IOSUD - Universitatea Politehnica Timişoara
Şcoala Doctorală de Studii Inginereşti

PATTERNS IN BIOINFORMATICS

Phd thesis – Abstract
For obtaining the scientific title of PhD at
Politehnica University Timișoara
In the field of Computers and Information Technology

author eng. Laura BROASCĂ
scientific supervisor Prof.Univ.Dr.eng. Horia CIOCÂRLIE

Abstract

The thesis is structured into 6 chapters. **Chapter 1** presents a short introduction into the domain of complex networks and of patterns Bioinformatics. More precisely, it approaches two directions on pattern analysis within the domain of Bioinformatics, based on complex networks: visual patterns – that can be observed and analysed with the help of visualization tools, and also numerical patterns – highlighted through the modeling, visualization and analysis of medical data and pulmonary pathologies.

Chapter 2 presents a study into complex networks, theoretical concepts, as well as the newest software algorithms used for complex networks visualization. This chapter ends with defining the chosen niche and the motivation for building a solution which would cover the missing pieces in the current field of visualizing and modeling complex networks.

Chapter 3 presents the proposed technical solution – a software application which is capable of transforming biological systems into complex networks that can be viewed in an innovative way – 2D and 3D graphics according to the user necessities so as to allow the highlighting of visual patterns in an efficient way. The algorithm is then assessed according to multiple criteria and then compared to other state-of-the-art algorithms, showing the benefits of using this application.

Chapter 4 deals with modeling numerical patterns from HRCT data (High Resolution Computer Tomography) coming from healthy and ill patients (suffering from Diffuse Interstitial Lung Diseases – DILD). For this purpose, I present a new algorithm developed for the transformation of HRCT images into complex networks that are able to model 3 types of pulmonary tissue. These types of tissue are the key to a fast and efficient medical treatment. The complex networks created through this algorithm are then analyzed from a pixel distribution point of view and then the proper mathematical functions are selected to fit the distributions. These functions make a difference between healthy and affected patients. Thus, the potential of this algorithm is showcased.

Chapter 5 is based on the algorithm described in chapter 4, and starting from the same model of affected pulmonary tissue (DILD), I propose a new method for the early detection of changes in the lung, as well as a new way of measuring illness progression speed. The proposed hypotheses are validated through some T-tests, demonstrating the potential of such algorithms in the process of medical diagnosis, based on two types of pulmonary lesions: GGO and Consolidation.

Chapter 6 covers the conclusions and the personal contributions on the two proposed directions: visual patterns and numerical patterns in Bioinformatics.

Objectives

The main objectives of this thesis cover two large topics.

Visual patterns and pattern analysis tools in Bioinformatoics based on complex networks:

1. An analysis of current state-of-the-art complex networks visualization tools in order to determine strong points as well as weaknesses in representing specific bioinformatics data.
2. Developing an algorithm based on complex networks which would facilitate visual pattern detection in the biomedical area (medium to large network dimension, with more than 200 nodes and 500 edges, high density). The algorithm will be able to function in 3 dimensions for a complex visual rendering. The number of dimensions can be 2 and/or 3 based on the particular necessities of the biologists / medical specialists/

bioinformatics specialists.
3. An analysis of the performance and visual quality of the proposed algorithm compared to other state-of-the-art tools.

Numeric patterns and methods of identifying numeric patterns in Bioinformatics:

1. Determining a domain in medicine suitable for a numerical analysis software application based on complex networks
2. Creating an innovative algorithm based on complex networks for processing HRCT lung imaging data.
3. Validating this algorithm from a medical and systems science perspective.
4. Comparing the algorithm with other existing algorithms
5. Integrating the algorithm into a multidisciplinary methodology that leads to the identification of patterns in Diffuse Interstitial Lung Diseases (based on at least one of the three pathological density types of lung tissue: Emphysema, Ground Glass Opacity, Consolidation).
6. Highlighting data patterns that allow differentiation between normal (healthy) and pathological patients.
7. Using the proposed algorithm for early identification of pathological changes in lung tissue.
8. Proposing a metric to evaluate the progression speed of lung damage.

Personal contributions

To achieve the mentioned objectives, I developed two algorithms based on complex networks: an innovative algorithm for visualizing complex networks, as well as a Computer Aided Diagnosis (CAD) algorithm based on complex networks.

The first part of the thesis (Chapter 3) describes the hybrid algorithm for complex networks visualization: a 2D and 3D algorithm designed for displaying and highlighting visual patters that would otherwise be potentially underestimated by other software applications.

Given the above, I have brought the following contributions to the field:

1. An analysis and evaluation of the state-of-the-art tools for complex networks visualization. Published in : [1][6][8]
2. I have proposed and implemented a new complex networks visualization algorithm [8][10]
3. I have tested the algorithm performance against multiple biological data sets of various sizes.
4. I have implemented a new covariance metric (similarity metric) for the network vertices and for their spatial positioning within the network layout. [9]
5. I have compared the algorithm against other popular software aplications for complex networks visualization. [9][10]

Considering the second and third parts of this thesis (Chapters 4 and 5) regarding the Novel Method for Computer Tomography image interpretation, the following contributions have been brought:

1. I participated in the multidisciplinary approach of creating a complex networks method for modeling lung HRCTs published in [23][25];
2. I proposed and implemented the HRCT processing algorithm based on complex networks published in [23];
3. I created and defined a model responsible for layering and analyzing DICOM

images according to the three dimensions: Emphysema, GGO, and Consolidation[23];

4. I processed and curated all data sets used for this approach, after their inclusion in the lot by the medical couterparts;[23]
5. I participated in defining a proper window size (crop dimensions of 65 x 65 pixels) for the analyzed data sets, so as to satisfy two major requirements: medical relevance (covering the basic lung unit – secondary pulmonary lobule) and delivering an adequate performance and throughput;
6. I conducted an in-depth analysis of various radial distance dimensions and the impact or relevance of such values for the final complex networks model of lung tissue [23][25];
7. I defined a similarity metric (delta) based on HU bands (Emphysema, GGO, Consolidations) and validated it against medical data, together with a medical team of specialists [23][25];
8. I identified the mathematical functions fitting the complex networks degree distributions associated with normal lungs and affected lungs [23][25];
9. I conducted an analysis for assessing the accuracy of such mathematical functions in the case of normal lungs and diseased lungs, as well as the extent to which they reflect System Science as well as Medical Science[23][25];
10. I validated the proposed model from a Network Science perspective[111], [139].
11. I analyzed the model for three different complex networks metrics: total count, average degree, and maximum degree[23][25];
12. I participated in defining a new measurement type and mathematical formula for fibrosis progression speed. Based on the classical notion of speed (defined as variation over time), the new relative variation speed formula was proposed[25].

Thesis summary by chapters

1.1. Introduction

The purpose of this thesis is to address some of the problems encountered in the field of Bioinformatics, specifically by making contributions (algorithms/software) based on complex networks that help discover patterns in Bioinformatics in two main directions: visual patterns and numerical patterns.

To this end, the current context was presented for both categories.

The datasets belonging to biological or genetic networks (DNA) have some characteristics that complicate and hinder discoveries in this niche: the datasets are very large and dense in size, but are also incomplete. This has led to the need for specialists to use tools or software applications that can generate as faithful graphical visualizations as possible of these systems, allowing them to clearly observe their structure, communities, as well as the connections or interdependencies between their elements.

In terms of visual patterns in the field of genetics and the existing software tools that can generate visualizations of different types of complex networks, there are several such tools, including Gephi, NetworkX (Python), Pajek, Igraph, etc.

However, although all of these can generate satisfactory representations of complex networks, they have different disadvantages either in terms of performance or in terms of compromising visual appearance. Especially in the field of biology or medicine, the need for rendering highly dense complex networks consisting of thousands of nodes and thousands or tens of thousands of connections is all the more stringent as, for example, in the case of DNA networks, they could help detect some undiscovered links between human genes. As such, compromises made on the graphical part are a disadvantage for specialists seeking visual details. On the other hand, the more detailed and loaded the visual representation, the more processing resources it requires, and this leads to poorer performance and longer generation times.

As a result of collaboration with specialists in genetics who wanted to study genetic anomalies that occur during the embryonic phase through complex networks, the motivation to develop an algorithm that could generate clearer visualizations of these networks in an innovative way arose: in a combined 2D and 3D manner.

In terms of numerical patterns, the niche proposed by the thesis is the analysis of data from pulmonary HRCTs through representation and modeling in the form of complex networks, and the detection of mathematical patterns that can characterize and differentiate healthy patients from those suffering from Diffuse interstitial lung diseases (DILD).

The need of pneumology specialists is to detect changes in lung tissue as quickly as possible so that they can intervene and treat pathologies more efficiently, given that some of these pathologies can have a virulent evolution and can cause death in a very short time. The medical diagnosis of these DILD is predominantly based on medical imaging, which can provide a clear representation of the lungs, corroborated with other parameters and medical tests. However, in terms of interpreting HRCTs, this process can be subjective, as the human eye can miss details of very small dimensions (a few pixels).

There are currently some software applications capable of identifying patterns through the analysis of HRCTs, but they are either still in an incipient state (Zrimec et al.), too complex, or costly from a financial standpoint (CALIPER). In this regard, the thesis proposes a new approach based on the analysis of numerical patterns observed in HRCTs, which can be helpful in consolidating diagnoses given by doctors in the case of DILD.

1.2. Theoretical background

Chapter 2 provides an overview of the most important theoretical concepts in the field of complex networks and software algorithms aimed at identifying visual patterns.

In the area of complex networks, several fundamental concepts are presented to understand the work at hand. Thus, the basic concepts underlying the characterization of a complex network are listed: degree distribution, clustering coefficient, modularity, or network density. At the same time, a classification of the most used metrics associated with complex networks is presented (degree centrality, betweenness centrality, closeness centrality), as well as the types of network models: the Albert Barabasi model, the Erdos-Renyi model, and the Watts-Strogatz model. Regarding clustering algorithms, some of the most used algorithms are listed, such as the Louvain Algorithm, the Leiden algorithm, or K-means clustering.

With regards to state-of-the-art algorithms in the niche of complex networks representation, some of the most used are presented, such as Force Atlas 2, Fruchterman-Reingold, OpenOrd, Yifan-Hu, Kamada-Kawai. Regarding applications that implement these algorithms, among these are Gephi (which implements all the algorithms listed above), NetworkX (developed in Python), iGraph (developed in the R language), Cytoscape, and Nodetrix.

At the end of Chapter 2, I define the niche addressed in this thesis. A comparison (in terms of performance but also visually) is made between the visualization applications listed above, and the directions for improvement and the criteria that a new algorithm developed by me should cover are identified: from a visual point of view, the algorithm should provide a clear representation of the complex network structure, with the ability to identify components and singular elements, which can process networks of medium and large dimensions, and which can function in a reasonable time, ideally better than the other applications.

1.3. Visual patterns in Bioinformatics

A hybrid algorithm for complex networks visualization in 2D and 3D

Chapter 3 presents a new algorithm based on complex networks for highlighting visual patterns, applied to biological or genetic networks. This algorithm was developed in Octave, a free version of the popular MatLab.

First, the criteria that can influence the graphical generation of complex networks and that this algorithm aims to meet are presented:

- Relevant and clear structural representation,
- Eye-catching graphics and color schemes for human perception
- Moderate resource consumption that allows for wide usage
- Acceptable execution time even though it depends on the size of the data set

The concept of the proposed algorithm is presented here: generating graphical visualizations of complex networks (in the field of biology) in which communities are distributed in a 3D space, but, depending on the user's needs, the components can be displayed in both 2D and 3D in a differentiated way.

Another innovative approach is the introduction of a graded reference system (xOyz axes) that provides a visual indicator of the distances between network elements, community sizes, and conveys a more precise idea of hierarchy between components.

The visualization structure is intended to be easy to understand, reduce overlaps between graph edges, and have well-defined communities.

The first published version of the algorithm is based on distributing the entire graph over a paraboloid composed of concentric circles, in which the network nodes are dispersed along the circles. The hierarchy is based on the degree of nodes, which will be associated with the radii of the circles. Nodes with a higher degree of connectivity will be positioned higher on the Oz axis, while nodes with a lower degree will be placed on the circles below. However, the drawback of the first version of the algorithm is that grouping nodes into distinct communities is visually difficult, and the resulting figure is congested. As such, this type of visualization does not meet the initially mentioned criteria.

Therefore, I have developed an improved version of the algorithm based on the following rules:

- Each community will be distributed on its own paraboloid in a 3D space.
- Within each community, nodes will be placed on concentric circles according to the rule: nodes with the highest degree are placed towards the top of the paraboloid (the position of the node on the Oz axis is given by the degree of the node, as well as the radius of the circle on which it will be placed)
- The node will be oriented towards the external communities it is most interconnected with – positioning based on the weighted center of gravity of the other communities it is connected to.
- Within its own community, the node will be placed closer to other nodes with which it has a high similarity (the similarity between two nodes is given by the proportion of neighbors they have in common).

The algorithm structure has 3 phases::

- Preparation of the dataset: complex network files from the field of biology divided into communities using the Louvain algorithm (edge lists).
- Calculation of the graph structure and determination of the spatial positioning of nodes and communities.
- Generation of the visualization based on the spatial structure established earlier.

Algorithm steps

The algorithm first parses the files containing the edge lists and stores or calculates information about each node: connectivity, degree, and the community to which it belongs.

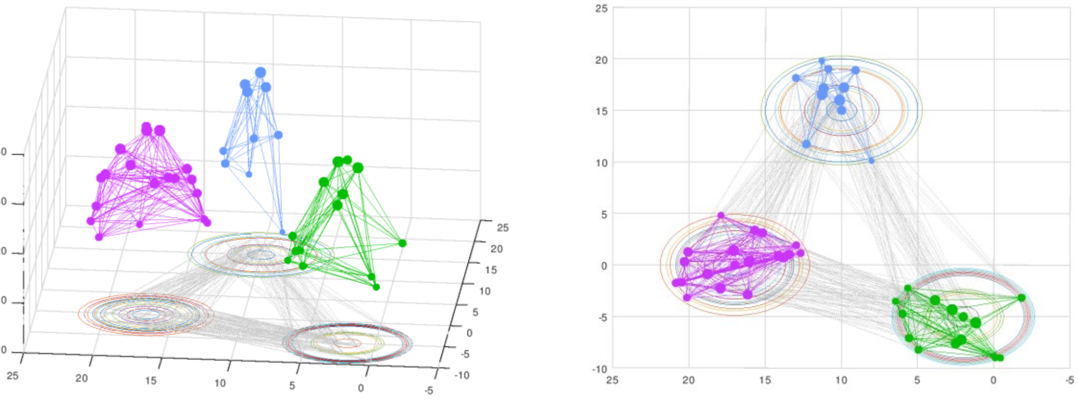
Then, a adjacency matrix is created to describe the graph. The algorithm then calculates the covariance of the nodes (the criterion that gives the degree of similarity) based on the adjacency matrix and stores it.

The 3D positioning of the nodes within a community is calculated based on the number of connections of the node to external communities, the weighted centroid of the communities to which the node is connected, the connections and similarity of the node with other nodes within the same cluster.

After calculating the two potential positions for placing the node on one of the concentric circles that form the support paraboloid for the current node's community, the position is adjusted proportionally with the co-expression between the current node and the previously placed node - to avoid creating a cluster of nodes in the same 3D area.

After calculating the positions of all nodes in the network, a visualization is generated with the precalculated nodes and edges. The user has the option to specify whether each of the communities should be displayed in 3D, or only one of them should be in 3D while adjacent communities are plotted in 2D. Additionally, there is the possibility of displaying the edges within clusters in 3D, while keeping the edges between communities in two dimensions to simplify the image.

Results



a) b)

Fig 1. 3D Visualization – communities in 3D, inter-cluster edges in 2D
a) 3D view b) View from above

Discussion

The proposed algorithm aims to be an improved alternative to existing applications in the field. Compared to Gephi or NetworkX (Python), the advantages consist of the visual clarity of the network structure, due to the separate distribution of communities and the elimination of some visual elements that can clutter the final image (inter-cluster edges are kept in 2D).

There is the advantage of a visual reference system (graded), which does not appear in any of the other mentioned applications (Gephi, NetworkX, iGraph, R, etc.) and which provides valuable information to the user about the ratio between the elements of the graph, as well as their hierarchy or size. This gives the network a true 3D aspect, unlike other software tools. The graph can be rotated along the 3 axes and zoomed in for better visualization.

Compared to other algorithms, in terms of performance, the results of the hybrid algorithm are better in some aspects. The runtime is proportional to the size of the network, but where the network density is very high, the proposed algorithm offers better execution times.

I compared the algorithms from several points of view: visual aspect, speed, relevance, node distribution, the existence of a clear reference system, the option to integrate user-defined metrics, adaptability, the ability to integrate machine learning algorithms, user-friendliness, interactivity - and the cumulative result of the score on these criteria has shown that the hybrid algorithm is indeed a viable, if not better, alternative to existing software applications, depending on the user's needs.

1.4. Modeling numeric patterns

4.1. A novel method of Computer Tomography image interpretation

Chapter 4 presents a brief introduction to Diffuse Interstitial Lung Diseases (DILD) and the challenges that the doctors face in diagnosing and efficiently treating these types of illnesses. Medical diagnosis, in this case, is largely based on medical imaging such as X-rays or HRCTs. However, this process is inherently subjective as identifying affected areas of the lungs relies on the visual observation of radiologists, and details can make a major difference.

To help improve diagnosis, there are currently a number of software applications that can analyze and interpret pulmonary medical HRCTs. Some of these are still in an early stage (Zrimec et al.), while others require a suite of other pulmonary tests to generate a CT analysis (CALIPER), and their use requires a paid license. Another disadvantage of these solutions is that they cannot quantify the degree or extent of lung lesions, but only classify the diagnosis without offering nuances.

Given the challenges and the need for more comprehensive solutions for medical diagnosis support in PID cases, I proposed and implemented an algorithm that transposes portions of the lung image (taken from a CT) into a complex network and analyzes it to highlight patterns in the types of lung tissue. The image is first divided into 3 layers, depending on the type of tissue: Emphysema, Ground Glass Opacity, Consolidation.

The result obtained is 3 separate images that contain only pixels belonging to one of the 3 types of lung densities. These images are then transformed into complex networks in which pixels are equivalent to nodes, and the connections between them are given by the degree of similarity. After obtaining the 3 complex networks, they are analyzed in terms of node degree distribution, and the types of mathematical functions that best map onto the two categories of patients are observed: patients with normal (healthy) lungs and patients with pathologies belonging to DILD.

Overview

This passage briefly presents the medical patterns on which medical diagnosis is based, in terms of spatial distribution in the lungs, as well as the position in the secondary lung lobe. The distribution of various types of lung tissue with different densities, their extent, and how they combine provide clues about the type of pathology. HRCT (High-Resolution Computed Tomography) is the modality by which specialists can visually analyze these patterns. HRCT is composed of multiple "very thin slices" or X-rays of the body, overlapped at a distance of no more than 1.25 mm. These images are then processed by reconstruction algorithms to produce a detailed and accurate presentation of a body part.

The DICOM format

The DICOM standard is an international standard commonly used for the digital representation and storage of HRCT images. Medical imaging equipment (such as MRI, X-rays, HRCT, and ultrasound) captures information about the patient, converts and stores it in this format. Files of this type have a *.dcm* extension. The standard consists of two sections: a header and the actual content of the file. The header stores metadata about the parameters of the equipment used to capture the sample, patient information, image resolution data, or data about the patient's position during the procedure. This information is necessary for software that reads the images to correctly interpret the rest of the content.

The actual content of the file consists of a binary sequence that describes the pixels of the image and their intensity.

Datasets

The data sets were selected by the team of specialists at the Victor Babeș Infectious Diseases and Pneumophthisiology Hospital in Timișoara. The total number of patients was 60. These CTs were divided into two categories: 30 cases of healthy lungs (control group) and 30 cases of patients with lungs suffering from ILD.

The inclusion criteria for patients in this study include the condition that a patient must have been diagnosed by at least 3 pulmonology specialists with a minimum of 5 years of experience in IPF, each CT must have the same characteristics and quality established for the entire batch, all patients must have annual imaging investigations, all must have a suite of investigations in their portfolio, such as DLco, FEV, or details of clinical evaluation and their outcome, all CTs must have been annotated by medical experts with indications and description of the affected lung areas.

Exclusion criteria include lack of written consent for use in scientific studies, lack of annual recurrent investigations, examples of CT images with quality below that established for this study, presence of other associated pathologies.

Algorithm overview

The algorithm steps consist in:

1. Dataset selection and preparation
2. Preprocessing CTs by selecting an affected lung sample (65 x 65 px) for each patient
3. The algorithm converts the pixels of each sample into its Hounsfield unit equivalent:
    a. Each sample is considered a pixel
    b. It converts the gradient of each pixel into its Hounsfield equivalent (HU)
    c. It copies the Hounsfield values into new matrices for every type of affected tissue: Emphysema, Ground Glass Opacity, Consolidation.
4. Each of the 3 matrices are converted into an adjacency matrix corresponding to a complex network by the following rules:

   a. Each pixel is considered a network node
   b. The edge connecting two nodes is created only if:
      i. The radial distance between the two is $rd \leq 4$
      ii. The Hounsfield gradient difference between the two pixels is $\Delta \leq 50$
      iii. The HU values of the two pixels pertain to the same Hounsfield band (the same type of affected tissue)
5. The 3 matrices are then analyzed from a mathematical perspective, defining functions to approximate the node degree distribution for all 3 networks.
6. The algorithm is executed for all CTs in the two sets: normal patients and DILD patients.
7. The resulting mathematical functions are then compared in order to determine similarities and differences.

Defining the sample

The chosen size for the analyzed samples was 65x65 pixels. Following the analysis of existing studies in the field and discussions with pulmonary specialists, this size was decided upon, which corresponds to their requirements from multiple perspectives: this area should be large enough to include the basic functional unit of the lung (SPL - secondary pulmonary lobule), which has a size between [1 cm; 2.5 cm]. Converting this size into pixels is done through a formula that depends on the pixel spacing (PS). For this experiment, PS = 0.74, so through a short calculation, the minimum size of the sample is 33.7837 px. However, this does not guarantee a perfect fit for an LPS, so we decided on a size almost double the minimum: 65 x 65 px.

Radial distance

We define radial distance (rd) as the maximum distance (Euclidean linear distance) at which two nodes can be located to be considered connected. Following an experiment in which I evaluated the complex network structures resulting from radial distances ranging from 1 to 8, I found that rd = 4 is the ideal distance with which the algorithm generates complex networks that can most faithfully represent the initial image. The chosen value of rd = 4 also corresponds to other studies in the field that refer to the minimal detectable lesion as having dimensions of 3-17 mm. Indeed, converting from pixels to physical size, 4 x 0.74 = 2.96 mm - this is the minimum size of the lesion that the proposed algorithm can represent, and it conforms to existing studies.

Hounsfield bands

Hounsfield Units are the basic units for quantifying X-ray absorption and attenuation by human tissue. Specialists use these units when they want to confirm certain diagnostic suspicions that would otherwise be difficult to evaluate visually. The HU value spectrum varies between -1000 HU in the case of air and 3000 HU in the case of hard metals like steel. For the human eye, these values are associated with a certain shade of gray (on a scale from white to black). For this study, three HU bands (three intervals) were chosen which correspond to three types of lung lesions: Emphysema [-1024, -977), Ground Glass Opacity [-703, -368), and Consolidation [-100, 5). An affected lung may have a combination of these lesions in different proportions.

Node similarity based on gradient delta

One of the conditions for two pixels to be considered connected is that the gradient difference between them should be at most 50 units. This value was chosen based on two criteria: the range of HU band values being very large, which requires dividing them into sub-intervals to better differentiate between pixels even within the same band. Secondly, from a biological and medical point of view, two points with a gradient difference greater than 50 do not represent the same type of tissue or the same degree of pulmonary damage, so this type of differentiation is necessary.

The formula by which the gradient of a pixel is converted to a Hounsfield value depends on the values of the medical equipment used:

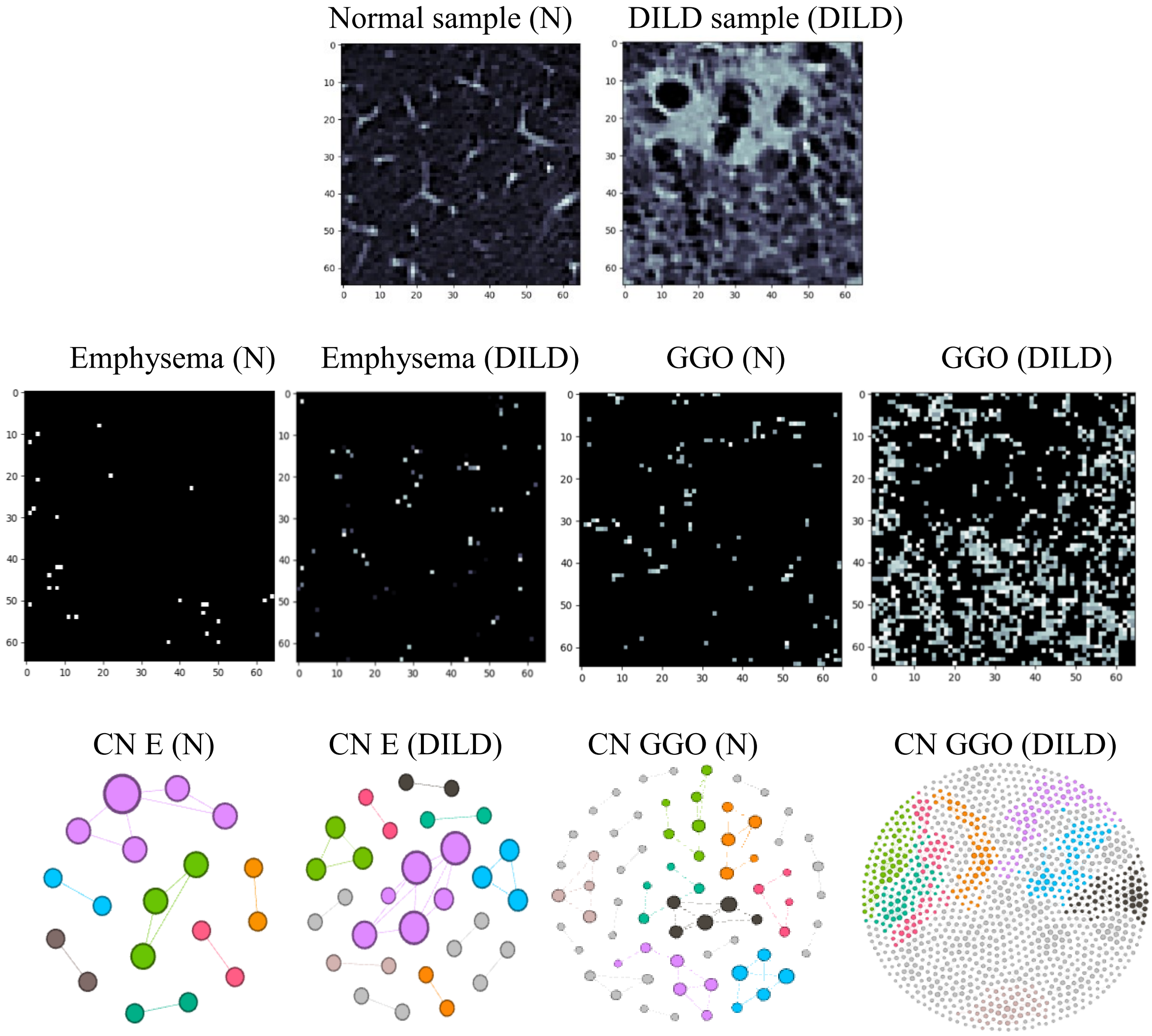
pixel_hu_value = pixel_value * RescaleSlope + RescaleIntercept

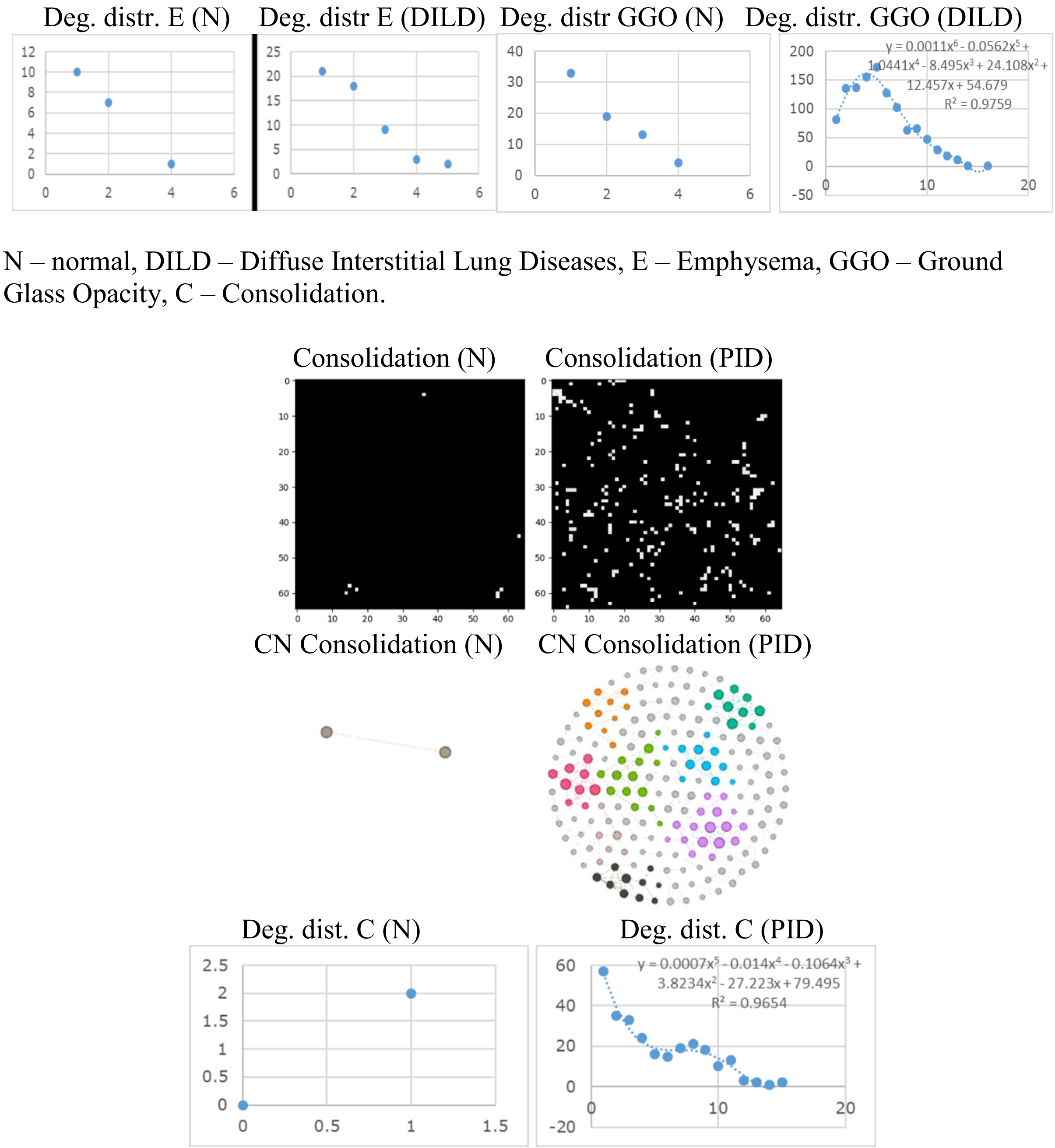
where RescaleSlope și RescaleIntercept are parameters stored in the DICOM header.

Converting HU bands into complex networks

The algorithm creates an adjacency matrix for each of the 3 selected HU bands (Emphysema, GGO, Consolidation) based on two conditions: two nodes are connected if they are within a maximum distance of 4 pixels and if the gradient difference between them is no more than 50 units.

Results


Normal sample (N)
DILD sample (DILD)
Emphysema (N)
Emphysema (DILD)
GGO (N)
GGO (DILD)
CN E (N)
CN E (DILD)
CN GGO (N)
CN GGO (DILD)



N – normal, DILD – Diffuse Interstitial Lung Diseases, E – Emphysema, GGO – Ground Glass Opacity, C – Consolidation.

Fig. 2 Converting a DICOM sample into complex networks on three dimensions: E, GGO, C

Discussion

As a result of the analysis, the function that best approximates normal lungs is logarithmic. Previous research in the field has shown that these types of ecosystems fall into the category of logarithmic or power law functions. In this case, however, since it is a biological system without a feedback loop, it indeed fits a logarithmic function.

In terms of diseased lungs, the distribution of degrees for these best fits with a polynomial function. Previous studies from the literature confirm this type of function, but it is also validated by the characteristic of PID pathologies to be proliferative processes, despite not being caused by a virus. It is worth noting here that the results do not show a single type of polynomial function that can characterize all cases of PID, so a conclusion regarding the maximum degree of the function cannot be drawn. This may be explained by the fact that PID pathologies are a combination of the 3 HU bands, in different proportions, which requires a more detailed study.

The thesis also presents the medical validation of these results, given that the entire study was developed in collaboration with the team of medical specialists from the Victor Babeș Infectious Diseases and Pneumophthisiology Hospital in Timișoara.

Furthermore, the thesis includes a statistical validation of the algorithm's results and analysis by calculating a T-test applied to the two sets of patients (normal and PID). This T-test confirms that the observed differences between the two data sets are not random.

Compared to other methods of CT analysis and evaluation (Zrimec et al, CALIPER, machine learning), the complex network-based method proposed here proves to have advantages from several points of view, such as being a method that covers 3 dimensions (E, GGO, C), being an analytical method, and offering a quantitative evaluation of lung lesions. Among the disadvantages, it is worth mentioning the need for a user-friendly graphical interface or the dependence on a pre-processed dataset with the help of radiologists.

1.5. Numeric patterns analysis

5.1. Enhancing Imagistic Interstitial Lung Disease Diagnostic

The algorithm proposed in chapter 4 is further used to provide even more clues necessary for an accurate medical diagnosis. Early detection is one of the pressing needs in PID lung pathologies. Currently, however, specialists visually evaluate successive CT scans of the same patient to detect the progression of these pathologies, but even this process is subjective and involves observing details of granularity that may exceed the human eye's capacity. For a more accurate diagnosis, doctors also rely on other pulmonary function tests.

Computer-aided diagnosis is a developing approach that attempts to integrate various computational techniques and algorithms to obtain the most accurate diagnoses possible. Of the current applications, CALIPER is capable of analyzing a single HRCT, but cannot correlate the results of several to provide a measure of disease progression.

This shortcoming opens up an unexplored niche for approaches such as pattern matching or complex networks.

The two major points that the algorithm described here aims to address are:

1. Quantitative evaluation of DILD progression using a complex network-based model
2. Contributing to early detection assistance of such pathologies

The analyzed subject group consists of 96 HRCT scans, divided into 31 healthy patients and 65 patients with a series of 2, 3, or even 4 CT scans taken at different times (different years). All patients were previously evaluated by specialist doctors and presented signs of multiple PID: sarcoidosis (S), idiopathic pulmonary fibrosis (IPF), non-specific interstitial pneumonia (NSIP). These pathologies are a combination of the three types of lung tissue analyzed by the proposed algorithm: Emphysema (E), Ground Glass Opacities (GGO), and Consolidation (C).

The thesis goes on to describe the medical specificities of these three tissue types and how they appear visually in an HRCT.

The data sets were processed with the algorithm described in Chapter 4, following the same steps:

1. Multiple 65 x 65 px samples are taken for each patient with affected lungs.
2. The pixels are then converted into their HU equivalent
3. The image is split into 3 layers: E, GGO, C
4. The resulting images are converted into adjacency matrices and complex networks
5. The specificities are then analyzed according to the proposed metrics

Relevant metrics

The resulting complex networks must respect the biological characteristics of the lungs in order to accurately describe the evolution, shape, and density of the affected tissue. Therefore, we chose the following complex network metrics for evaluation: maximum degree, total degree, and average degree. From a medical point of view, these metrics represent the following features: total degree - the cumulative effect on the entire lung section, average degree suggests the location of the lesions, while maximum degree shows the maximum intensity of a lesion.

Lesion progression

The proposed function for evaluating the progression of lung lesions was inspired by the well-known formula for velocity in physics. However, it was defined as a relative velocity, because a patient can realistically only be compared to themselves in terms of disease progression.

$$v = \begin{cases} \dfrac{(s - s_0)}{s_0 \times t}, for\ s_0\ != 0 \\ \dfrac{s}{t} \qquad\qquad, otherwise \end{cases}$$

where *s* is the evaluated metric (total degree, average degree, maximum degree), *s0* is the initial point used for normalization, and t is a measure of time expressed in years, since patients with PID undergo annual check-ups.

Results

As relevant examples, we have selected two cases of patients with DILD whose progression speed was evaluated and compared using the three metrics mentioned based on consecutive annual imaging evaluations. The results showed the progression pattern of the three HU bands (E, GGO, C) based on the three metrics, compared to the pulmonary function tests performed by medical specialists during their medical analyses.

To validate the results obtained, we also calculated a T-test (statistical method). This test showed that the relative progression speed detected using the proposed method, compared to clinical evaluations (DLco), is valid for GGO and C bands, but not for the E band.

Furthermore, to validate the hypothesis of early detection of these types of pathologies, patients were divided into two categories: normal patients and patients with early-stage PID and acceptable functional parameters. We created a context where the proposed method could detect early signs of DILD.

The evaluated metrics are still the three mentioned earlier (maximum degree, total degree, average degree). Based on these metrics, another T-test was conducted, and the results showed that in the case of the GGO and C bands, the results are valid, but not in the case of the Emphysema band.

Discussion

Given the results of the T-tests, it can be concluded that this CT analysis method can accurately and quantitatively evaluate the progression rate of lung lesions in two HU bands (GGO and Consolidation). However, in the case of the Emphysema band, it cannot detect significant variations, or at least ones that can be correlated with the functional tests (DLco) of the patients.

In the case of the second proposed hypothesis, that of early detection, the proposed method demonstrates that on the GGO bands (all three metrics) and C (two of the metrics), notable statistically significant differences can be observed between normal patients and borderline patients, but not in the case of the Emphysema band.

In conclusion, with regard to the second proposed hypothesis, the complex network model has proven to be effective in detecting early changes on the GGO band, partially valid in the case of the Consolidation band, and false in the case of the Emphysema band.

Conclusions

1. Conclusions – Visual Patterns (Hybrid 3D Network Layout Visualization Algorithm)

The Hybrid layout proposed in Chapter 3 aims to become a viable and appealing choice when talking about 2D and 3D graph layout solutions. This type of algorithm has proved that it can successfully fill the gap or the missing piece in network software which was until now lacking in terms of network structure, from a visual point of view. This approach offers a special type of view (2D and 3D) into the network elements, and also considers the user's point of interest (centering the user's cluster of interest as a 3D entity). It also enhances the network view with a gridded aspect, which gives a more practical approach to graph representation, offering clear quantifiable visual queues (graded xOyz axes). 3D graph interaction and manipulation are also possible with this solution, giving users the possibility to turn the resulting image on all axes.

The additional third dimension (Oz axis) offers more space for a better node distribution across the entire 3D canvas, as well as providing a layering of elements, giving a different significance to each of them.

Further improvements aim at improving generation time, especially in terms of edge plotting, as well as reducing edge crossings, thus enhancing graph readability.

2. Conclusions – Modeling numeric patterns (A novel method for Computer Tomography image interpretation)

A novel complex networks-based method that transforms and interprets HRCTs has been developed and tested. This approach analyzes medical data in a three dimension manner, involving mathematical function fitting. The overview and algorithm development sections in Chapter 4 describe the algorithm stages in a detailed manner.

There is a solid argumentation regarding the analyzed sample size (65 x 65 px) and a comparison with existing field tools justifies taking a step further and enlarging the interest area. Sample dimensions are also consistent with the anatomical details (secondary pulmonary lobule).

Vertex connectivity and the associated radial distance selection are supported by an extensive experiment regarding pixel similarity criteria corroborated with complex networks attachment principles, as well as evaluating the role of network density and clusterization in the process of medical diagnosis.

The proposed algorithm uses Hounsfield Unit intervals both for image layering and simultaneously as similarity criteria for potentially linked nodes, allowing for more granularity when observing lung injuries. These ranges are also dependent on the device and resolution of the machine used to perform the HRCTs.

The results section presents a full algorithm execution and its comparative outcome for two sample patients, a normal one (baseline sample) as well as a pathological one.

Furthermore, the discussion section justifies the coherence and correctness of the complex networks-based algorithm from a Systems Science standpoint, by entailing the metric of degree distribution as a central device for system representation. We also showcase clusterization as a network measurement which shows distinct discrepancies between the two studies HRCT lots: healthy (normal) and pathological patients. From a Medical Science viewpoint, the model is validated by its faithful and fine-grained representation of clinical data and this can prove to be crucial in the diagnosis process.

Finally, the comparisons with other present days tools underline the advantages of using the proposed method: offering a comprehensive measuring instrument for qualitative and quantitative analysis.

Among the drawbacks, we mention its inability to work as off-the-shelf software yet, as well as the particularly modest lot size used for testing it. Improvements regarding the aforementioned are to be addressed in future research, with a much larger training set, as well as user-friendly customizations (a graphical user interface) which would offer it a more appealing look.

In conclusion, the new complex networks algorithm has been shown to be extremely useful in the DILD diagnosis process, by transforming lung HRCTs into quantifiable and qualifiable structures.

3. Conclusions – Numeric Patterns Analysis (Enhancing Interstitial Lung disease diagnostic)

This complex networks approach has been developed as a means of support for the process of diagnosing and managing DILDs. For this purpose, there were two hypotheses being evaluated: early detection and accurately evaluating disease progression, as these are the two main factors that medical specialists have been struggling with. Especially when it comes to IPF, for instance, existing techniques or technologies in the field have, so far, not been able to provide effective answers.

For the first of the proposed hypotheses, regarding progression, the proposed complex networks algorithm and overall approach have proven to be a success. Given its precision and 3 mm granular lesion detection, this approach has shown a very good connection with the clinical symptoms at such a level that could not be reached by the usual functional tests. It is worth noting here that the Emphysema layer constitutes an exception to this conclusion (average count measurement), however, this is easily surpassed by the other five metric axes.

As for the second hypothesis, regarding early detection, the best results have been for the GGO and Consolidation band. From a medical perspective, the GGO and Consolidation layers and their expression are very important in detecting common DILD states. This is an essential capability demonstrated by an algorithm of this kind, and is very fitting to DILD, unlike other software in the field such as Caliper.

In terms of challenges that are still considered for improvement, this approach still takes a fair amount of time, which is directly correlated to the dimension of the analyzed window. Added to this is the time spent preprocessing the HRCT slices, but this can be overcome through the use of other CAD rather than manually.

Future improvements involve integrating this algorithm into a larger software solution and combining swifter ML segmentation and pattern recognition competencies with the steadier but more granular and precise complex networks in-depth analysis.\

Selective bibliography

1. V. M. Ancusa and L. Broasca, “A Method to Pinpoint Undiscovered Links in Genetic and Protein Networks,” Stud. Health Technol. Inform., vol. 210, pp. 771–775, 2015.
2. Y. Li and L. Chen, “Big biological data: challenges and opportunities.,” Genomics Proteomics Bioinformatics, vol. 12, no. 5, pp. 187–189, Oct. 2014, doi: 10.1016/j.gpb.2014.10.001.
3. H. Jeong, Tombor, R. Albert, Z. Oltvai, and A.-L. Barabási, “The large-scale organization of metabolic networks,” Nature, vol. 407, pp. 651–654, 2000.
4. R. Albert and A.-L. Barabási, “Topology of Evolving Networks: Local Events and Universality,” Phys. Rev. Lett., vol. 85, pp. 5234–5237, 2000.
5. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” J. Stat. Mech. Theory Exp., vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/p10008.
6. L. Broască, V. Ancuşa, and H. Ciocârlie, “Bioinformatics Visualisation Tools: An Unbalanced Picture,” Stud. Health Technol. Inform., vol. 228, pp. 760–764, 2016.
7. O.-H. Kwon, T. Crnovrsanin, and K.-L. Ma, “What Would a Graph Look Like in this Layout? A Machine Learning Approach to Large Graph Visualization,” IEEE Trans. Vis. Comput. Graph., vol. 24, no. 1, pp. 478–488, Jan. 2018, doi: 10.1109/tvcg.2017.2743858.
8. L. Broasca, V.-M. Ancusa, and H. Ciocarlie, “A Qualitative Analysis on Force Directed Network Visualization Tools in the Context of Large Complex Networks,” in 2019 23rd International Conference on System Theory, Control and Computing (ICSTCC), Oct. 2019, pp. 656–661. doi: 10.1109/ICSTCC.2019.8885641.
9. L. Broască, V.-M. Ancușa, and H. Ciocârlie, “Towards a Hybrid Layout for Complex Networks Visualization,” in 2020 24th International Conference on System Theory, Control and Computing (ICSTCC), 2020, pp. 118–123. doi: 10.1109/ICSTCC50638.2020.9259656.
10. L. Broasca, V. Ancusa, and H. Ciocarlie, “A 3D Surface Fitting Layout for Complex Networks Visualization,” Stud. Health Technol. Inform., vol. 272, pp. 362–365, Jun. 2020, doi: 10.3233/SHTI200570.
11. A. Christe et al., “Computer-Aided Diagnosis of Pulmonary Fibrosis Using Deep Learning and CT Images.,” Invest. Radiol., vol. 54, no. 10, pp. 627–632, Oct. 2019, doi: 10.1097/RLI.0000000000000574.
12. L. Vulliard and J. Menche, “Complex Networks in Health and Disease,” Syst. Med., pp. 26–33, 2021, doi: 10.1016/B978-0-12-801238-3.11640-X.
13. A. Trușculescu, L. Broască, V. M. Ancușa, D. Manolescu, E. Tudorache, and C. Oancea, “Managing Interstitial Lung Diseases with Computer-Aided Visualization,” in Hybrid Artificial Intelligence and IoT in Healthcare, A. Kumar Bhoi, P. K. Mallick, M. Narayana Mohanty, and V. H. C. de Albuquerque, Eds. Singapore: Springer Singapore, 2021, pp. 245–271. doi: 10.1007/978-981-16-2972-3_12.
14. J.-Z. Cheng et al., “Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans,” Sci. Rep., vol. 6, no. 1, p. 24454, Apr. 2016, doi: 10.1038/srep24454.
15. W. Shi et al., “A deep learning-based quantitative computed tomography model for predicting the severity of COVID-19: a retrospective study of 196 patients.,” Ann. Transl. Med., vol. 9, no. 3, p. 216, Feb. 2021, doi: 10.21037/atm-20-2464
16. L. Li et al., “Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy,” Radiology, vol. 296, no. 2, pp. E65–E71, Aug. 2020, doi: 10.1148/radiol.2020200905.
17. M. Molina-Molina et al., “Importance of early diagnosis and treatment in idiopathic

pulmonary fibrosis," Expert Rev. Respir. Med., vol. 12, no. 7, Art. no. 7, Jul. 2018, doi: 10.1080/17476348.2018.1472580.

18. A. A. Trusculescu, D. Manolescu, E. Tudorache, and C. Oancea, "Deep learning in interstitial lung disease-how long until daily practice," Eur. Radiol., vol. 30, no. 11, Art. no. 11, Nov. 2020, doi: 10.1007/s00330-020-06986-4.
19. Q. Li, W. Cai, and D. D. Feng, "Lung image patch classification with automatic feature learning," Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf., vol. 2013, pp. 6079–6082, 2013, doi: 10.1109/EMBC.2013.6610939.
20. T. Zrimec and S. Busayarat, "Computer-aided Analysis and Interpretation of HRCT Images of the Lung," 2011. doi: 10.5772/14507.
21. A. Depeursinge, T. Zrimec, S. Busayarat, and H. Müller, "3D Lung Image Retrieval Using Localized Features," SPIE Med. Imaging 2011 Lake Buena Vista Orlando Fla. U. S., Oct. 2011, doi: 10.1117/12.877943.
22. G. V. L. de Lima, T. R. Castilho, P. H. Bugatti, P. T. M. Saito, and F. M. Lopes, "A Complex Network-Based Approach to the Analysis and Classification of Images," in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Cham, 2015, pp. 322–330. doi: 10.1007/978-3-319-25751-8_39.
23. L. Broască et al., "A Novel Method for Lung Image Processing Using Complex Networks," Tomogr. Ann Arbor Mich, vol. 8, no. 4, pp. 1928–1946, Jul. 2022, doi: 10.3390/tomography8040162.
24. R. Grassi et al., "COVID-19 pneumonia: computer-aided quantification of healthy lung parenchyma, emphysema, ground glass and consolidation on chest computed tomography (CT)," Radiol. Med. (Torino), vol. 126, no. 4, pp. 553–560, Apr. 2021, doi: 10.1007/s11547-020-01305-9.
25. A. A. Trușculescu et al., "Enhancing Imagistic Interstitial Lung Disease Diagnosis by Using Complex Networks.," Med. Kaunas Lith., vol. 58, no. 9, Sep. 2022, doi: 10.3390/medicina58091288.