Tunisian Lung Cancer Dataset: Collection, Annotation and Validation with Transfer Learning

Omar Khouadja[1], Mohamed Saber Naceur[2], Samira Mhamedi[3], Anis Baffoun[4]
Laboratoire De Télédétection Et Systèmes d'Information à Référence Spatiale (LTSIRS)[1, 2]
Institut National Des Sciences Appliquées Et De Technologie (INSAT), Université De Carthage, Tunis, Tunisia[1, 2]
Hôpital militaire principal d'instruction de Tunis, Tunisia[3, 4]

***Abstract*—Globally, lung cancer remains the leading cause of cancer-related deaths, with early detection significantly improving survival rates. Developing robust machine learning models for early detection necessitates access to high-quality, localized datasets. This project establishes the first lung cancer dataset in Tunisia, utilizing DICOM CT scans from 123 Tunisian patients. The dataset, annotated by experienced radiologists, includes diverse forms of lung cancer at various stages. Using transfer learning with pre-trained 3D ResNet models from Tencent's MedicalNet, our tests showed the dataset outperformed previous models in specificity and sensitivity. This demonstrates its effectiveness in capturing the unique clinical characteristics of the Tunisian population and its potential to significantly enhance lung cancer diagnosis and detection.**

***Keywords*—Lung cancer; Tunisia; dataset; transfer learning; medical imaging; annotations**

I. Introduction

The lungs are the main organs of respiration. They regulate breathing, ensuring that each and every cell in the body receives oxygen [1]. The human body's specifically designed defense mechanisms shield the organs. However, they are unable to completely eliminate the risk of contracting specific lung diseases. Infections, inflammation, or even more serious conditions like the emergence of a malignant tumor can affect the lungs. One of the main causes of death in industrialized countries is lung cancer. Toxic surroundings, long-term inflammation, and smoking are only a few of the variables that frequently cause long-term harm. Phlegm is one of many strategies that the lungs use to clean their airways on their own. However, this is not enough for a smoker [2].

The latest advancements in imaging and sequencing technology have resulted in tremendous progress in the clinical investigation of lung cancer. However, the human mind's ability to comprehend and utilize the collection of such enormous amounts of data is limited. Machine learning-based techniques enable the integration and analysis of these large and complex datasets, which have extensively described lung cancer through the application of diverse viewpoints from these acquired data [3].

Developing precise and trustworthy diagnostic tools, especially in the areas of cancer detection and medical imaging, requires a rich dataset. Researchers can train sophisticated machine and deep learning models that can effectively generalize to a wide range of populations by using comprehensive datasets that include a wide array of patient demographics, imaging modalities, and annotated examples. These datasets are essential for enhancing individualized treatment strategies, early diagnosis, and early detection of diseases like lung cancer [4].

Nevertheless, there is a dearth of such extensive medical records in Tunisia. The creation of specialized diagnostic instruments that can cater to the unique requirements and traits of the community is hampered by this scarcity. Due to variations in genetics, environment, and demography, diagnostic models trained on data from other locations could not perform as well in the absence of localized datasets. To close this gap and improve the precision and efficacy of lung cancer diagnosis and treatment in the area, high-quality, annotated medical datasets must be created and shared immediately in Tunisia.

This study intends to overcome these shortcomings and offer a useful resource that can aid in the development of AI-driven diagnostic tools customized for the Tunisian population by producing the first dataset from Tunisia for intelligent lung cancer detection. Such initiatives are necessary to ensure that medical technology improvements benefit all regions equally and to improve healthcare outcomes.

We start this article with a definition of lung cancer where we present the lung anatomy and explain the origin of the disease, its types, and stages. Next, Section III describes lung cancer detection and diagnoses using imaging techniques. Moving on to Section IV, highlights the importance and impact of data in cancer detection, introducing the challenges we face in finding Tunisian datasets for regional analysis. Section V describes related work, positioning our research in the context of other studies and highlighting how our approach differs and contributes to the field. Then, the process of building our Tunisian lung cancer dataset is described. We go over how we found and collected the data, how the images were prepared and annotated, and the stringent quality control procedures put in place to guarantee data integrity. In Section VII, we present the model used to validate our created dataset, the results are then compared with literature models and discussed in Section VIII. Finally, the paper is concluded in Section IX.

II. Lung Cancer Definition

The lungs are two sponge-like organs inside the chest. Three lobes, or parts, make up the right lung. Two lobes make up the left lung. It is smaller on that body side because the

heart occupies more space there. When we inhale, air enters our nose or mouth and goes to our lungs via the trachea (windpipe). The trachea divides into bronchi, which enter the lungs and divide further into smaller bronchi. Bronchioles are tiny branches that divide from them. There at the tip of the bronchioles are small sacs of air known as alveoli. When we breathe air, the alveoli transport oxygen in the blood and expel carbon dioxide. Our lungs' primary functions are to take in oxygen and expel carbon dioxide. Lung cancers typically develop in the cells that make up the bronchi and other parts of the lung, like the alveoli or the bronchioles. The pleura is a thin layer of membrane that surrounds the lungs. As the lungs expand and contract during breathing, the pleura shields them and aids in their sliding back and forth against the chest wall. A narrow, dome-shaped muscle known as the diaphragm, separates the chest from the belly beneath the lungs. As we breathe, the organ contracts and expands, propelling air into and out of the lungs [5]. Cancer arises when the body's cells begin to proliferate uncontrolled. When it's in the lungs, we talk about Lung Cancer. For both sexes, lung cancer is one of the most common cancer-related causes of death [6].

The most common indicator of this type of cancer is coughing, which needs to be treated carefully because most lung cancer patients also have chronic obstructive pulmonary disease, which can cause coughing on its own. More importantly, the cough's characteristics change—becoming more intense, persistent, and possibly accompanied by expectoration or bloody sputum. Lung cancer also manifests as expectoration, chest pain, shortness of breath, anorexia, fever, hemoptysis, and weight loss [7].

A pulmonary nodule, often known as an abnormal growth, forms in the lung. Respiratory problems and infections can lead to the development of nodules in the lungs. Most lung nodules are not indicative of lung cancer and do not require medical attention. On X-rays or scans, these growths could show up as a shadow or spot on the lung. One or more nodules may form in one lung or more in both [2].

A. Lung Cancer Types

Two primary forms of lung cancer exist [10]:

Non-small cell lung cancer (NSCLC): The most common type, approximately 80–85% of instances of lung cancer are caused by non-small cell lung cancer (NSCLC). Adenocarcinoma, Squamous cell carcinoma and Giant cell carcinoma are the three main types of non-small cell lung cancer.

- The most prevalent subtype of NSCLC, Adenocarcinoma is typically present in the lung's outer regions. It affects women and non-smokers more frequently.
- Squamous Cell Carcinoma is frequently associated with smoking, typically begins in the middle of the lungs, close to a bronchus.
- Large Cell Carcinoma is a rarer variety that can develop anywhere in the lung and has a rapid growth and dissemination rate.

Small cell lung cancer (SCLC): It is less common and more aggressive than NSCLC, this type of lung cancer accounts for 10% to 15% of all cases of lung cancer. It's also known as oat cell cancer at times. Compared to NSCLC, this type of lung cancer develops and spreads faster. In most patients, the cancer has already exited the lungs when they are diagnosed with SCLC. Because it spreads quickly, this cancer usually responds well to chemotherapy and radiation treatments. Unfortunately, most patients will experience recurrent cancer. SCLC is heavily associated with smoking and it has two main subtypes which are:

- Small Cell Carcinoma: Sometimes referred to as Oat cell cancer, is the most aggressive type and frequently spreads to other body areas.
- Combined Small Cell Carcinoma: It consists of both non-small cell and small cell cancer.

B. Lung Cancer Stages

Comprehending the distinct forms of lung cancer is imperative in order to comprehend the progression of each type through its varied phases. Depending on the severity of the disease, each kind of lung cancer—small cell lung cancer (SCLC) or non-small cell lung cancer (NSCLC)—follows a different course of development and dissemination and is divided into phases.

1) NSCLC stages: NSCLC develops in a number of stages, each of which indicates how far the disease has spread. The tumors' stage is determined by their size and whether or not they have spread to adjacent lymph nodes or other organs [9]:

a) First stage: A 5 mm diameter tumor was discovered; it has not spread to any organs or lymph nodes. Usually, these tumors can be removed surgically.

b) Second stage: The tumor has grown to neighboring lymph nodes and is no more than 7 mm across. As an alternative, there can be more than one distinct tumor nodule visible. These tumors can usually be surgically removed.

c) Third stage: Any size tumor is possible, and it has spread to the lymph nodes. It might have also extended to nearby regions. It is possible for a single lung to have two or more tumors in separate lobes. At this point, it is not possible to remove the tumors.

d) Fourth stage: Characterized by pleural effusion or metastasis (spread) to other body parts. Any size lung tumor has progressed to the fluid surrounding the lungs, lymph nodes, and other distant organs.

Fig. 1 shows the different NSCLC stages explained.

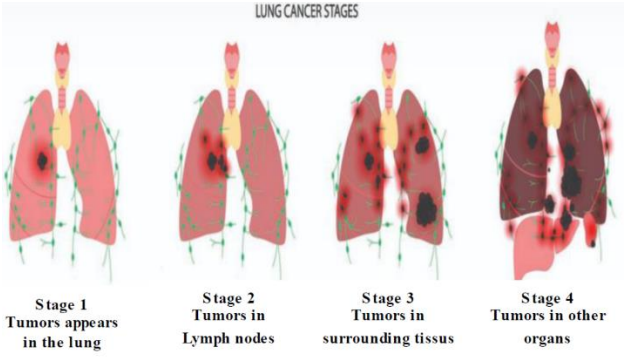

Fig. 1. NSCLC stages [8].

Fig. 2 presents the two SCLC stages.

2) SCLC stages: Because SCLC is aggressive in character, it splits into two primary stages [8]:

a) Limited stage: One radiation field can be used to treat cancer that is limited to one side of the chest, including just one lung and adjacent lymph nodes.

b) Extensive stage: The cancer has progressed to distant organs or other areas of the chest. Because SCLC progresses quickly, the majority of cases are diagnosed at this point.

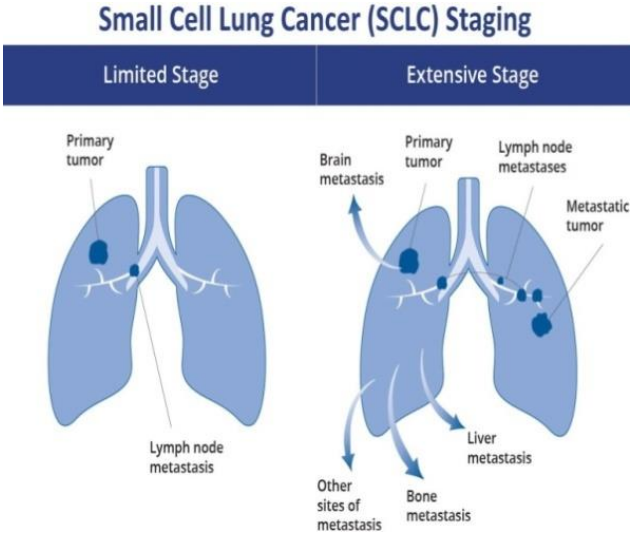

Fig. 2. SCLC stages [11].

Understanding the stage of your lung cancer is essential to determining course of therapy. Even though advanced-stage cancer might potentially prolong a person's life, earlier-stage tumors are usually easier to treat.

III. Lung Cancer Diagnosis and Imaging

A general practitioner (GP) will talk about the patient's overall health and symptoms in order to identify lung cancer. If the patient's physical examination and history suggest that they may have lung cancer, more testing will be done. Imaging studies may be one of them. Imaging tests produce pictures of the internal organs. There are several reasons to undergo imaging tests, both before and after being diagnosed with lung cancer [12], such as: investigating suspicious or possibly malignant areas; estimating the extent to which cancer may have spread; evaluating the efficacy of the treatment; and looking for any signs that the disease might recur following treatment.

When it comes to lung cancer detection, staging, and treatment, imaging is essential. Comprehensive information regarding lung tumors' existence, size, location, and extent—as well as their potential to spread to other body parts—can be obtained using a variety of imaging methods. Chest X-rays, computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI) are the main imaging modalities used in lung cancer[13]:

When lung cancer is suspected, X-rays of the chest are frequently the first imaging tests carried out. Large tumors and notable anomalies may be seen, but smaller or less noticeable lesions may go unnoticed. The majority of lung cancers appear as a white-gray mass on X-rays like shown in Fig. 3.

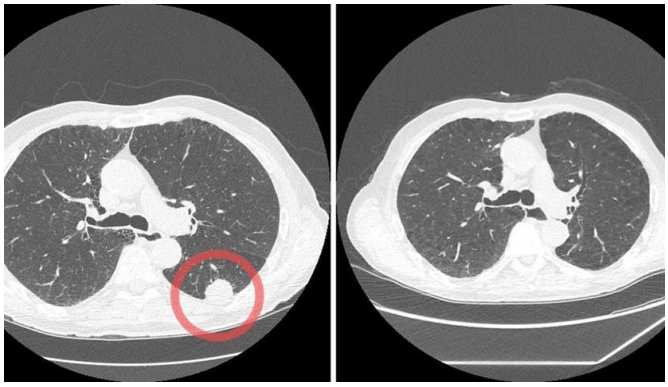

Fig. 3. Lung X-ray image [15].

More precise cross-sectional images of the lungs and other chest tissues are provided by CT scans, which aid in the detection of smaller tumors as well as the localization and size of malignancy. It is common practice to perform a CT scan after a chest X-ray. A CT scan uses X-rays and a computer to create detailed images of the inside of the body. It creates complex images of the body in cross-section. A CT scan gathers many images, as opposed to a typical X-ray, which only captures one or two. These images are then combined by a computer to create a slice of the body portion under study [14]. The Fig. 4 shows an example of a lung cancer CT scan.

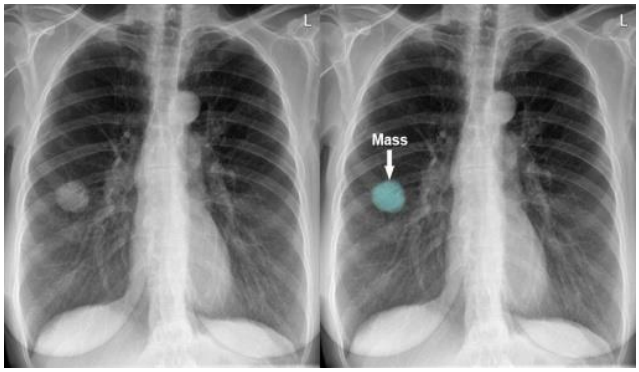

Fig. 4. Lung CT scan image [16].

Usually, PET scans are performed to find metastases and evaluate the metabolic activity of lung cancer cells. A tiny quantity of radioactive glucose is injected using this approach, and because cancer cells have a greater metabolic rate than other cells, they absorb the glucose. In order to improve diagnostic precision, PET scans are frequently coupled with CT scans (PET/CT). The Fig. 5 shows an example of a lung cancer PET scan.

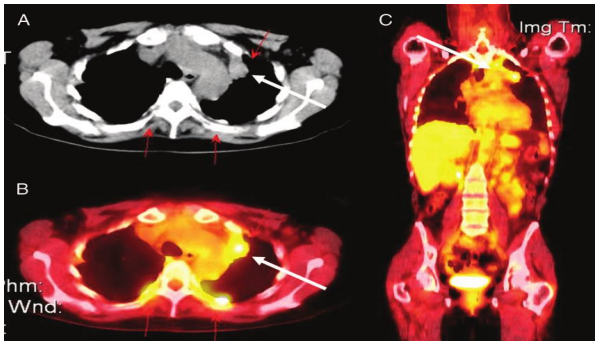

Fig. 5. Lung PET image [17].

While less frequently used for lung cancer, MRI is especially helpful for analyzing tumors close to important blood vessels and determining whether cancer has progressed to the brain or spinal cord. Similar to CT scans, MRIs produce finely detailed images of the body's soft tissues. However, MRI scans employ strong magnets and radio waves in place of X-rays. The most prevalent use for MRI scans includes the detection of possible brain or spinal cord metastases from lung cancer.

Out of all the approaches outlined, the CT image technique is the most widely used since it may give a view without showing structures that overlap. It might be difficult for physicians to diagnose and interpret cancer. The use of CT imaging allows for the accurate diagnosis of lung cancer [14].

On imaging studies, lung cancer can appear as a single microscopically small nodule, ground-glass opacity, lung collapse, pleural effusion, numerous nodules, or multiple opacities. Simple and tiny lesions are extremely hard to locate. Due to their late diagnosis, lung cancer patients usually have a poor prognosis. Due to the unpredictability of imaging results, and histology, it is challenging for doctors to choose the best course of treatment for lung cancer [1]. Because so many images need to be analyzed, radiologists must rely largely on their years of expertise to spot anomalies. Even highly qualified individuals may overlook tiny indications of cancer. The process is made more difficult by the variety in tumor appearance, which includes variations in size, shape, and density. Tumors can be hidden by overlapping bodily structures, making it challenging to identify them. Furthermore, determining the difference between benign and malignant lesions necessitates meticulous examination, which is laborious and prone to human mistakes. Patient outcomes may be impacted by missed diagnoses or false positives due to human error and fatigue.

These constraints can be overcome by machine learning models, especially deep learning algorithms. These algorithms are able to understand and identify complicated patterns linked to different types of malignancies because they have been trained on largedatabases of annotated medical images.

Using these models contributes to early tumor diagnosis, which is essential for bettering patient outcomes. Large volumes of imaging data can be processed and analyzed swiftly by automated methods, which can deliver reliable results quickly. This improves overall diagnostic efficiency by relieving radiologists of some of their duty and freeing them up to concentrate on more complex patients. Additionally, as these models are exposed to additional data, they can get better over time, increasing their capacity to identify even the smallest and most subtle problems. Thus, the application of AI to medical imaging marks a substantial breakthrough in the early diagnosis and treatment of cancer, resulting in a quicker, more precise, and more easily accessible diagnostic procedure.

IV. Data Importance

The development and success of machine learning models, particularly in medical imaging and cancer diagnostics, depend heavily on rich and comprehensive datasets. The quality and comprehensiveness of the data directly affect the models' performance, accuracy, and reliability. Extensive datasets with a wide range of patient demographics, imaging modalities, and detailed annotations are crucial for capturing the entire spectrum of disease presentations and variations.

Machine learning models need to be trained on datasets that accurately represent the variety of real-world medical cases in order for them to identify and categorize cancers. This covers differences in imaging methods and instruments in addition to variances in tumor sizes, forms, locations, and stages. Rich datasets let the model understand intricate patterns and characteristics linked to various tumor forms, improving generalization and improving prediction accuracy across a range of patient populations.

A large dataset helps reduce the possibility of overfitting, in which a model performs well on training data but poorly on new, unseen data; by exposing the model to an extensive variety of examples, it learns to generalize well, improving its robustness and accuracy. This is especially important in medical imaging, where variation in patient anatomy and imaging conditions can be significant. A high-performing model requires a large amount of data to be trained on.

Reducing biases resulting from training models on small or homogeneous datasets is another benefit of having an extensive data set. The model's capacity to discriminate between benign and malignant lesions is enhanced when it is trained on a dataset that encompasses a wide variety of cases. This is especially crucial for early detection, as tiny irregularities could be readily missed in the absence of an extensive collection of training data.

Furthermore, thorough annotations from knowledgeable radiologists and oncologists improve learning by giving exact labels and classifications. The model uses these annotations as a vital source of information when it is being trained, which enables it to link particular imaging characteristics to related diagnostic categories. Extensive data

means that the model is exposed to a broad range of scenarios, increasing its robustness and ability to manage challenging cases in actual practice.

A. Localized Data Importance

Not only data is necessary for training machine learning models, but having a unique dataset for every location is also essential. Local data document the population's distinct demographic, genetic, and environmental features, all of which have a substantial impact on how diseases like lung cancer manifest, develop, and react to therapy. Diagnostic techniques and treatment plans might not work as well without localized data because they are frequently created using data from other areas with distinct demographic characteristics.

The access of such extensive medical databases is restricted in Tunisia for a number of reasons:

- First, many healthcare organizations lack the infrastructure and resources necessary for the systematic gathering and processing of data. This includes a lack of financing for the staff and equipment required to compile and manage huge datasets.
- Second, issues with privacy and regulations may make it difficult to integrate and share data throughout various medical facilities. The capacity to gather and exploit big, centralized datasets is frequently hampered by stringent data protection regulations and worries about patient confidentiality.
- Third, incomplete or inconsistent datasets are frequently the result of a lack of qualified individuals who can appropriately annotate and curate medical images.
- Fourth, different healthcare facilities lack uniform standards for data collection and interpretation. This discrepancy can result in inconsistent and fragmented data, which makes it challenging to assemble a coherent and extensive dataset.
- Sixth, the low acceptance and knowledge of electronic health records (EHRs) in many healthcare settings is another difficulty. A considerable portion of patient data is still in unstructured or paper formats in the absence of widespread use of EHRs, making it difficult to access for in-depth analysis and model training.
- Seventh, healthcare facilities sometimes face financial barriers that keep them from making the investments in the equipment and training needed for efficient data handling.
- Eighth, a large number of healthcare practitioners may also give priority to short-term clinical demands over long- term data gathering initiatives, which adds to the dearth of thorough datasets.
- On top of that, competitive tactics or a lack of incentives to exchange data might hinder collaboration between institutions, resulting in underutilized information silos.

B. The Impact of Limited Data

The creation and application of sophisticated diagnostic techniques in Tunisia are severely hampered by the absence of good datasets. Due to variations in demographics, genetics, and environmental factors, machine learning models trained on datasets from other regions would not function as well in the Tunisian setting without extensive local data. Poorer patient outcomes and less accurate diagnosis may result from this. Personalized medicine, which depends on comprehensive patient data to customize therapies to specific needs, is also constrained by the incapacity to use large amounts of data.

Improving Tunisia's healthcare will require addressing these data constraints. The creation and dissemination of superior annotated medical datasets would improve the precision and dependability of diagnostic models, resulting in improved identification and management of conditions like lung cancer. Tunisia can make sure that its healthcare system takes advantage of the advances in medical technology and offers its people equitable care by making investments in data infrastructure, hiring qualified staff, and creating frameworks for data sharing.

V. Related Work

Localizing populations in medical datasets is crucial for ensuring that diagnostic models are accurate and applicable to specific demographic groups. Many existing machine learning models for medical diagnosis are trained on datasets with a broad demographic range, which may not capture the unique characteristics of specific populations, such as those in Tunisia. This lack of localization can limit the effectiveness of these models in particular clinical settings. For instance, while the Lung-PET-CT-Dx dataset is extensive, it predominantly includes data from diverse regions and may not reflect the specific clinical characteristics seen in the Tunisian population.

Generalizing machine learning models across diverse populations is essential for robust performance. However, this generalization must be balanced with localization to ensure that models remain effective for specific demographic groups [33]. Localized datasets are tailored to capture the nuances of a particular population, leading to improved diagnostic accuracy within that group. Researchers stress the importance of creating more localized and representative datasets to address gaps in current research and ensure that models can accurately diagnose within specific populations [34].

A study titled "Optimizing double-layered convolutional neural networks for efficient lung cancer classification," published by BioMed Central, underscores the importance of localized datasets in training robust models. This research demonstrates that incorporating data from specific regions enhances a model's ability to accurately diagnose within those populations, thereby improving diagnostic accuracy and reliability. The authors found that models trained on localized datasets perform better in real-world scenarios, emphasizing the need for datasets like ours that focus on the Tunisian population [34].

Regarding Tunisian data, significant advancements have been made in understanding the epidemiological profile and risk factors specific to Tunisia. The study "Lung Cancer in Central Tunisia: Epidemiology and Clinicopathological Features" details the clinical and pathological characteristics of lung cancer cases in Central Tunisia over a 15-year period. It reveals that lung cancer is the most common cancer among Tunisian men, typically presenting at advanced stages, with squamous cell carcinoma being the most prevalent histological type in men and adenocarcinoma in women. These findings highlight the need for effective lung cancer control and prevention programs tailored to the Tunisian context [35].

The initial study is outdated, with its dataset created for a specific purpose that is now rather limited. Our new study, on the other hand, utilizes a more comprehensive and current dataset specifically designed for lung cancer detection, a crucial medical application. As a retrospective study covering the years 1993 to 2007, there may be biases from inaccurate or incomplete historical medical records. Additionally, the study does not consider several potential confounding variables that could affect lung cancer incidence and prognosis, such as genetic predispositions, environmental exposures, or socioeconomic factors. The outdated diagnostic equipment may not adequately capture the complexities of lung cancer progression and treatment response, making some findings less applicable to modern clinical practices. These limitations should be considered when interpreting the study's results and recommendations.

Another valuable resource is the RECIST PFS/OS lung cancer dataset, available on Mendeley Data. This dataset includes annotated CT scan images of lung cancer cases. The Salah Azaiez Institute in Tunisia provided data for creating a dataset that includes, for each patient, age, sex, treatment, presence of mass and nodules, censoring information, objective response, and survival time in days, using CT scans and reports from radiologists at the institute [36]. The dataset primarily uses the RECIST criteria to evaluate tumor response to treatment, which, while standardized, may not fully account for all the subtleties of tumor biology and patient outcomes. Additionally, as a retrospective and observational dataset, it may suffer from biases such as selection bias and information bias. Lastly, although it includes key variables such as sex, age, type of therapy, and survival times, it may lack other significant factors like genetic data, detailed treatment plans, and environmental exposures.

These datasets are limited by their lack of diversity and clinical settings that may not fully represent the unique characteristics of lung cancer in Tunisia.

Our study addresses this gap by creating the first Tunisian lung cancer dataset, which includes DICOM CT scans from 123 Tunisian individuals, annotated by experienced radiologists to cover various types of lung cancer at different stages. This comprehensive dataset ensures a more accurate representation of the Tunisian demographic, making it better suited for developing localized diagnostic tools. The aim of our dataset is to provide the necessary data to detect lung cancer accurately.

VI. Tunisian Lung Cancer Dataset

In order to meet the urgent demand for localized medical datasets in Tunisia, we sought the advice of experts at the esteemed "Military Hospital of Instruction of Tunis (HMPIT)" [31], an institution renowned for its competence in medical care and research. It is one of Africa's biggest and most prominent university hospitals. The Tunisian Ministry of National Defense is in charge of this medical center. The partnership with Military Hospital of Instruction of Tunis (HMPIT) played a pivotal role in procuring the superior CT scans required for our project.

We obtained DICOM-formatted CT scans from 123 individuals, where 80% or 98 persons have lung cancer and are treated at this hospital. For the purpose of compiling an extensive and representative dataset of lung cancer cases unique to the Tunisian population, these scans were essential. Because of the medical experts, the dataset was made more robust and applicable by included a variety of patients that represented different stages and forms of lung cancer.

Along with the radiologists and oncologists involved, the annotation process was carried out to guarantee the dataset's quality and dependability. Their knowledge was extremely helpful in precisely identifying and dividing up the lung nodules and other pertinent elements in the CT scans. This cooperative method helped the concerned teams create capacity and transmit expertise, in addition to improving the quality of the annotations.

The DICOM format CT scans of 98 lung cancer patients, encompassing adenocarcinoma, squamous cell carcinoma, and small cell lung cancer in all stages, are included in the collection from the Military Hospital of Instruction of Tunis (HMPIT) along with the CT scans of the other 25 healthy indivduals representing 20% of the dataset. Nodule counts, sizes, types, features, follow- up status, tumor volume, density, growth rates, involvement of lymph nodes, and documentation of metastases are all included in the annotations. The scans provide high-resolution images with an average of 350 slices per scan, and they are obtained utilizing advanced imaging modalities including Siemens SOMATOM Perspective and GE Healthcare Lightspeed VCT. Included are demographics, clinical data on symptoms, past medical histories, and results.

A. Scanners Used

Modern CT scanners commonly found in hospitals through- out Tunisia—the Siemens SOMATOM Perspective and the GE Healthcare Lightspeed VCT—were used to carefully generate the lung cancer dataset images. Our dataset contains high-quality and consistent data because these scanners were selected due to their extensive use in clinical settings, advanced imaging capabilities,and dependability.

1) Siemens SOMATOM perspective: The Siemens SO-MATOM Perspective is well known for its remarkable image quality and low radiation dosage, which makes it perfect for the in-depth imaging needed to diagnose lung cancer. With the use of cutting- edge technology like iterative reconstruction, this scanner greatly improves image clarity while lowering the

patient's radiation dose. By reducing distortions brought on by metal implants, the metal artifact reduction feature helps to improve diagnostic precision. Moreover, the scanner offers adaptability in identifying a variety of illnesses and supports a broad range of clinical applications.

However, like presented in Table I, the scanner does, have certain drawbacks, including high operating costs because of maintenance and operating costs, the requirement for thorough training for best use, potential problems with image quality because of patient movement (motion artifacts), and the lack of advanced features in older models.

TABLE I. SIEMENS SOMATOM PERSPECTIVE STRENGTHS AND WEAKNESSES

Siemens SOMATOM Perspective	
Strengths	**Weaknesses**
High image quality	High operational costs
Low radiation dose	Complex operation
Iterative reconstruction	Susceptibility to movement artifacts
Metal artifact reduction	Limited advanced features in older models
Versatile clinical applications	

2) GE Healthcare Lightspeed VCT: Another high-performance scanner with a reputation for quick and high-resolution imaging is the GE Healthcare Lightspeed VCT. It has cutting-edge technology like low-dose imaging protocols and the Volume Imaging Protocol (VIP), which guarantee thorough lung scans with little radiation exposure. Because of its quick picture acquisition capabilities, this scanner is perfect for use in high-throughput clinical settings and emergency situations where time is of the essence. Its advantages include quick image acquisition, improved workflow efficiency, detailed images appropriate for in-depth analysis, low-dose protocols that minimize radiation exposure while preserving image quality, and quick and easy image acquisition in hectic clinical settings.

Its shortcomings include the necessity for frequent calibration to preserve picture accuracy, the high cost of maintenance and consumables, the vulnerability to artifacts caused by patient movement, and the limited availability of specialized imaging modes in certain configurations. Table II summarizes the GE Healthcare Lightspeed VCT Strengths and Weaknesses.

TABLE II. GE HEALTHCARE LIGHTSPEED VCT STRENGTHS AND WEAKNESSES

GE Healthcare Light speed VCT	
Strengths	**Weaknesses**
Rapid image acquisition	High operational costs
High-resolution imaging	Susceptibility to movement artifacts
Low-dose protocols	Requires frequent calibration
Efficient workflow	Limited specialized imaging modes

To guarantee data integrity, quality and patient confidentiality, the generated images were safely stored in DICOM format on encrypted external drives. The external drives were kept in a safe, climate-controlled environment, and frequent backups were made to guard against data loss. Extensive metadata was recorded to make retrieval and analysis simple.

B. DICOM Format CT Scans

The images in the collection are kept in the DICOM (Digital Imaging and Communications in Medicine) format, which is a commonly utilized format for organizing, transferring, and storing data related to medical imaging. DICOM is made to make sure that systems that create, show, transmit, store, query, process, retrieve, print, and manage medical pictures can communicate with one another. The DICOM format was selected primarily because it can maintain excellent picture quality without adding compression artifacts, which is essential for preserving the images' diagnostic integrity.

In addition to the image data, DICOM files include an abundance of metadata, such as patient demographics, scan parameters, imaging modality specifics, and facts on the hospital and its equipment. This metadata is immediately included into the DICOM file, offering a thorough record that is necessary for precise diagnosis, study repeatability, and other purposes. For instance, for comparison research and to ensure uniformity between scans, scan characteristics including slice thickness, resolution, and radiation dose are essential.

Furthermore, DICOM is a flexible option for multi-modality imaging investigations since it supports a broad variety of imaging modalities, such as CT, MRI, ultrasound, and X-ray. Sensitive information is safeguarded since the format complies with global standards for patient privacy and medical data security.

By using the DICOM format, the dataset can be used in a variety of clinical and research settings because it is compatible with a wide range of medical imaging applications and systems. In order to provide easy access and analysis by medical professionals, this compatibility is especially crucial for integration with Picture Archiving and Communication Systems (PACS), which are utilized in clinics and hospitals. Furthermore, DICOM's ability to handle sophisticated imaging capabilities including 3D reconstructions and multi-frame functionality increases its usefulness for in-depth research of lung cancer.

CT scans are especially useful in the detection of lung cancer because they provide high-resolution images that can detect tumors and small nodules that might not be visible with other imaging modalities. Additionally, CT imaging offers excellent contrast between different types of tissue, which is crucial for precisely identifying and characterizing lung nodules, as well as determining their size, composition, and size.

CT scans provide 3D reconstruction of the lung structure, offering a thorough perspective that facilitates accurate tumor location and evaluation in relation to adjacent tissues. Planning surgical procedures, directing biopsies, and tracking

the efficacy of treatments over time are all made possible by this capacity.

By choosing CT scans in DICOM format, we ensure that the dataset meets the highest standards of image quality, data integrity, and interoperability, making it a robust and valuable resource for ongoing and future research in lung cancer diagnosis and treatment. Detailed information about the dataset is provided in the Table III.

TABLE III. DATASET DETAILS FROM MILITARY HOSPITAL OF INSTRUCTION OF TUNIS (HMPIT)

Attribute	Details
Number of Patients	123 (including both healthy and sick patients)
Health Status	• Healthy Patients: 25 • Sick Patients: 98
Format	DICOM
Types of Lung Cancer	• Adenocarcinoma: 45 patients • Squamous Cell Carcinoma: 33 patients • Small Cell Lung Cancer: 20 patients
Stages	• Stage I: 34 patients • Stage II: 26 patients • Stage III: 28 patients • Stage IV: 10 patients
Annotations	• Number of nodules: Detailed count per patient • Size of nodules: Measurements in millimeters • Nodule type: Solid, part-solid, or ground-glass • Nodule characteristics: Margin, shape, and calcification status • Follow-up status: Monitoring of nodule changes over time • Tumor volume and density • Tumor growth rate (if multiple scans available) • Identification and annotation of affected lymph nodes • Documentation of any metastasis to other parts of the body visible in scans
Imaging Modalities and Scanners	• **GE Healthcare LightSpeed VCT:** High-speed imaging capabilities and good spatial resolution • **Siemens SOMATOM Perspective CT Scanner:** Precise imaging and dose efficiency
Technical Details	• Resolution: Typically 512 x 512 pixels • Slice Thickness: 1-5 mm, ensuring consistency and quality • Number of Slices per Scan: Average of 300 slices per CT scan • Scan Duration: Approximately 5-15 minutes per scan • Imaging Dates: January 2020 - April 2024 • Scanner Settings: Voltage (120 kVp), Current (200-400 mA), Exposure Time (0.5-1 seconds perslice)
Demographics	• Age: Range from 20 to 80 years, with an average age of 60 • Gender: 73 males, 50 females • Relevant Medical History: Includes smoking history (80% of patients), family history of lungcancer (30% of patients)

Clinical Data	• Symptoms:
	  ○ Cough: 84 patients
	  ○ Chest pain: 53 patients
	  ○ Shortness of breath: 76 patients
	• Treatment History:
	  ○ Surgery: 42 patients
	  ○ Chemotherapy: 69 patients
	  ○ Radiation therapy: 50 patients
	• Outcomes:
	  ○ Survival rate
	  ○ Recurrence
	  ○ Cancer-free
	• Additional Annotations:
	  ○ Histopathological findings
	  ○ Genetic mutations (e.g., EGFR, ALK)
	  ○ Biomarker levels (e.g., PD-L1 expression)

C. Data Quality

The lung cancer dataset was created with the highest priority on ensuring high data quality. Advanced imaging technologies such as Siemens SOMATOM Perspective and GE Healthcare LightSpeed VCT CT Scanners were used to obtain all CT scans. This ensured that all images were high-resolution, with a typical resolution of 512 x 512 pixels and consistent slice thicknesses ranging from 1 to 5 mm. In order to protect the quality and integrity of the images and prevent compression artifacts, they were stored in DICOM format.

Qualified radiologists painstakingly analyzed the images, recording in-depth information regarding every nodule, such as numbers, millimeter diameters, kinds (solid, part-solid, or ground- glass), and features including margin, shape, and calcification status. The volume, density, and growth rates of the tumor were annotated, as well as the lymph nodes that were afflicted and any obvious metastases. These thorough annotations were saved as structured CSV files, which offer a common format for simple analysis and integration with different data processing applications.

A thorough validation procedure was put in place to guarantee the highest level of accuracy. Peer evaluations of the annotations, several cross-checks, and consistency checks against accepted medical norms were all part of this process. To make sure that data management procedures were being followed, the dataset also went through routine audits and quality reviews. To stop data deterioration, the external disks holding the data were encrypted, often backed up, and kept in a safe, climate-controlled location.

The dataset is an important and dependable resource for research and development in lung cancer diagnosis and therapy because of the exact, well-documented annotations and high-quality photos. Table III contains comprehensive details about the dataset.

D. Data Preparation

Making ensuring the raw data is appropriately structured, cleaned, and arranged is the goal of data preparation so that it may be used for additional processing and analysis. There were several important steps in this phase. To enable uniform analysis, all CT scans were first standardized to a uniform resolution and slice thickness, usually between 1 and 5 mm. The maintenance of uniformity across images acquired from various scanners, such as the Siemens SOMATOM Perspective and the GE Healthcare Lightspeed VCT, required this standardization. The quality of the scans was then improved by applying image noise reduction techniques, which included algorithms to filter out aberrations and improve the visibility of minute features.

After that, radiologists performed a preliminary examination of the scans to find and fix any irregularities, like motion artifacts or partial images. This quality check made sure that the dataset contained only the best images. Critical data, including patient demographics, scan parameters, and gear characteristics, were included in the metadata and carefully checked for accuracy.

1) Data cleaning: Data cleaning in the data preparation stage of our Tunisian lung cancer dataset entailed finding and fixing mistakes or inconsistencies in the DICOM images. Before processing the dataset further, this involved correcting missing values, standardizing data formats, and eliminating duplicate records in order to guarantee its integrity and quality.

2) CT imaging parameters: There are 123 subjects' worth of CT images in DICOM format available. Since this is a ret-respectively gathered dataset, various subjects were scanned with different scanners, protocols, and parameters: slice thickness of 1-4 mm (median: 3 mm) and an X-ray tube current of 200-400 mA (mean 250 mA) at 100-140 kVp (mean 120 kVp). Specific scanning parameters, such as the make and model of the scanner, are specified in the DICOM headers. The subjects were scanned while supine, and the scans were obtained from the apex of the lung to the adrenal gland in a single breath-hold.

3) Image segmentation: Further, a segmentation procedure was applied to each scan, defining the lung regions in order to isolate the key areas of the image and concentrate on the area of interest. For it to enable accurate annotations and lessen computing load during analysis, this step was crucial. As part of the preparation stage, imaging protocols were standardized to reduce variability brought about by various scanning configurations and methods.

E. Image Annotations

The process of locating and labeling pulmonary nodules in CT scans and other medical imaging studies is known as "nodule annotation." Medical professionals must conduct a thorough examination of 3D volumetric data in order to identify, quantify, and categorize nodules that might be signs of lung cancer. The purpose of nodule annotation is to produce an accurate and thorough dataset that can be utilized to train machine learning models for dependable and accurate lung nodule detection.

The first step in annotating our our Tunisian lung cancer dataset CT scans, is selecting the right annotation software. We employed software such as 3D Slicer and ITK-SNAP, which are well acclaimed in the medical imaging field for their feature-rich and intuitive interfaces. For the purpose of designating pulmonary nodules, these technologies are perfect because they enable the thorough inspection and annotation of 3D volumetric data.

- 3D Slicer is an open-source software platform that can handle a wide range of imaging formats and is highly adaptable for medical image informatics, image processing, and three-dimensional visualization. Users can load DICOM images, and it offers strong visualization tools for precisely locating and labeling nodules [20].
- Another well-liked software with a focus on 3D medical image segmentation is ITK-SNAP. Experts can more easily annotate nodules with precision thanks to its

semi-automated segmentation capabilities and user-friendly manual segmentation features [21].

Second step, the CT scans for our Tunisian dataset were imported into the selected tools which were set up to show the pictures in a way that makes it simple to identify nodules. This setup comprised:

- Changing the Window and Level Settings: To bring the nodules out against the background lung tissue, adjust the brightness and contrast.
- Multi-Planar Reconstructions (MPR) are made possible: permitting views in the axial, sagittal, and coronal planes to give a thorough understanding of the nodule's structure.
- Effective Navigation: Guaranteeing that images are easily panned, zoomed in on, and navigated so that specialists can examine them in-depth.

After that, each scan was examined by radiologists and oncologists from hospitals in Tunisia to look for lung nodules. Being able to differentiate nodules from other anatomical features and possible artifacts needed a high level of knowledge. In order to obtain a thorough grasp of the nodule's properties, we collaborated with the specialists and made use of the tools' features to zoom in on areas of interest, change the contrast of the image, and switch between different views.

The following was a part of the annotation process:

- Exact Position: The nodule's x, y, and z coordinates were noted. Mapping the nodule's location in the lung's three-dimensional space requires these coordinates.
- The size measurement: A measurement of the nodule's diameter was made. This measurement aids in the classification of the nodule and determines whether it is potentially cancerous.
- Classification: Based on its features, each nodule was categorized. Nodules were often categorized as malignant or benign (0).
- Existence of Nodules: In order to clearly distinguish between scans with and without nodules, this was additionally noted if there were none.
- Disease Stage: Each patient's disease stage was recorded, which added further context for the severity and course of the sickness.

More annotations are found in the dataset descriptive Table I.

For convenience of access and integration with machine learning techniques, the annotated data was stored as a CSV file. The CSV file contained multiple distinct columns, each of which represented a single nodule. Table IV lists and describes some annotations in our CSV file.

Several experts examined the annotations to guarantee their uniformity and accuracy. After disagreements between reviewers were reviewed and settled, the final annotations underwent validation and quality control to make sure they adhered to the required accuracy requirements. In order to confirm the validity and suitability of the Tunisian lung cancer dataset for machine learning model training, this step was essential.

TABLE IV. ANNOTATION CSV FILE COLUMNS

Column	Description
patient_id	A distinct identitfier for each patient.
series_id	A unique and distinct identifier for the collection of images ofeach patient.
coordX, coordY, coordZ	The nodule coordinates within the lungs.
diameter_mm	The nodule diameter in millimeters.
class	The nodule classification (0 for benign, 1 for malignant).
nodule_present	A boolean representing whether nodules are present (1) or absent (0).
stage	The cancer stage (Stage I, Stage II, Stage III, Stage IV).

F. Data Compliance and Standards

One of the main tenets for establishing the lung cancer dataset was adhering to legal and ethical guidelines. The relevant institutional review board granted ethical approval to the project prior to data collection, guaranteeing that the study complied with all relevant ethical standards. Every patient gave their informed consent, ensuring that they understood exactly how their information would be used, maintained, and safeguarded.

In accordance with the Tunisian National Instance for the Protection of Personal Data (INPDP) [18], the dataset was painstakingly de-identified to remove any personal identifiers. The integrity and usability of the data were preserved while patient privacy was protected thanks to this de-identification procedure.

1) De-Identification of imaging DICOM data: Before being analyzed at the Military Hospital of Instruction of Tunis (HMPIT), all imaging data were de-identified. Using XNAT (eXtensible Neuroimaging Archive Toolkit), we were able to de-identify the imaging data. With the help of XNAT, medical imaging data can be securely managed and made anonymous, guaranteeing that DICOM objects no longer contain protected health information (PHI).

Every personal identification was eliminated from the dataset in order to preserve patient confidentiality. By de-identifying the data, privacy laws were satisfied with by the dataset. To further improve the dataset's resilience, several versions of the preexisting photos were produced using data augmentation techniques. To give context and make it easier for other researchers to use the dataset, thorough documentation about its creation, properties, and annotations was produced.

We used again XNAT to execute a second round of de-identification before releasing the data for research, ensuring that all identifying information had been completely removed. With options like Clean Pixel Data, Clean Descriptors, Retain Longitudinal with Modified Dates, Retain Patient Characteristics, Retain Device Identity, and Retain Safe

Private Options, this de- identification procedure conforms to international requirements for medical data privacy.

2) Data encryption: The DICOM standard, which offers a framework for the interchange and storage of medical images and related data, is one of the worldwide standards for medical imaging that the dataset was created to comply with. The broad use and integration of medical imaging devices and software is facilitated by compliance with DICOM standards, which guarantees interoperability.

The dataset was also stored using encryption and frequent backups, which followed the best practices for data security and integrity. We made sure that the dataset respects patient rights and privacy in addition to meeting high-quality benchmarks by closely adhering to these ethical and regulatory norms. This makes it a dependable and morally sound resource for lung cancer research.

To preserve data quality, the external disks were kept in a safe, climate-controlled environment. Extensive metadata documentation made it simple to retrieve and analyze the dataset, which made it a strong and useful tool for studying lung cancer.

VII. Tunisian Lung Cancer Dataset Model Selection

To make certain our lung cancer dataset is high-quality and useful for training cutting-edge machine learning models, it must be tested. Extensive tests enable us to assess the dataset's robustness and detect any potential biases or restrictions that can impair model performance. We can learn a great deal about the dataset's suitability for lung nodule identification and diagnosis by carefully evaluating it. Our dataset's value in practical applications is demonstrated by benchmarking it against well-established models, which also reveals its potential to increase diagnostic accuracy. The results of these studies will serve as a strong basis for upcoming investigations, propelling the creation of more accurate and effectivemedical imaging instruments.

A. Comparative Analysis of Model Architectures

In this section, we compare and contrast a number of renowned model architectures from the field of medical imaging, including CNN, U-Net, VGG, and ResNet. These models were selected for comparison because they are widely used and have a track record of success in a variety of image processing applications, including medical imaging. It is crucial to comprehend these models' performance and applicability for lung nodule identification in order to choose the best architecture for our dataset. We hope to determine the advantages and disadvantages of each model through this comparison, giving a convincing explanation for our selection. This comparison analysis aids in our decision-making process for choosing the most suitable model for our application by offering a thorough grasp of how various architectures function in the contextof lung nodule identification.

Table V shows the different architecture and various use cases of each model mentioned.

Table VI lists the multiple advantages and also disadvantagesof each model.

TABLE V. Models Architecture and Use Cases

Model	Architecture	Use Case
CNN [25]	sequence of convolutional layers, pooling layers, and fully connected layers in order of succession	General image classification
U-Net [26]	symmetric layer encoder-decoder design with skip connections	Biomedical image segmentation
VGG [27]	16 or 19-layer deep architecture with tiny (3x3) convolution filters	Large-scale image classification
ResNet [19]	Identity mapping can be achieved with a deep architecture featuring residual blocks.	Complex image classification and detection

TABLE VI. Models Advantages and Disadvantages

Model	Advantages	Disadvantages
CNN [24]	Simple and efficient for extracting features, well-established and straightforward to develop	Vanishing gradient causes Problems with highly deep networks, which may necessitate extensive tweaking for complicatedtasks.
U-Net [28]	Great for segmenting images, very accurate for localization tasks	Computationally demanding, could not adapt well to tasks requiring classification without adjustments
VGG [29]	Robust large-scale image classification performance with a straightforward and deep architecture	High memory consumption, high computational expense, and less useful for very deepnetworks
ResNet [30]	Residual learning reduces the vanishing gradient issue and enables the formation of extremely deep networks with exceptional performance on challenging tasks.	Can have a more complicated architecture and be computationally demanding than conventional CNNs.

The comparison study draws attention to the unique traits and functionalities of the CNN, U-Net, VGG, and ResNet models. Every architecture has advantages and disadvantages that affect which medical imaging tasks they are best suited for.

Based on the unique needs of lung nodule detection—which necessitates a deep architecture capable of capturing delicate and detailed features—ResNet models were chosen over CNN, U-Net, and VGG. We have collected high-resolution CT scans from 123 individuals 80% from them have lung cancer in Tunisia and 20% are healthy. This large and heterogeneous dataset demands a model that can efficiently identify and learn from intricate patterns and minute differences in the data. A model that can successfully capture and learn from intricate patterns and minor variations in the data is required because of this large and diverse dataset.

The vanishing gradient issue is successfully addressed by ResNet's residual learning framework, which makes it especially suitable for this purpose and makes it possible to train very deep networks—which are necessary for high-accuracy detection tasks. ResNet is perfect for managing the complex characteristics in our dataset because of its ability to retain performance in deep networks by alleviating the vanishing gradient issue [19], making it possible to extract detailed features from complicated data. Even with deeper

network architecture, steady training and enhanced performance are guaranteed by the incorporation of residual blocks and skip connections. CNNs lack the depth required for more sophisticated tasks, even if they are simple and efficient for basic picture categorization [24]. U-Net performs quite well in segmentation, but its large processing overhead increases when applied to classification tasks [28]. Though powerful, VGG's high memory needs make it computationally costly and less useful for very deep networks [29].

The most balanced method for creating a lung nodule identification model that can effectively utilize the rich and extensive data in our Tunisian dataset is ResNet, thanks to its depth, resilience, and performance. Accurate and dependable lung nodule detection in a variety of clinical scenarios can be efficiently supported by its ability to handle complicated data structures and retain high accuracy [30].

B. Resnet Models

In their 2015 publication"Deep Residual Learning for Image Recognition", Kaiming He et al. [19] introduced ResNet, short for Residual Network, a kind of deep neural network. ResNet's main breakthrough is residual learning architecture, which makes it possible for the network to train considerably deeper models than it could have before. With this invention, the vanishing gradient problem—a prevalent difficulty in deep learning—is addressed. As network depth increases, gradients become increasingly small, making learning ineffective. ResNet models were initially created to classify 2D images; however, they have since been expanded to 3D versions to handle volumetric data, including CT scans. To be more specific, ResNet models have been expanded to 3D versions [23] in the context of medical imaging, particularly for 3D data such as CT and MRI scans. These models make use of 3D convolutional layers, which perform three-dimensional convolution operations to capture spatial data in the dimensions of depth, height, and width. This is crucial for activities that depend on the geographical context in three dimensions, such as lung nodule detection.

Multiple residual blocks, each having a set of convolutional layers, make up ResNet models. The input is added back to the original input after passing through convolutional layers in a residual block, creating a skip or shortcut link. This facilitates the learning of identity mappings by the model and aids in maintaining the gradient flow, which facilitates the training of deeper networks.

Bypassing one or more layers, the skip connections add the input straight to the stacked layers' output. This lessens the degradation issue, which occurs when a sufficiently deep model gains more layers, increasing training error.

The multiple layers that make up the architecture of 3D ResNet models are intended to capture varying degrees of abstraction from the input data. The essential elements consist of [19]:

- 3D Convolutional Layers: These layers use three-dimensional convolution processes to capture spatial data related to the input volumes' depth, height, and width.
- Layers for batch normalization: These layers speed up training and increase the stability of the model by normalizing the output of convolutional layers.
- Layers of ReLU Activation: The Rectified Linear Unit (ReLU) activation adds non-linearity to the model so that it may pick up intricate patterns.
- Residual Blocks: By allowing the model to learn residual functions in relation to the layer inputs, these blocks make it possible to build extremely deep networks without experiencing any degradation.
- Pooling layers: These layers help to downsample the data and lower computational complexity by reducing the spatial dimensions of the input.
- Fully Connected Layers: These layers create final predictions at the conclusion of the network by combining features that were extracted by earlier levels.

ResNet models come in a number of depths: ResNet10, ResNet18, ResNet34 and ResNet50. The number denotes the total number of layers in each model. To depict varied levels of complexity and detail, these models feature different arrangements of leftover blocks [19].

- ResNet10:

Architecture: Ten-layer ResNet's most basic model. It is effective at capturing important information during training even with constrained computational resources.

Use: Fits well with activities that need faster inference times and less complexity.

- ReNet18:

Architecture: An eighteen-layered, relatively deeper model. Its ability to strike a balance between performance and complexity qualifies it for a variety of uses.

Use: Frequently applied to tasks involving generic medical picture classification.

- ResNet34:

Architecture: A 34-layer, deeper model that enables more precise feature extraction.

Use: Perfect for jobs like segmentation and tiny anomaly identification that call for in-depth analysis and excellent accuracy.

- ResNet50:

Architecture: A complex model with 50 layers, offering the highest capacity for capturing intricate patterns in the data.

Use: Best suited for highly detailed tasks that require extensive computation, such as multi-class segmentation and advanced diagnostic analysis.

VIII. DATASET ROBUSTNESS TESTING

We ran thorough tests using multiple 3D ResNet models to assess the resilience of our lung cancer dataset. CT scan pictures with annotations for lung nodules were used to

train the models. Several ResNet designs (ResNet10, ResNet18, ResNet34, and ResNet50) were used in the training process, and the outcomes were contrasted with those attained using the Tencent MedicalNet models.

A. Tencent MedicalNet Models

A set of pre-trained models created especially for medical imaging tasks are available through Tencent's MedicalNet initiative [22]. The models are optimized for certain tasks, such lung nodule identification, after having undergone extensive and varied pre-training on a vast collection of medical images.

We painstakingly duplicated Tencent MedicalNet's experimental setting to verify the reliability of our lung cancer dataset. To guarantee a direct and impartial comparison between the MedicalNet models' and our dataset's performance, this required sticking to the same 3D ResNet models and training parameters.

A wide range of modalities, target organs, and diseases were covered by the 23 datasets that were combined for the MedicalNet project. The models can acquire universal feature representations through this thorough pre-training, which they may then apply to a variety of medical imaging tasks. The models were tested for adaptability and high performance on a variety of tasks, such as lung segmentation and pulmonary nodule classification. The research ensured a thorough and diversified dataset for pre-training by compiling data from multiple sources, such as MRI and CT scans. The study made use of a variety of 3D ResNet designs (ResNet10, ResNet18, ResNet34, and ResNet50) to capture varying degrees of intricacy and detail in the data. MedicalNet used spatial and intensity normalizing approaches to address the diversity in spatial resolution and intensity distributions. This improved the training process by guaranteeing that the data given into the models was consistent.

We selected Tencent's MedicalNet to showcase the resilience of our lung cancer dataset, thanks to its pre-trained 3D ResNet models. Pre-trained on an extensive and varied collection of medical images, MedicalNet's models improve their generalization and performance on a range of tasks. We are able to assess our dataset's quality and its potential to help construct high-performance diagnostic tools by using these pre-trained models. This thorough assessment highlights the contribution of our dataset to the advancement of medical image processing in general and lung nodule detection specifically.

B. Transfer Learning

Transfer learning is a potent deep learning technique in which a pre-trained model is refined on a smaller, task-specific dataset after it was first trained on a larger dataset. By utilizing the knowledge gained from the lengthy pre-training phase, this method improves generalization and increases the model's efficiency in learning from the smaller dataset. In medical imaging, where it might be difficult to gather big annotated datasets, transfer learning is very helpful [32]. We may greatly improve our models' performance by utilizing pre-trained models, like Tencent's MedicalNet, since they gain from the wide range of feature representations that are acquired during the pre-training stage. This methodology enhances the models' accuracy and robustness when used for particular tasks, such lung nodule identification in our dataset, while also lowering the computational resources needed for training.

We used pre-trained 3D ResNet models from Tencent's MedicalNet to implement transfer learning in our study. Since a big and varied collection of medical images served as the initial training set, the models were able to pick up a wealth of attributes pertinent to medical imaging. With the help of our lung cancer dataset, we adjusted these pre-trained models so they could be specifically used for lung nodule detection.

The procedure entailed starting with the MedicalNet models' pre-trained weights and completing the training on our dataset. By using this method, the models were better able to identify and categorize lung nodules because they could make use of the generic traits that they had acquired during the first training phase. Our goal in fine-tuning these models was to bring together the unique characteristics of our dataset with the advantages of thorough pre-training.

C. Pre-processing and Training

We were able to use the same architectures—ResNet10, ResNet18, ResNet34, and ResNet50—and apply comparable pre-processing methods, optimization tactics, and evaluation criteria by coordinating our experiments with those carried out by MedicalNet. We were able to provide a thorough and consistent review thanks to this strategy, which also made sure that any discrepancies in performance could be traced back to the datasets themselves instead of deviations in methodology.

To guarantee consistency and enhance the learning process, the CT scan images had been processed before being used for the training and assessment of the 3D ResNet models. Among the preprocessing actions were:

- Format Conversion: DICOM CT scans were programmatically transformed to NEFTII format. This modification made handling and processing of both our and Tencent Medicalnet volumetric data more efficient.
- Normalization: To make sure that the intensity values were scaled correctly for the neural network, each CT scan was normalized to a range of [-1, 1].
- Resizing: To standardize the input size and lower processing needs, the scans were downsized to a uniform shape of 64x64x64 voxels.
- Data Augmentation: During training, data augmentation techniques like random rotations and flips were used to improve the models' capacity for generalization.
- Data Division: The dataset was divided into training, validation, and testing subsets in order to guarantee the efficient training and assessment of machine learning models. This tactical separation is essential to creating reliable and accurate models. The full range of variations seen in the entire dataset was carefully reflected in these divides, which were made to maintain diversity and balance. Training is for 70% of the split, validation for 15%, and testing for 15%.

The preliminary actions made to collect, arrange, and structure the raw data in order to make it suitable for analysis or modeling are referred to as data preparation. A more detailed step called "data pre-processing" entails getting the cleaned and sorted data ready for the real machine learning or data analysis work. The goal of this step is to change the data in order to improve the models' accuracy and performance. Therefore after making sure that our data is ready for the models we proceed to the training.

There were multiple steps in the training:

1) Data loading: After being loaded, CT scan images underwent preprocessing to standardize and resize them into a form that would work with the models. Tencent MedicalNet's pre-trained ResNet models, including ResNet10, ResNet18, ResNet34, and ResNet50, are loaded. These models have a good pattern recognition capacity because they have already been trained on big datasets. The final layers are adjusted to meet our classification requirements in order to customize these models for our particular task of lung nodule identification. By utilizing the power of transfer learning, this phase enables the pre-trained models to efficiently apply the features they have learnt to our dataset.

2) Hyperparameter configuration: One important stage in the training process is configuring the hyperparameters. The learning rate, which regulates the step size during gradient descent, the number of epochs, or full runs through the training dataset, the batch size, which establishes the quantity of samples processed before the updating of the model's internal parameters, and the loss criteria, which direct the optimization procedure, are important hyperparameters. Appropriate hyperparameter selection is essential to maximize model performance and guarantee effective training.

3) Training of the models: Using the training set, the models' weights are modified during the training phase. The validation set is used to assess the model's performance at each epoch in order to keep an eye out for overfitting. When a model performs well on training data but poorly on unknown data, this is known as overfitting. We can reduce overfitting by using early stopping or other regularization strategies by evaluating the validation set. Every ResNet model underwent ten epochs of training, during which the accuracy and loss were noted.

4) Evaluation: Accuracy served as the main performance indicator for each model. Ten epochs were required to record the final accuracy. The test set is used to assess the final models' performance after training, giving an objective appraisal of the model's capabilities. The model's accuracy is assessed to assess how well it detects lung nodules. To verify the reliability and efficacy of our dataset and model modifications, these outcomes are then contrasted with the performance metrics of the previously trained models on comparable datasets

D. Results and Analysis

The resilience and good quality of the dataset were demonstrated by the models trained on it, which repeatedly displayed excellent performance.

The accuracy trends of each model trained on our dataset, are clearly represented visually in the Fig. 6.

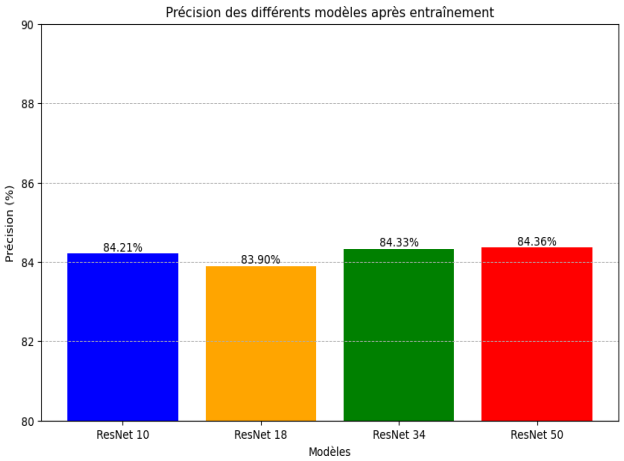

Fig. 6. Comparison of the accuracy achieved by different 3D ResNet models over the training epochs

Following training, each model's final accuracy is listed in Table VII below which shows a comparison between the accuracy of each model trained on our dataset and the models trained on Tencent MedicalNet datasets.

TABLE VII. COMPARISON OF ACCURACY BETWEEN OUR DATASET AND TENCENT MEDICALNET MODELS

Model	Your Dataset Accuracy	MedicalNet Accuracy
ResNet10	84.21%	96.56%
ResNet18	83.90%	94.68%
ResNet34	84.33%	94.14%
ResNet50	84.36%	89.25%

All of the models that were trained on our dataset performed admirably, with an accuracy rate of above 84%. ResNet50 demonstrated the best accuracy of 84.36%, demonstrating the resilience of our dataset in the identification of lung nodules. Nevertheless our models' accuracy was slightly lower than Tencent's MedicalNet models', which were pre-trained on a larger and more varied collection of medical images. For example, the MedicalNet ResNet10 model attained an astounding 96.56% accuracy, while our dataset only managed 84.21%. There are various reasons for this disparity.

- First off, MedicalNet has a big edge because to its thorough pre-training on a variety of medical images. By learning a wide range of characteristics that are applicable to many tasks, the models benefit from this pre-training, which improves their performance on new datasets. Even though our dataset is strong, it is smaller and less varied than MedicalNet's, which restricts the models' capacity to generalize to previously undiscovered data.

- Second, the improved performance of the MedicalNet models can be attributed to the variety in the dataset, which encompasses numerous modalities and target organs. This variety enhances the models' accuracy and resilience across a range of tasks by enabling them to gain a more thorough grasp of medical imagery.
- Thirdly, more thorough training and fine-tuning are made possible by MedicalNet's large computational resources and longer training periods, which can have a big impact on the final performance. We could get better outcomes if we extend the training period and increase our computational resources.
- Fourthly, the discrepancies in accuracy seen might have been caused by the scanners we utilized to obtain our dataset, including the Siemens SOMATOM Perspective and GE Healthcare Lightspeed VCT. Variations in imaging techniques and scanner features may result in inconsistent image quality and resolution, which could have an impact on the performance of the model.

E. An Overview of the Tunisian Lung Cancer Dataset Creation Workflow

To ascertain the quality, reliability, and usability of the Tunisian lung cancer dataset for the development of sophisticated machine learning models, a number of crucial procedures have to be taken during the creation process, like shown in Fig. 7.

1) Data collection: We started by gathering DICOM images from the Military Hospital of Instruction in Tunisia (HMPIT). Siemens SOMATOM Perspective and GE Healthcare Lightspeed VCT were the two scanners used to capture the images.

2) DICOM image storage: The integrity and confidentiality of patient data were then preserved by importing these images onto a safe external device.

3) Data preparation: The data preparation stage began along with gathering and safely storing the DICOM images from the Military Hospital of Instruction of Tunis (HMPIT). Setting up parameters, fixing mistakes, getting rid of duplication, and standardizing formats are all part of this phase. Furthermore, segmentation is done to divide the data into areas that make sense, allowing for more focused and effective analysis. By doing this, we guarantee that the dataset is reliable, consistent, and prepared for the thorough annotation and pre-processing stages necessary for training a machine learning model.

4) Nodule annotation: We worked together with specialized software like 3D Slicer and ITK-SNAP to annotate the CT images. Strong capabilities and intuitive user interfaces were offered by these tools for in-depth examination and annotation.

5) Nodule annotation validation: Several experts examined the annotations to guarantee uniformity and accuracy. Consensus meetings were used to settle disagreements.

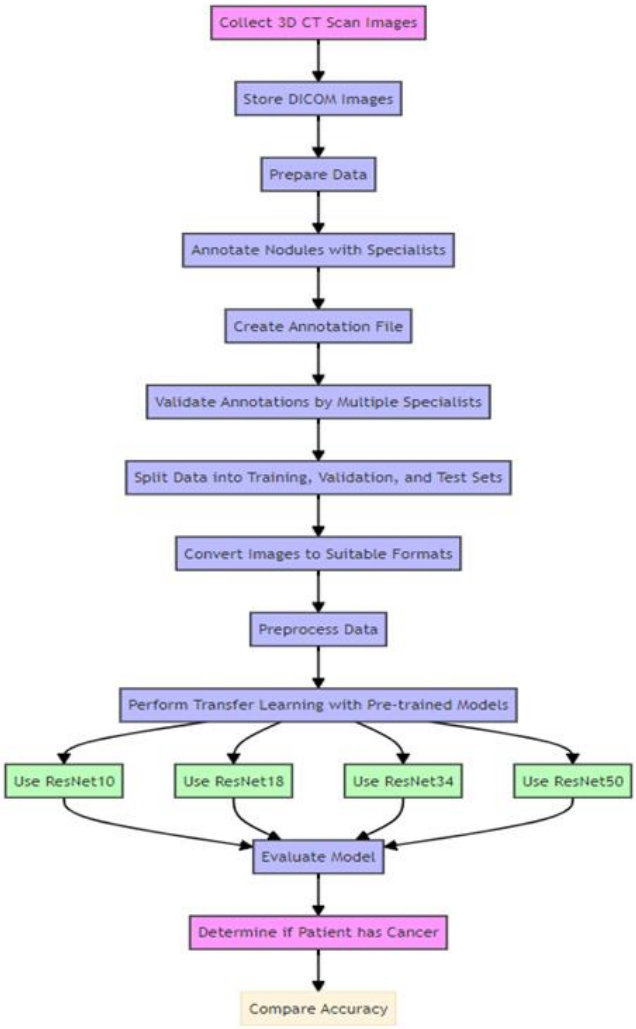

Fig. 7. The workflow of the creation and validation of the lung cancer Tunisian dataset.

6) Dataset splitting: To make sure that each set accurately reflected the diversity of the full dataset, it was divided into training, validation, and testing sets. Typically, training would account for 70% of the split, validation for 15%, and testing for 15%.

7) Data pre-processing: To ensure compatibility with Tencent MedicalNet models, data pre-processing for our Tunisian lung cancer dataset project entailed converting DICOM pictures to NEFTII format. Standardized image resolutions and normalized intensity values were achieved. Rotation and flipping are examples of data augmentation techniques that produced additional training samples. To ensure accurate model evaluation and peak performance, the dataset was finally divided into 70% training, 15% validation, and 15 % testing.

8) Transfer learning with ResNet models from tencent MedicalNet: We used Tencent MedicalNet's pre-trained ResNet models (e.g. ResNet10, ResNet18, ResNet34, and ResNet50). We adjusted these models with our Tunisian lung

cancer dataset. Using the knowledge from pre-trained models, transfer learning was used to improve performance on our particular dataset.

9) Experiments and validation: We ran experiments to assess how well the refined ResNet models performed using our dataset defining how well the models can detect lung cancer patients thus the accuracy was measured.

10)Compare accuracy: The robustness of our dataset was evaluated by contrasting its results with those obtained from the MedicalNet models. To make sure the models translate effectively to fresh, untested data, they were verified using the testing set. Upon contrasting our accuracy outcomes with those obtained from MedicalNet, we discovered that although our dataset had strong performance, the models trained on MedicalNet data demonstrated slightly greater accuracy. This demonstrates that in order to match the performance of existing datasets, additional improvements in data quality and diversity are required.

Every stage of the dataset creation procedure, including data preparation, annotation, pre-processing, training of models, and collection, was thoroughly documented. This documentation guarantees reproducibility and offers precise instructions for further study and advancements. In order to provide transparency and promote cooperation with other researchers, it also includes metadata regarding the dataset, annotation processes, and pre-processing techniques utilized. Following ethical and privacy rules, the dataset and model results were shared and archived securely.

IX. Conclusion

To enhance lung nodule detection and develop diagnostic techniques tailored to the local population, it is crucial to address the lack of a lung cancer dataset in Tunisia. We assembled a comprehensive dataset of 123 well-annotated DICOM-format CT images from various locations within Tunisia. By utilizing pre-trained 3D ResNet models from Tencent's MedicalNet and applying transfer learning, we validated the robustness of our dataset. After refinement, these models exhibited outstanding performance, demonstrating the effectiveness of our approach.

The significance of broad and varied pre-training on a variety of datasets is shown by the superior performance of MedicalNet models. Future work will focus on several key areas to enhance the dataset and its applicability. First, improving pre-processing and augmentation techniques will be crucial to improve the quality and robustness of the dataset. Additionally, we aim to expand the dataset by including more diverse and comprehensive data sourced from additional medical institutions across Tunisia. Incorporating multi-modality imaging, such as MRI and PET scans, will provide a more holistic view of lung cancer characteristics, enhancing the depth and scope of the dataset. Finally, we will seek collaboration with international research bodies to standardize annotation protocols and integrate the Tunisian dataset with global datasets, facilitating broader applicability and creating new research opportunities.

References

[1] H. B. Schiller, D. T. Montoro, L. M. Simon, E. L. Rawlins, K. B. Meyer, M. Strunz, F. A. Vieira Braga, W. Timens, G. H. Koppelman, G. R. S. Budinger, J. K. Burgess, A. Waghray, M. van den Berge, F. J. Theis, A. Regev, N. Kaminski, J. Rajagopal, S. A. Teichmann, A. V. Misharin, and M. C. Nawijn, "The human lung cell atlas: A high- resolution reference map of the human lung in health and disease," Am. J. Respir. Cell Mol. Biol., vol. 61, no. 1, pp. 31–41, Jul. 2019.

[2] O. Khouadja and M. S. Naceur, "Lung Cancer Detection with Machine Learning and Deep Learning: A Narrative Review," 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC ASET), Hammamet, Tunisia, 2023, pp. 1-8, doi: 10.1109/IC ASET58101.2023.10150913.

[3] H. Zhang, D. Meng, S. Cai, H. Guo, P. Chen, Z. Zheng, J. Zhu, W. Zhao, H. Wang, S. Zhao, J. Yu, and Y. He, "The application of artificial intelligence in lung cancer: a narrative review," Translational Cancer Research, vol. 10, no. 5, 2021.

[4] Jiang X, Hu Z, Wang S, Zhang Y. Deep Learning for Med- ical Image-Based Cancer Diagnosis. Cancers (Basel). 2023 Jul 13;15(14):3608. doi: 10.3390/cancers15143608. PMID: 37509272; PMCID: PMC10377683.

[5] Chaudhry R, Omole AE, Bordoni B. Anatomy, Thorax, Lungs. [Updated 2024 Apr 20]. In: StatPearls [Internet]. Treasure Is- land (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470197/.

[6] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012," Int. J. Cancer, vol. 136, no. 5, pp. E359–86,Mar. 2015.

[7] J. P. W. Julie A Barta, Charles A Powell, "global epidemiology of lung cancer," Annals of global health, vol. 85, no. 1, Jan. 2019.

[8] Health365. (n.d.). Stages of lung cancer. Accessed May 25, 2024, from https://www.health365.sg/stages-of-lung-cancer/.

[9] S. Shyamala and M. Pushparani, "Pre-processing and segmentation techniques for lung cancer on ct images," International Journal of Current Research, vol. 8, no. 05, pp. 31 665–31 668, 2016.

[10] Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong KK. Non-small-cell lung cancers: a heterogeneous set of diseases. Nat Rev Cancer. 2014 Aug;14(8):535-46. doi: 10.1038/nrc3775. Erratum in: Nat Rev Cancer. 2015 Apr;15(4):247. PMID: 25056707; PMCID: PMC5712844.

[11] Diaz, Gerald. "Lung Cancer Staging and Diagnosis (Small Cell and Non-Small Cell)." Accessed May 25, 2024. https://www.grepmed.com/images/12465/lung-cancer-staging-diagnosis-smallcell.

[12] N. Duma, R. Santana-Davila, and J. R. Molina, "Non-small cell lung cancer: Epidemiology, screening, diagnosis, and treatment," Mayo Clin. Proc., vol. 94, no. 8, pp. 1623–1640, Aug. 2019.

[13] A. Panunzio and P. Sartori, "Lung cancer and radiological imaging," Curr Radiopharm., vol. 13, no. 3, pp. 238–242, 2020.

[14] D. S. Gierada, W. C. Black, C. Chiles, P. F. Pinsky, and D. F. Yankelevitz, "Low-dose CT screening for lung cancer: Evidence from 2 decades of study," Radiol. Imaging Cancer, vol. 2, no. 2, p. e190058, Mar. 2020.

[15] Radiology Masterclass. "Lung Cancer - Radiotherapy." Accessed May 25, 2024.https://www.radiologymasterclass.co.uk/.

[16] AUSRAD. "Current or Former Heavy Smoker? CT Lung Screening Saves Lives." Accessed May 25, 2024. https://www.ausrad.com/current-or-former-heavy-smoker-ct-lung-screening-saves-lives/.

[17] Young, Lisa Franz, David Nagarkatte, Preeti Fletcher, Christopher Wikenheiser-Brokamp, Kathryn Galsky, Matt Corbridge, Thomas Lam, Anna Gelfand, Michael Mccormack, Francis. (2009). Utility of [F-18]2-Fluoro-2-Deoxyglucose-PET in Sporadic and Tuberous Sclerosis-Associated Lymphangioleiomyomatosis. Chest. 136. 926-33. 10.1378/chest.09-0336.

[18] National Instance for the Protection of Personal Data. (n.d.). Accessed June 3, 2024, from https://www.inpdp.tn/

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv preprint arXiv:1512.03385, Dec. 2015.

[20] Slicer. "Slicer: 3D Slicer Platform." Accessed June 5, 2024. https://www.slicer.org

[21] ITK-SNAP. "ITK-SNAP: Interactive Medical Image Segmentation." Accessed June 5, 2024. http://www.itksnap.org/pmwiki/pmwiki.php

[22] Chen, Sihong, Ma, Kai, and Zheng, Yefeng. "Med3D: Transfer Learn- ing for 3D Medical Image Analysis." arXiv preprint arXiv:1904.00625,2019.

[23] Singh SP, Wang L, Gupta S, Goli H, Padmanabhan P, Gulyás B. 3D Deep Learning on Medical Images: A Review. Sensors (Basel). 2020 Sep 7;20(18):5097. doi: 10.3390/s20185097. PMID: 32906819; PMCID: PMC7570704.

[24] Yan, Jiamiao. (2024). Application of CNN in computer vision. Applied and Computational Engineering. 30. 104-110. 10.54254/2755-2721/30/20230081.

[25] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8.

[26] Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018). 2018 Sep;11045:3-11. doi: 10.1007/978-3-030-00889-5-1. Epub 2018 Sep 20. PMID: 32613207; PMCID: PMC7329239.

[27] Simonyan, K., Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.

[28] Zhang, Shuai and Niu, Yanmin. (2023). LcmUNet: A Lightweight Network Combining CNN and MLP for Real-Time Medical Im- age Segmentation. Bioengineering. 10. 712. 10.3390/bioengineering10060712.

[29] Zhou, Yuepeng Chang, Huiyou Lu, Yonghe Lu, Xili Zhou, Ruqi. (2020). Improving the Performance of VGG Through Different Gran- ularity Feature Combinations. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.3031908.

[30] Sahota,H.(2023).An Intuitive Guide to Convolutional Neural Networks.Comet Blog Retrieved from https://www.comet.com/site/blog/an-intuitive-guide-to-convolutional-neural-networks/.

[31] Medilsys. (2020). The Military Hospital of Tunis. Retrieved from [https://medilsys.com/the-military-hospital-of-tunis/]

[32] Hosna A, Merry E, Gyalmo J, Alom Z, Aung Z, Azim MA. Transfer learning: a friendly introduction. J Big Data. 2022;9(1):102. doi: 10.1186/s40537-022-00652-w. Epub 2022 Oct 22. PMID: 36313477; PMCID: PMC9589764.

[33] The Cancer Imaging Archive, "Lung-PET-CT- Dx," accessed: Jul. 20, 2024. [Online]. Available: https://www.cancerimagingarchive.net/collection/lung-pet-ct-dx/

[34] P. Bajpai, A. Ghosh, S. Gupta, and M. K. Tiwari, "AI-powered decision support systems for precision medicine: A review and perspective," BMC Medical Informatics and Decision Making, vol. 24, p. 253, 2024, doi: 10.1186/s12911-024-02553-9. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911- 024-02553-9.

[35] N. Missaoui, S. Hmissa, H. Landolsi, S. Korbi, W. Joma, A. Anjorin, S. Ben Abdelkrim, N. Beizig, and M. Mokni, "Lung Cancer in Central Tunisia: Epidemiology and Clinicopathological Features," Asian Pacific Journal of Cancer Prevention (APJCP), vol. 12, pp. 2305-2309, 2011.

[36] T. Hamdeni, "Lung Cancer RECIST PFS/OS data," Mendeley Data, V1, 2023, doi: 10.17632/rxsw3f69yc.1. [Online]. Available: https://data.mendeley.com/datasets/rxsw3f69yc/1.