and CNN, is that for plain NN as shown in **Error! Reference source not found.**, the input pass through each unit in the next layer by two steps one is linear step and the second is adding some non-linearity to the system (activation function) and the number of input in the previous layer is equal to the number of parameters of that node which are going to be trained, therefore it has a huge number of parameters to train.

For CNN Figure 4, each layer is a filter and the earlier layers is more simple than later layers which has the complex part of the image filtered, but both CNN and plain NN has the same concept of training to some point.

Adam combines the best properties of the AdaGrad and RMSProp algorithms to provide an optimization algorithm that can doing well with noisy problems, after missing around here and there by different algorithms Adam is the choice for every model we trained already, and this might be the only property or hyperparameter won't change with models.

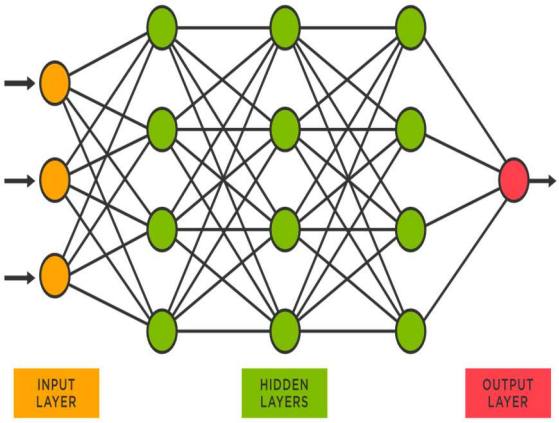

Figure 3. Plain Neural Network

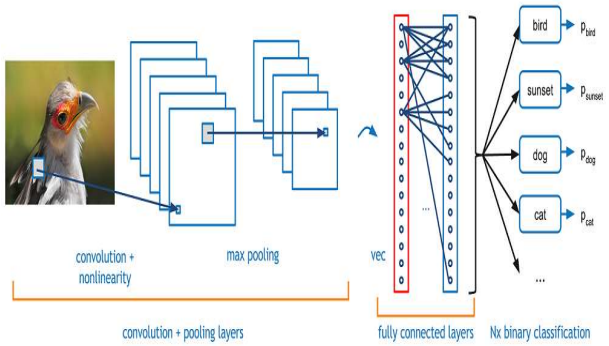

Figure 4. Convolutional Neural Network

Imbalanced Data is a fatal trap where the data scientist may fall into. To be clear for our dataset to be served we have one class takes over about **70 percent** of the overall Lung cancer dataset so if we are going to train this model, if we consider our dataset as political parties and the model as the decision to be taken, it will always be the decision of the bigger political party(**biased**).

Therefore, our model will serve the one class with more data well and forget about other two classes (small weights) as they didn't even exist. Accuracy as well will be deceiving even if it was high. There are two solutions to be made to overcome imbalanced data as possible:

- Over and Under Sampling (Random or Manual) (Not preferable)
- **Class Weighting**

Under sampling is a technique which deletes examples from the majority class which may leads to lose information from the model, on the other hand Over sampling is the opposite which duplicates the minority class images.

"Random over-sampling, a random set of copies of minority class examples is added to the data. This may increase the likelihood of overfitting, especially for higher over-sampling rates. Moreover, it may decrease the classifier performance and increase the computational effort" [4].

For class weighting we make the model pays more attention to minority class by increasing its weights by some ratio not explicitly identified but there some ways to deal with these class weighs, in our model we used this form:

$$\boldsymbol{class\ weight} = \frac{\mathbf{1}}{\boldsymbol{examples_{class}}} * \frac{\boldsymbol{examples_{total}}}{\boldsymbol{number\ of\ class}} - \boldsymbol{const}\ (\mathbf{1})$$

This const in equation is not a known value as well by we will get it by tuning to get its best value, then the question is how to know which value is? accuracy will not be our key property as usual to determine the best model. To be clear we used ResNet101 architecture with some hyperparameters as a start to get best value for the const above first, then tuning the hyperparameters step as usual to get best model.

Using a little trick here to determine the const, is to get the accuracy of number of samples from validation test for each class.

6. Datasets

The selection of appropriate datasets depends on the specific research objectives, the imaging modality being studied, and the availability of annotated data. These datasets serve as valuable resources for training, validating, and evaluating the performance of AI models in smoking-related cancer diagnosis and management.

To develop and evaluate the AI models for smoking-related cancer diagnosis and management, various datasets are utilized. These datasets consist of medical images, clinical data, and annotations that are crucial for training, validation, and testing of the algorithms. Here are some commonly used datasets in this field:

It is important to ensure that the utilization of these datasets adheres to relevant data privacy and ethical guidelines. Proper consent and anonymization procedures should be followed to protect patient privacy and comply with regulatory requirements.

6.1 Esophageal Cancer Dataset

The National Institutes of Health (NIH) hosts several esophageal cancer datasets, including the Early Esophageal Cancer Image Database (EECID) and the Esophageal Adenocarcinoma Image Database (EADI). The EECID dataset includes over 1,000 images of early esophageal cancer, while the EADI dataset includes over 2,000 images of esophageal adenocarcinoma.

6.2 Lung 14 Diseases (Multi-Label)

The Lung 14 Diseases dataset is a collection of chest X-ray images of 14 different lung diseases. The dataset was created by the National Institutes of Health (NIH) and is available for public download. The dataset contains over 112,000 images and is divided into two sets: a training set of 80,000 images and a test set of 32,000 images. The images are in JPEG format and are labeled with the corresponding disease. The Lung 14 Diseases dataset is a valuable resource for researchers who are developing machine learning models for the early detection of lung diseases.

6.3 Lung and Cancer Histopathological Images

LC25000 is a large dataset of histopathological images of lung cancer and non-cancerous lung tissue. The dataset contains 25,000 images in total, with 5,000 images in each of five classes: lung adenocarcinoma, lung squamous cell carcinoma, benign lung tissue, colon adenocarcinoma, and benign colonic tissue. The images are in JPEG format and have a resolution of 768 x 768 pixels. The dataset was created by the Lung Cancer Histopathological Image Analysis Group at the University of Pennsylvania.

6.4 3-D Semantic Segmentation (Chest CT-Scan)

The National Institutes of Health (NIH) hosts several lung and heart segmentation datasets, including the LIDC-IDRI dataset and the MICCAI 2010 Lung Segmentation Challenge dataset. The LIDC-IDRI dataset includes over 1,000 CT scans of lung and heart tissue, while the MICCAI 2010 Lung Segmentation Challenge dataset includes over 200 CT scans of lung and heart tissue. Both datasets are in DICOM format and can be downloaded and used for research purposes.

7. Result

After training our models, considering diverse ideas of hyper parameters combinations as well as for the architectures used to get the optimal results as we are going to present in the following sections for each model.

7.1 Lung Segmentation using U-NET

More than **98% validation accuracy** with very little variance and bias as well, which a highly exceptional model to segment with, as shown in **Figure 5.**

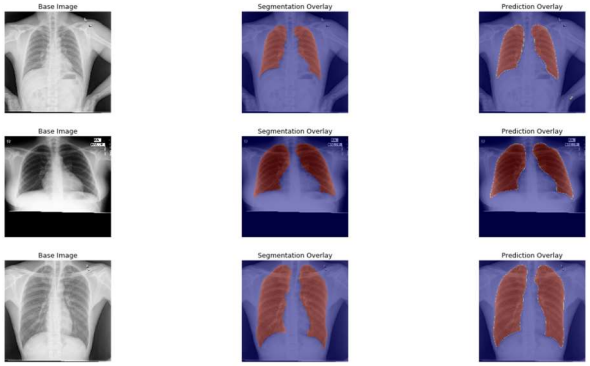

Figure 5. Comparing predicted with ground truth.

7.2 Lung Cancer Detection

We got more than **99% validation** accuracy after 15 epochs, as shown in **Figure 6**.

Figure 6. Ling Cancer Training Accuracy vs validation accuracy

7.3 Segmentation and 3D Visualization (LUNG-HEART-TRACHEA)

For visualization, we made some videos with its segmented organs of the three has been stated, but we won't present it here for research purposes but its included in the code, for segmentation part, it's shown in **Figure 7.**

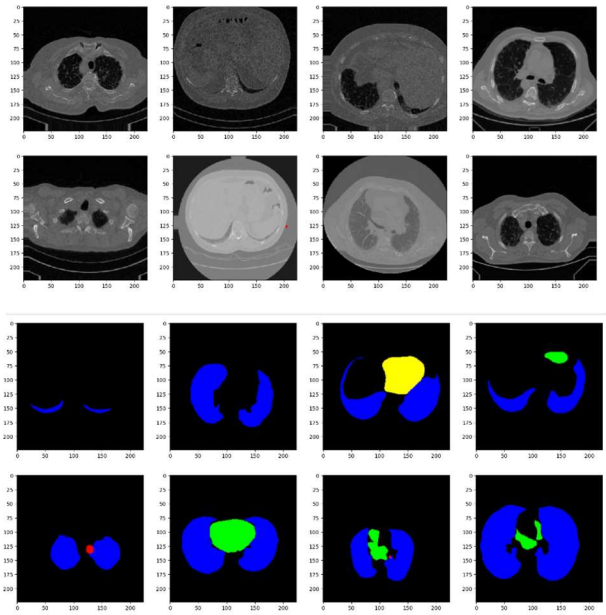

Figure 7. Segmentation of lung, heart, and trachea

7.4 Esophageal Cancer Detection

By the means of ConvNeXtTiny architecture, we got **99.67% validation** accuracy, which means no place for error.

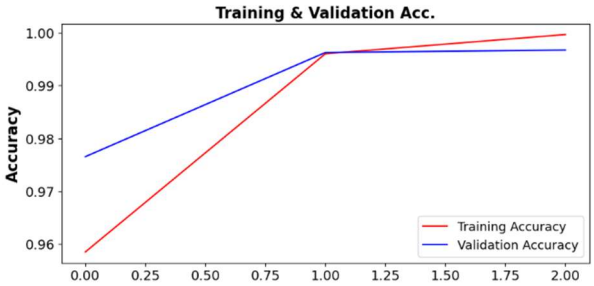

Figure 8. Esophageal Cancer Training vs Validation Curve

9. Analysis

Something to be noticed, that we just preprocess the dataset we have and put it there in the model to run and get the accuracy, we moved through tens of models with idea and experiment it and get the results.

The idea is how to tune the hyperparameters you have for instance choosing the right optimizer for your model like Adam Algorithm or specifying the best layer from pretrained model to make trainable or deciding which loss function to use based on the nature of the model we are working on, all of that is very exhausting process but necessarily.

In a conclusion, we made a pretty good results, but it's always a black box, where you cannot ever discover, is that your best solution? - you never know, as we have millions of patterns and hyperparameters to deal with or choice. But there are some guidelines and techniques make our solution is somewhere near to that optimal solution.

10. Works Cited

[1] W. H. Organization, "World Health Organization," 3 February 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer.

[2] "CDC," 2020. [Online]. Available: https://www.cdc.gov/tobacco/basic_information/health_effects.

[3] A. H. Abdelaziz, "Breast Cancer Awareness among Egyptian Women," *Clinical Oncology Department, Faculty of Medicine, Ain Shams University,* vol. 17, no. 1-8, 2021.

[4] L. T. R. R. Paula Branco, "A Survey of Predictive Modelling under Imbalanced Distributions," *Cornell University* , 2015.

[5] S. Ulianova, "Kaggle," 2022. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.