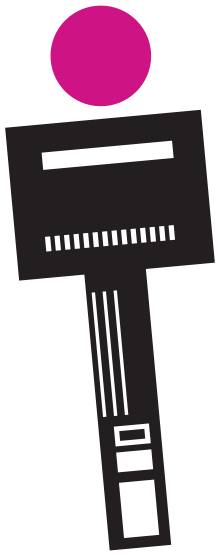

RAM ROBOTICS AND MECHATRONICS

DIFFICULTY MEASURE FOR RADIOLOGY CASES WITH USE OF ITEM ANALYSIS AND DEEP LEARNING

M.H. (Meike) van Benthem

MSC ASSIGNMENT

Committee:
prof. dr. ir. C.H. Slump
E.I.S. Hofmeijer, MSc
dr. C.O. Tan
dr. J. Veltman
dr. M. Poel

June, 2022

016RaM2022
Robotics and Mechatronics
EEMCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | TECHMED CENTRE UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Summary

A vast and extensive radiology education is fundamental for the diagnosis, monitoring, and treatment of lung cancer. Image synthesis, with Generative Adversarial Networks (GANs), can be a powerful tool in radiology education by its ability to diversify training cases for medical students and radiology residents. By controlling the image synthesis, images can be produced with a specific difficulty or complexity level that fits the student's level. To achieve optimal personalization of education, the knowledge gap should be defined by a concrete measure. This research focuses on a measure of difficulty for the detection of lung nodules. Item analysis is a statistical method that can give an indication of the difficulty based on the responses of a group of individuals. To automate the calculation of a measure of difficulty, deep neural networks are used to perform item analysis on lung nodule cases. The method is validated by comparing the measure of difficulty with a subjective subtlety score given by experienced radiologists. The ordinal logistic regression analysis shows a statistically significant relationship between the calculated measure of difficulty and the subtlety scores of nodules. A measure of difficulty is defined that has the potential to be applied to image synthesis for the design of computer-assisted learning systems.

Samenvatting

Intensief radiologie onderwijs is fundamenteel voor de diagnose en behandeling van longkanker. Image synthesis met gebruik van Generative Adversarial Networks (GANs) is een veelbelovende techniek voor radiologie onderwijs, vanwege de mogelijkheid om oefencasussen voor radiologie studenten te produceren. Wanneer de image sythesis gecontroleerd kan worden, is het mogelijk om casussen te produceren op een bepaald moeilijkheidsniveau. Om de personalisatie van educatie te optimaliseren, wordt in dit onderzoek een moeilijkheidsgraad voor long nodule casussen gedefinieerd. Item analyse is een statistische methode dat inzage kan geven in moeilijkheid op basis van responsie van een groep individuen. Door gebruik te maken van deep learning kan het berekenen van een moeilijkheidsgraad geautomatiseerd worden. Deze methode is gevalideerd door de moeilijkheidsgraad te vergelijken met een subjectieve subtiliteitsscore gegeven door ervaren radiologen. De ordinale logaritmische regressieanalyse laat een statistisch significant verband zien tussen de berekende moeilijkheidsgraad en de subtiliteitsscores. Dit onderzoek heeft een moeilijkheidsgraad berekent voor long nodule casussen en heeft the potentie om toegepast te worden bij image synthesis.

Contents

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Context . . . 2
1.3 Item analysis . . . 4
1.4 Research objectives . . . 4
1.5 Thesis outline . . . 4

2 Theoretical Background **6**
2.1 Lung nodules in CT scans . . . 6
2.2 Computer Aided Detection for Lung Nodules . . . 6
2.3 Item analysis . . . 7
2.4 Validation . . . 8

3 Method **11**
3.1 Data . . . 11
3.2 Deep Neural Network . . . 14
3.3 Item analysis . . . 15
3.4 Evaluation . . . 17

4 Results **19**
4.1 lung nodule detection network . . . 19
4.2 Item analysis . . . 21
4.3 Ordinal categorical logistic Regression . . . 24

5 Discussion **27**
5.1 Interpretation of the results . . . 27
5.2 Limitations . . . 30

6 Conclusion **32**
6.1 Conclusions . . . 32
6.2 Recommendations and Future Work . . . 32

7 Appendix 1 **34**
7.1 Descriptive values of data set . . . 34
7.2 Additional Methods Binary score and results . . . 36
7.3 Difficulty-Net results . . . 36

Bibliography **42**

1 Introduction

1.1 Motivation

Lung cancer is the third most common cancer in the Netherlands and has caused over 10.000 deaths in 2020. This makes lung cancer together with Chronic Obstructive Pulmonary Disease (COPD) and Cardiovascular Disease (CVD) part of the so-called Big-3 diseases that are in the top ten global causes of death (37; 39; wor). Trials have shown that the mortality rate of lung cancer can be reduced with early detection by screening risk groups with Low-Dose Computed Tomography (LDCT) scans (4; 8; 37). Consequently, screening trials produce large medical image databases that require interpretation by highly trained radiologists. The detection of lung nodules is challenging due to their varying size and can easily be missed in the 3D volume of a LDCT. Recent interest in machine learning and deep learning has lead to the development of computer aided detection (CAD) systems that exceed the sensitivity of a radiologist (27). Yet, radiologist are of great importance in the diagnosis of lung cancer. Radiologic errors have considerable impact on a patients chance of recovery (9). A vast and extensive radiology knowledge is fundamental for the diagnosis, monitoring and treatment of lung cancer.

The development of such a radiologic knowledge base and a proficient skill set is a time consuming and costly process. A residency in the practical environment of a hospital allows a radiology resident to train on a variety of available radiological cases. In a framework of knowledge and skills of a radiologist the importance of discriminating between healthy and unhealthy images is emphasized (13). After the residency, a life-long learning is an essential aspect for radiologist to preserve knowledge and skills, especially with the rapid technological advancement in medical imaging. By reasoning, the knowledge and level of each radiologist depends on the quantity and variety of radiologic cases the radiologist has examined. A hospital accumulates a unique distribution of radiologic cases with different level of complexity, affecting the education of an individual radiologist. The cases a radiologist examines during training is limited by the patient safety and privacy concerns, as well as low prevalence of certain disease types. The efficiency and quality of the training process of a radiologist can be optimized when the radiology residents are presented with cases suitable to their level, or fit to their knowledge gap.

The field of radiology is based on image interpretation, which is suitable to for computer-assisted learning (CAL) as an addition to traditional radiology education. In this report, a CAL system is defined as a computer based education method that facilitates interaction during a students learning process. With the use of CAL, education can be tailored and personalised to the trainees needs. Image synthesis has been successfully implemented for medical imaging after the introduction of Generative Adversarial Networks (GAN) (35). Image synthesis can be a powerful tool in radiology education by its ability to diversify training cases for medical students and radiology residents. By controlling the image synthesis, images can be produced with a specific difficulty or complexity level that fits the students level. To achieve optimal personalization of education, the knowledge gap should be defined by a concrete measure. This research focuses on a measure of difficulty for detecting lung nodules.

Difficulty is defined as: "the quality or state of being hard to do, deal with, or understand: the quality or state of being difficult" (mer). A measure of difficulty is a subjective property depending on the level, skill and experience of an individual. However, when a group of individuals find a radiological case difficult, one can assume that the difficulty is embedded in the radiological scan. A method that can determine a level of difficulty as a property of a lung CT scan is explored in the scope of this research.

1.2 Context

Computer-assisted learning systems are a promising innovation for the field of radiology. To design a radiology CAL system, a good understanding of the knowledge and skills of radiologists is vital. Image synthesis with the use of GAN has the potential to produce data that can be used for educational purposes. To achieve optimal personalization of education, which fits an individual's knowledge gap, educational material can be synthesized specifically for any individual. Yet, in order to make a knowledge gap concrete, educational material should be ordered to a scale of complexity or difficulty. Item analysis is a statistical method that can give an indication on the difficulty based on responses of a group of individuals. The group of individuals can replaced by deep neural networks to automate the calculation of the measure of difficulty. To validate the method, the calculated measure of difficulty is compared to a subjective measure of the lung nodule case.

1.2.1 Knowledge and skills in Radiology

Radiology is a relatively new field in medicine and over the past 50 years it has become a key component in diagnosis and treatment of many maladies. The numerous techniques that are currently used to visualize the inside of the human body in a non-invasive way has lead to a fast evolving specialty in medicine. On an annual basis, approximately 1 billion radiologic imaging examinations are performed world wide (9). This asks for highly trained radiologists that have to work accurately and efficiently to give a patient the best care possible. The extensive radiology residency is 5 years where the following themes are addressed:

- neuro- and head/neck radiology
- cardio and thorax radiology
- abdominal radiology
- mamma radiology
- pediatric radiology
- musculoskeletal radiology
- intervention radiology
- nuclear medicine
- molecular radiology.

The work of (13) defines a comprehensive framework of knowledge and skills required for two-dimensional and multi planar interpretation in radiology. The combination of a literature study, a semi-structured interview with a multidisciplinary expert panel and empirical data of think-aloud experiments has lead to four knowledge items and thirteen skill items, summarized in figure 1.1. The framework can serve as a guideline for training and assessment as they can be seen as learning objectives.

The framework shows that radiological image interpretation is a complex cognitive process and it should be stated that many knowledge and skill items are partly integrated. For diagnosis and treatment of diseases the reliance on the knowledge and skills of radiologist is high, but the process of radiologic interpretation is variable and therefore prone to errors (9).

There exists a discrimination between two broad categories of radiologic errors: perceptual errors and cognitive/interpretative errors (9; 42; 24). Perceptual errors have an elevated occurrence, making 60% to 80% of the radiologic error. The perceptual error is defined as an abnormality that is retrospectively present in the image, but was missed by the interpretation radiologist in the initial interpretation phase. The cognitive or interpretative error can be defined

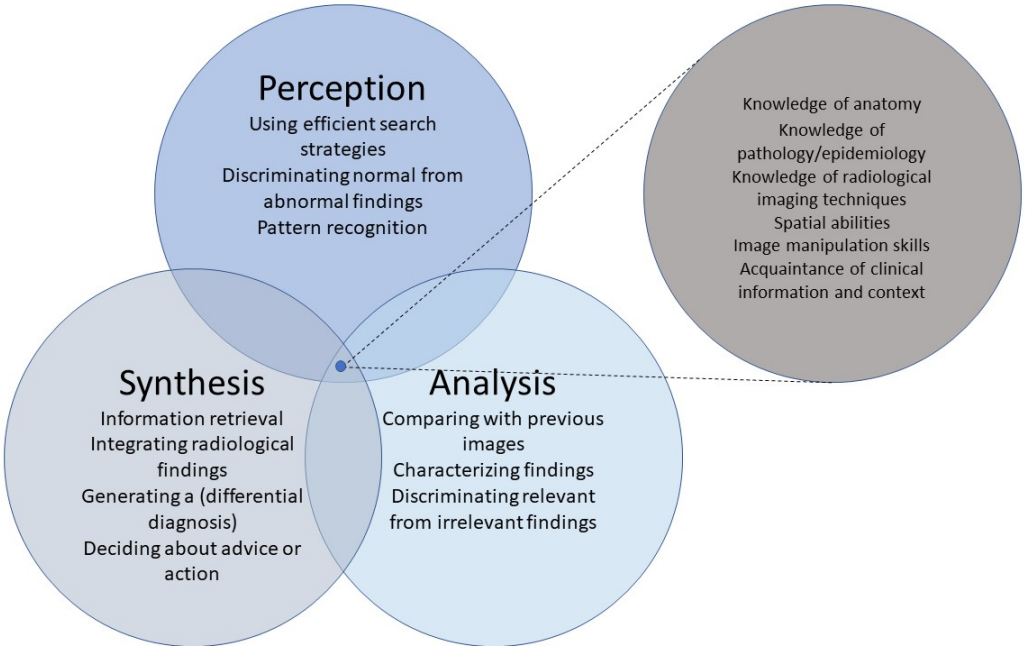

Figure 1.1: Summary framework representing the most important skills for radiological image interpretation.

as an detected abnormality a diagnostic image, but the significance is incorrectly interpreted resulting in an incorrect final diagnosis (9). Despite the technological advances in medical imaging that lead to improved image quality, the perception of the human eye and brain is not evolving with the same agility (24). Furthermore, in a medical image the diagnostic information is embedded in a background of high uncertainty. Every feature of the image may or may not represent a finding that is of clinical value (9).

In the development of strategies for error reduction, various efforts have been made, which predominantly focus on intensive education for radiologists-in-training and retraining of practising radiologists. With the focus on performance improvement of the radiologist, there is also eye for work-life-balance to limit fatigue and to mitigate pressure of the rapid pace of work. By automatizing the process of medical image interpretation with the use of computer aided detection (CAD), the radiological error is hoped to be mitigated.

1.2.2 Computer Assisted Learning

The advancement in technology for healthcare has expanded the knowledge and skills a healthcare professional should posses. It has created many niches within the healthcare branch, which lead to specializations that require extensive education. The continuous improvement of healthcare leaves specialists with the challenge of keeping up with the increasing scope of medical knowledge. This demands a systematic approach in the acquisition, assimilation, organisation and processing of knowledge. Computer assisted learning (CAL) can have the potential to aid professionals to meet the requirements.

CAL has been defined as a range of computer-based packages that focus on providing interactive instruction in a specific subject area (22). The work of Schittek (32) defines computer based instruction (CBI), computer aided (CAI) and computer aided learning (CAL) as the learning procedures and environments facilitated through computers where interaction during the learning process is the key element (32).

In this report a computer assisted learning system is defined as a computer based education method that facilitates interaction during a students learning process.

Over the scope of many years, various amounts of CAL systems have been developed, specifically for healthcare. Within the peeks of the Covid-19 pandemic CAL has shown to be indispensable in today's educational system, since distanced learning was the only option available for many students to continue their education. Other advantages of CAL is that it allows for asynchronous learning, where individuals can study at their own pace and direction (34). Every student adopts its own learning style and applies strategies for studying. A CAL system is able to anticipate on the students learning by personalising and tailoring the learning experience and by giving students feedback where needed, which promotes active and self-directed learning (22). The digitisation of education opens doors to innovate teaching methods because of the flexibility that software can offer. Research in cognitive psychology has benefited massively as advancements in software development allows only the developer's creativity to limit the innovation of didactic methods. A CAL system is inexhaustible and non-judgemental in presenting the relevant teaching material, resulting in a patient system that has time for every student. (32).

CAL is limiting in direct interaction with other students and teachers. Live teaching allows teachers to engage with the audience that can stimulate learning. Having a teacher or mentor at hand that provides adequate instruction, supervision and help is of value that cannot be underestimated (34). The use of CAL also comes with a set of technical challenges, as the development of CAL software is labor intensive, requiring appropriate hardware, backup and frequent upgrading (22; 34). The implementation of CAL systems is not straightforward and will take motivated leadership for effective integration in educational systems (34).

1.3 Item analysis

Item analysis is a term for statistical analysis on items in a test. An item can be an exercise, question, or problem on a test, questionnaire or performance assessment (16). The purpose of item analysis is to provide information about the test items, not about the competence of the test taker. Item analysis is performed retrospectively and uses data consisting the responses by individuals on a group of items (16; 28). Item analysis is able to give information about the difficulty and discriminative power of an item. In this project, Item analysis is used to calculate the level of difficulty of detecting a lung nodule.

1.4 Research objectives

In this project, a method to determine the difficulty as a property of a lung CT scan is explored. The main research goal is defined as: "How can the level of difficulty of a lung nodule case be defined as a property of the scan?" The research is supported by two subquestions.

How can item analysis with the use of deep learning determine a measure of difficulty for lung nodule cases? How does a measure of difficulty, as a property of the scan, relate to the subjective difficulty of a lung nodule case?

1.5 Thesis outline

This thesis constructs in six chapters the conducted research. The first chapter introduces the research and the research objectives are stated. Chapter 2 gives theoretical background on the research and the methods that will be used to answer the research objectives. The use of Item analysis and deep learning will be thoroughly explained. Chapter 3 explains the outline of the methods used in the research. Firstly, a description of the data used in this thesis is provided, followed by the preprocessing steps that were undertaken. Next, the deep learning process is elucidated by a description of the network architecture and parameters. Additionally, the training procedure is described. Item analysis, the application of this statistical method is delineated. Lastly, the evaluation method for the level of difficulty is explained. The results are presented in chapter 4, where an overview of the findings is given. The performance of the lung

nodule detection networks of the Difficulty-Net is presented. The variation in performance of the difficulty nets is exploited in the item analysis. The evaluation method will be presented. In chapter 5 the results will be discussed and clarified. Also the limitations of this research are discussed. The interpretation of the results will lead towards the conclusion in chapter 6. Recommendations for improvement of the research are summarized and recommendations for future work are described. The appendix contains additional information supporting decisions made during the scope of this thesis.

2 Theoretical Background

2.1 Lung nodules in CT scans

For screening and diagnosis of lung cancer low dose CT scans is a widely used imaging modality. It has the ability visualize lung nodules, because of their specific attenuation properties compared to surrounding tissue. According to Fleishner Society the definition of a lung nodule is an approximately rounded opacity more or less well-defined measuring up to 3 cm in diameter (18). When the measure of 3 cm is exceeded the Fleishner Society defines them as a lung mass; any pulmonary, pleural or mediastinal lesion seen on chest radiographs as an opacity greater than 3 cm in diameter (without regard to contour, border, or density characteristics) (18).

Dependence on the attenuation of lung nodules in CT scans, they can be categorized in three different types:

1. solid nodules
2. ground-glass nodules
3. part-solid nodules

Solid nodules are characterized by their homogeneous soft-tissue attenuation and are prevalent. Ground-glass nodules show a nonuniform appearance with hazy increase in local attenuation of lung parenchyma not obscuring the underlying bronchial and vascular structures. The part-solid nodules are a constitution of solid and ground-glass attenuation components (26). Lung nodules occur solitary, when they are fully surrounded by normal lung tissue. Mostly ground-glass nodules are found in within multiple nodular lesions. Studies have showed that within multiple nodules the malignant dominant tumor also present benign satellite lesions, but it should be noted that not always is the larger or multiple nodules the most dominant malignant nodule (26).

The size of the nodule is strongly associated with the probability of malignancy. Additionally, spiculation of the anatomical morphology and the prevalent occurrence in the upper lung lobes are identified as predictors of malignancy. Lung nodules with a diameter below 6 mm are considered to be of low malignancy risk (26). Other predictors of benign etiology are nodules (<10 mm) adjacent fissures or the pleural surface, as they are likely to depict intra-pulmonary lymph nodes (26). Calcification is present in 10% of lung cancer cases, but calcified lung nodules are generally not considered malignant.

2.2 Computer Aided Detection for Lung Nodules

With the use of machine learning algorithms lung nodules can automatically be detected. The performance of these algorithms has improved over the last years and nowadays the performance exceeds the accuracy of a radiologist. Initially, feature-based algorithms like support vector machines (SVM), random forest and regression trees classifiers are used. The workflow of these algorithms operate in the following fashion: first, a lung segmentation is performed to remove structures like the vessels, the bronchi and the rib cage. Lung nodule candidates are detected and their location is identified in the second stage. In the third stage, the features like shape size and texture is extracted from lung nodule candidates. Lastly, the features are combined using clustering techniques where the candidates are classified as true positives. The SVM classifiers obtain the high results with respect to accuracy, sensitivity and sensitivity (27; 43). The rise of Deep Learning algorithms have shown to obtain better results than feature based methods for automatic lung nodule detection, because of the advantages deep

learning has in terms of the processing of low dose CT image data(29; 15). Still, many researchers use conventional machine learning methods combined with deep learning methods, such as convolutional neural networks (CNN), recurrent neural networks, deep belief networks, auto-encoders and general adversarial networks (15). Automatic lung nodule detection can be achieved with object detection algorithms, comprising two categories: two-stage and one-stage algorithms. The two-stage category first uses an algorithm to select lung nodule candidates, then another algorithm analyses the candidates and reduces the false positives. The two-stage algorithms have good performance and high sensitivity for lung nodules (40; 38; 10; 44). The two-stage algorithm can be extended with 3D information of the lung nodule to obtain higher performance (19; 45). To use 3D information of in the low dose CT is challenging because of the limited computational power for deep neural networks. In the work of Jenuwine (21) 3D lung nodule detection without candidate selection is attempted. The system is not accurate enough to be usable, but the paper gives proof for one-stage lung nodule detection. To investigate the possibility of one-stage lung nodule detection, the work of (12) uses a YOLO based deep learning network for lung nodule detection. The YOLO network is an object detection network that predicts a bound box coordinates. The system uses depth information by implementing the preceding and succeeding images of the scan image of interest. The performance of the network on the LIDC/IDRI dataset is 89% sensitivity at 6 false positives per image. The work of (25) uses the latest YOLO v3 algorithm as a CNN implementation (31) for the detection of lung nodules in simulated data and patient data. The simulation study showed a sensitivity of 99.3% with 4 false positives per scan. The patient study shows a sensitivity of 90.0%. Besides a high performance the YOLO v3 network has high computational efficiency. Khosravan (23) tackles the computational power problem with a different one-stage detection algorithm and proposes S4ND: a deep learning based method for lung nodule detection. The 3D deep network architecture is designed to detect lung nodules in a singe shot using a single-scale network. The S4ND method achieved an average Free-Response operating curve score (FROC-score) of 0.897 (23).

2.3 Item analysis

Item analysis is often applied to educational sciences to monitor the quality of questions on tests (16). To illustrate the interpretation of item analysis a student exam analogy is used. In the context of this research, a question on an exam corresponds to a lung nodule case. The performance of all students participating in the test correspond to the performances of multiple deep neural networks. The items are regarded as a stack of slices from a lung CT on which the following question could be posed: "Does the stack contain a lung nodule?" The question is answered with yes or no.

Item analysis is able to give information about the difficulty and discriminative power of an item. These features can be calculated with several methods. One must beware of the pitfalls of such methods to be able to draw valid conclusions.

2.3.1 Item difficulty

The difficulty of an item is simply how hard the item is. Some items can be too hard to answer which results in almost no correct responses, whereas some items are so easy that every test taker can answer them correctly. By having insight in the difficulty of every item, the test maker has more control over the overall level of difficulty of the test (16).

The most obvious measure of difficulty of an item for a group of test takes is the average score on the item. This is only possible when the item is dichotomous, meaning that the answer can only be correct or incorrect. The average score on the test is only useful when prior knowledge about the group taking the test is available. The difficulty can also be calculated by dividing the number of test takers answering the item correctly N_c by the total number of test takers

answering the item N_t. In equation 2.1 the proportion is showed.

$$p = \frac{N_c}{N_t}. \tag{2.1}$$

For Item analysis, it should be considered that the p-value from 2.1 is a behavioural measure, because difficulty is defined in terms of relative frequency with those taking the test to choose the correct response. The p-value here is also a characteristic of both the item and the sample taking the test (16; 28). The value ranges from 0 to 1, where 0 represents the most difficult item and 1 represents the least difficult item. In this context, p-value is a behavioural measure of neural networks.

2.3.2 Item discrimination

The discriminative power of an item is the tendency of the item to be answered correctly by test takers that are strong in the skills and knowledge, necessary for the item to be answered correctly, and the test takers that are not. In item analysis the discriminative power of an item can be evaluated with a measure of the proficiency and competency, which is called the criterion. In this research the criterion is the performance of a neural network on the entire test set.

The discriminating power can be measured by the product moment correlation in which one variable is continuous and the other variable is binary (dichotomous). In the context of this research the point-biserial correlation would be a measure of discrimination by calculating the correlation between the scores on the item Y and the performance on the test set or criterion X (16; 36). The simplified formula for the point-biserial correlation r_{pb} follows:

$$r_{pb} = \frac{\overline{X_1} - \overline{X_0}}{\sigma_x} \sqrt{\frac{n_1 n_0}{n(n-1)}}, \tag{2.2}$$

where $\overline{X_1}$ is the mean on X of those who scored 1 on Y,
$\overline{X_0}$ is the mean on X of those who scored 0 on Y,
σ_x is the standard deviation of all n scores on X,
n_1 is the number of test takers scoring 1 on Y,
n_0 is the number of persons scoring 0 on Y,
and $n = n_1 + n_2$ (14).

2.4 Validation

To validate a level of difficulty for the detection of a lung nodule, a subjective variable is necessary. This project will use the Lung Imaged Database Consortium - Image Database Resource Initiative (LIDC/IDRI) data set which is provided with metadata containing annotations of nodules of 4 radiologist. The construction of the data set will be elaborated in section 3.1. Each radiologist was asked to assess several characteristics of their annotated nodule. The following characteristics were scored on a scale from 1-5.

- Subtlety - in terms of its difficulty in detection
- Internal structure - or expected internal composistion of the nodule
- Calcificiation - pattern of calcification present
- Sphericity - the three dimensional shape of the nodule in terms of its roundness

- Margin - – description of how well defined the margins of the nodule is
- Spiculation – amount of speculation present in nodule
- Texture – internal texture or composition of nodule in terms of solid and ground glass components
- Malignancy - Radiologist subjective assessment of likelihood of malignancy of this nodule

A subjective score on the difficulty is provided in the subtlety score, where higher values indicate easier detection.

1. 'Extremely Subtle'
2. 'Moderately Subtle'
3. 'Fairly Subtle'
4. 'Moderately Obvious'
5. 'Obvious'

The objective difficulty score developed in this thesis can be validated by using the subjective difficulty score provided by the LIDC/IDRI dataset. The subtlety score is seen as a variable with an ordered categorical scale, as there clear ordering in the levels. Yet, the absolute distances between the categories is unknown. To analyze subtlety score and its relation to the determined level of difficulty score, ordinal regression analysis will be applied. This type of regression restricts analysis solely to the methods that use only the ordering information about the categories (6).

2.4.1 Ordinal regression

Ordinal logistic regression is used to predict an ordinal dependent variable given one or more independent variables. An ordinal dependent variable can be predicted by one ore more independent variables using ordinal logistic regression. An ordinal dependent variable is a variable with an ordered categorical scale and the independent variables are continuous, ordinal or categorical. For example, ordinal regression can be used to predict the likelihood a bachelor student will do a masters, based on a 4-point Likert scale describing 4 levels; "very likely","somewhat likely", "somewhat unlikely" and "very unlikely". The regression model is able to predict the category based on a number of independent variables. In this example it could be; "Age" and "Savings". Additionally, ordinal regression could be used to analyse whether an independent variable can predict the ordinal dependent variable (6).

In this research it is tested whether the independent variable "level of difficulty" can predict the ordinal dependent variable "subtlety of a lung nodule". By realizing ordinal regression model, it can be tested if the model is statistically significant. The independent variable "level of difficulty" the test will show how a single unit increase or decrease in that variable is associated with he odds of the dependent variable "subtlety" having a higher or lower value.

For ordinal regression the data has to satisfy the following assumptions.

- **Assumption 1:** The dependent variable has to be measured at the ordinal level. This means that the dependent variable consists of ranked categories.
- **Assumption 2:** The independent variable(s) are either continuous ordinal or categorical.

- **Assumption 3:** For two or more independent variables there is no multicollinearity, which could occur when two or more independent variables are correlated with one another. It could lead to problems with understanding which variable contributes to the explanation of the dependent variable.
- **Assumption 4:** The independent variable has an identical effect at each cumulative split of the ordinal dependent variable. This can also be described as proportional odds.

Assumptions 1 and 2 should be tested first, as they are fundamental for ordinal regression. Assumption 3 is important for two or more dependent variables, which is not the case in this research. Assumption 4 can be checked with the test of parallel lines.

The notation of a ordinal regression model is as follows. Suppose Y is an ordinal dependent variable with J categories. The cumulative probability of Y less than or equal to a specific category $j = 1, .., J-1$ is noted as $P(Y \leq j)$. From the cumulative probability follows that $P(Y \leq J) = 1$. The odds of being less than or equal to a particular category is defined as

$$\frac{P(Y \leq j)}{P(Y > j)} \tag{2.3}$$

for $j = 1, ..., J-1$ as $P(Y > J) = 0$. Here dividing by zero is undefined. The log odds, also known as logit, can be described in the following equation, where $P(Y > j) = 1 - P(Y \leq j)$.

$$\text{logit}(P(Y \leq j)) = \log \frac{P(Y \leq j)}{P(Y > j)} \tag{2.4}$$

The model of ordinal logistic regression can be defined as

$$\text{logit}(P(Y \leq j)) = \beta_{j0} + \beta_{j1} x_1 + ... + \beta_{jp} x_p, \tag{2.5}$$

where $\beta_{j0}, \beta_{j1}, ..., \beta_{jp}$ are model coefficient parameters with p predictors for $j = 1, ..., J-1$. In many software packages the model coefficient parameters are subtracted. This makes it easier to interpret the relationship between the independent variables and the relationship with the ordered variables. The computed probabilities can be revalidated with the actual outcome to validate if high probabilities are associated with events and low probabilities non-events.

A logistic model is said to provide a better fit to the data if it shows an improvement over the null model, which can be seen as a baseline model without predictors. The model is tested against the null hypothesis that all observations belong to the largest outcome category. An improvement on the baseline model can be tested with the -2 Log Likelihood test, which are often calculated by statistical programs or packages (30).

3 Method

3.1 Data

3.1.1 Data characteristics

For the development of comuter aided detection (CAD) methods for lung nodules the Lung Imaged Database Consortium (LIDC) and the Image Database Resource Initiative (IDRI) have completed a publicly available database with a well-characterized repository of thoracic CT scans. The LIDC and IDRI consist of the Weill Cornell Medical College, University of California, Los Angeles, University of Chicago, University of Iowa, University of Michigan, MD Anderson Cancer Center and Memorial Sloan-Kettering Cancer Center and were broadened out with 8 medical imaging companies (AGFA Healthcare, Carestream Health, Inc., Fuji Photo Film co., GE Healthcare, iCAD, Inc., Phillips Healthcare, Riverain Medical and Siemens Medical Solutions). As the expertise of academic centers was merged with the medical imaging companies, the database collection will be referred to as the LIDC/IDRI Database.

The LIDC/IDRI Database contains 1018 helical thoracic CT scans collected from screening trials. The data is anonymized to remove all protected health information. To achieve a heterogeneous range of scanner models and technical parameters, no scan was performed specifically for the purpose of the database. Furthermore, the Database includes one scan from one patient, such that the scans in the LIDC/IDRI Database are not correlated.

To ensure relevance of the scans for the development of CAD systems for lung nodules, the database includes (1) an image repository of screening and diagnostic thoracic CT scans, (2) metadata associated with the scan such as technical scan parameters and patient information, and (3) nodule truth information based on the subjective assessments of a panel of experienced radiologists.
The nodule truth information was obtained in a two-step image annotation process, where each case was interpreted by 4 radiologists. In the first "blinded read phase" every radiologist independently reviewed a scan and marked lesions. The lesions where placed in one of the following categories:

1. nodule > 3 mm
2. nodule < 3 mm
3. non-nodule > 3 mm

A lesion defined as nodule > 3 mm is considered to be a nodule with greatest in-plane dimension in the range 3 to 30 mm, regardless of presumed histology. The lesions defined as nodule < 3 mm are considered to be nodules with greatest in-plane dimension less than 3 mm, that is not clearly benign. The non-nodules are any other pulmonary lesions greater than or equal to 3 mm that does not possess features consistent with those of a nodule. (7) The radiologist did not receive a definition of the concept "nodule", such that each radiologist provided their own interpretation of the "noduleness" of a lesion during the blinded read phase. For each nodule >3mm the centre-of-mass location has been indicated. In the second "unblinded read phase" the results of all radiologists were anonymously revealed to each of the radiologists, who then independently reviewed the other radiologists marks, as well as their own. This makes it optional for the radiologist to leave its own marks unchanged, switched in terms of lesion category, deleted or additional marks can be added. The database contains 19 CT scans from 19 patients where no nodule > 3mm or nodule < 3mm was marked by the radiologist. For 268 of the 10180 patients, pathological information was collected retrospectively. This includes the

patients diagnosis (nonmalignant disease, primary lung cancer or metastatic disease) along with the method of diagnosis.
The two-phase process was developed for the asynchronous interpretation of thoracic CT scans without forced consensus, which incorporates real-world variability of image interpretations. The lack of ground-truth creates the challenge for researchers in how to define targets for the training and testing of CAD methods. The annotations made for the database differ with a radiologists accustomed routine assessment in routine practice. Here the radiologists encounters 3 steps;

1. Detection of a lesion.

   Is the observed structure an abnormality or normal anatomy?

2. Determination of lesion size.

   Is the lesion greater than or less than 3 mm?

   Is the lesion in the range of 3 mm to 30 mm?

3. Evaluation of the lesion features.

   Does the lesion represent a "nodule"?

   If the lesion is less than 3 mm, is it clearly benign?

This creates many possible combinations for one single lesions, and any option assigned by the different radiologist should be considered reasonable due to the inherent subjectivity. The LIDC/IDRI database has several obvious applications, but is only limited by the creativity of those who use it. (7)

3.1.2 Data preprocessing

The LIDC/IDRI database consists of an image repository of screening and diagnostic CT scans that requires preprocessing before deep learning techniques can be applied. With the use of pylidc, the data queried in an sQL-like fashion (17). The steps necessary for training the lung nodule detection algorithm are described. The preprocessing methods are adapted to the network architecture and the purpose of this research.

The LIDC/IDRI consists of 1018 helical CT scans from 1010 patients, where only one scan is considered per patient. Scans with a slice thickness larger than 3 mm were excluded. The remaining scans are resampled, with the use of interpolation, such that each voxel is 1x1x1 mm. The transversal slices are cropped and padded to a size 512x512. The scans were set to the lung window ranging from -1250 HU to 250 HU.

In the total data set, 19 cases are healthy and have no lung annotations. During the training of computer assisted detection algorithms problems arise because of the ratio of positive (lung nodule) data and negative (no lung nodule) within one scan. The enrichment of data with lung nodules in data sets is necessary for the training of a CAD algorithm to achieve sufficient sensitivity. Instead of training the CAD algorithm on the full scan, a stack of slices containing a nodule is used. Essentially, a level of difficulty is determined for a stack of slices and not the complete lung CT scan. It was chosen to have a fixed amount of slices per stack such that the performance of the neural networks could be expressed based on a number or slices. Figure 7.1 in appendix 7 shows the thickness of nodules in slices in a histogram. By dividing the distribution in quartiles it was found that approximately 75% of the nodules had a thickness of 11 slices. The average thickness was 6 slices. For easy handling the size of the stack was set to 10 slices per stack. Additionally, to show that all lung nodules in the data set are not accumulated at a certain location within the scan, the locations of all nodules along the cephalad direction are shown in appendix 7 figure 7.2. Furthermore, the stacks are restricted to having 1 nodule

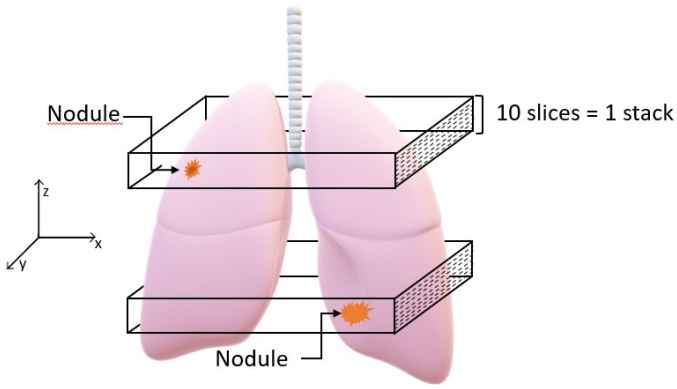

Figure 3.1: Schematic representation of the construction of two stacks from one pair of lungs. Each stack consists of 10 slices where the centeroid of the nodule is position on the fifth or sixth slice in the transversal direction (z-direction). For stacks where nodules are smaller than 10 slices in thickness, healthy slices were added up on the bottom and top, to aqcuire stacks of 10 slices.

per stack, such that a measure of difficulty would be dependent on one nodule in the stack. The presence of multiple nodules in one stack has an effect on the level of difficulty, when it would be presented to human subjects.

In short, a stack was created from the following conditions:

1. The stack has a thickness of 10 slices
2. The stack contains 1 nodule
3. The nodule is annotated by at least one radiologist

From these restrains follows that nodules thicker than 10 slices are initially excluded. Also nodules that show an overlap in the z-direction are not considered to form a stack. In a later state of the research, stacks with nodules larger than 10 slices were included to achieve variation in the networks. Figure 3.1 gives a schematic illustration of a pair of lungs with two nodules. The nodules are at different locations and have a thickness smaller than 10 slices, so from this scan 2 stacks are created. The slice containing the the center of mass of the nodule is taken and the 4 preceding stacks and 5 succeeding stacks are taken to form a stack of 10 slices. It must be noted that within a stack, the top and/or bottom slices might not contain features of the nodule.

The bounding box of each annotation is stored in the data set and from each nodule one annotation was selected. Coordinates of the bounding box were given in the following format and resampled and padded according to the sampling and padding of the scan.

$$\text{bounding box} = [x\_min \;\; x\_max \;\; y\_min \;\; y\_max \;\; z\_min \;\; z\_max]$$

Figure 3.2: Architecture of the deep neural network.

3.2 Deep Neural Network

As described in section 2.2, there are several methods for the detection of lung nodules with deep learning algorithms. The two-step and one-step methods each have advantages. For the purpose of a determining the diagnostic difficulty of a stack, a one-step lung nodule detection algorithm has been selected without a sliding window approach. In order to perform Item analysis on the stacks, multiple neural networks of varying performances are required. The network architecture and parameters are presented in 3.2.1 and in 3.2.2 the training procedure is outlined. The variation in performance is also explained.

3.2.1 Proposed architecture

The chosen network architecture is a network for lung nodule detection inspired by the YOLO algorithm, but is drastically simplified and has an initial accuracy of 65%. In figure 3.2 the architecture of the network is presented. The network uses an input of 512 x 512 x 1 pixels and a label of 16 x 16 x 1 pixels. The architecture employs 18 2D convolutional layers with filtersizes of 3 by 3 and takes strides 1,1 strides. The convolutional layers implement padding on all borders and apply a Rectified Linear Unit (ReLU) activation function given in 3.1. Each convolutional layer is followed by a batch normalization. In the network 5 2D MaxPooling layers are used, which also apply padding. The network provides a 16 x 16 label map as an output. The network is build with Keras and Tensorflow (5; 11). The loss function, given in equation 3.2 is based on the mean squared error (MSE) with an Adam optimizer using a learning rate of $1e^{-5}$.

$$\mathrm{ReLU}(z) = \max(0, z) \tag{3.1}$$

$$\mathrm{Loss} = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{3.2}$$

3.2.2 Training procedure

During preprocessing 710 nodules have been found that meet the requirements stated in section 3.1.2. For the training process the data is divided into a train set and a test set, that contain 80% and 20% respectively. The train set and test set are disjoint, meaning that multiple stacks from the same patient are in the same set. This results in a train set of 659 stacks with 6590

slices and a test set of 151 stacks with 1510 slices. The summary can be found in appendix 7 table 7.1.

The ensemble of networks will be referred to as the Difficulty-Net ensemble. The Difficulty-Net consists of 10 neural networks with the same architecture as described in section 3.2.1, but with varying performances. A parameter that is known to influence the performance of a neural network is the size of the data set. To acquire 10 neural networks that are comparable in architecture and parameter, but with varying performances, the networks are trained on different sizes of train sets. All ten networks are tested with the same test set.

The initial network (NetA 100) was trained with the train set for 100 epochs with a batch size of 64. It takes 160 minutes to train the proposed network on a 2x Tesla A40 GPU. The network is tested on the test set and the performance is expressed in the Area Under Curve (AUC). The AUC of NetA 100 is 0.67. The restrictions stated in 3.1.2 for the stacks has excluded 1429 nodules. Each nodule has been created into a stack of 10 slices and was incrementally added to the train set. In total 10 networks were trained with increasing size of train set. In table 3.1 an overview of the sizes of the train sets has been given

Difficulty-Net	Train set size (slices)
NetA 100	6590
NetB 90/10	7320
NetC 85/15	7750
NetD 80/20	8230
NetE 75/25	8780
NetF 70/30	9410
NetG 65/35	10130
NetH 60/40	10980
NetI 50/50	13180
NetJ 45/55	14640

Table 3.1: Train set sizes of the neural networks that are part of the Difficulty-Net ensemble.

3.3 Item analysis

The purpose of the Item analysis is to map the stacks in the test set to a ranking based on their level of difficulty as a property of the scan. The item analysis will be performed on each of the stacks in the test set, where every stack in the test set represents an item. Item analysis can only be performed retrospectively, meaning the performance of the 10 neural networks on the whole test set is used to determine the difficulty of one stack. The value of the Point biserial correlation on the stack is used as a measure for the difficulty of the stack. In this context, the value of the point-biserial correlation gives in indication of how well the stack can be used to discriminate high performing neural networks from low performing neural networks in the Difficulty-Net ensemble. Figure 3.3 shows which values are used to calculate the Point Biserial correlation. The stack from the test set is fed to all 10 neural networks from the Difficulty-Net ensemble to give a prediction on detecting the nodule. The prediction is transformed to a binary score, where 1 indicates that the nodule has been detected and 0 indicates the nodule has not been detected. The transformation of the prediction to a binary item can be performed in several ways and will be discusses in section 3.3.1. The formula of the point-biserial correlation takes the square root of the number of neural networks that were able to detect the nodule multiplied by neural networks that were not able to detect the nodule divided by the total number of neural networks in the ensemble. This is multiplied by the mean value of the performance of the neural networks detecting the nodule minus the mean of the performance of the neural networks not detecting the nodule, divided by the standard deviation of the performance of all

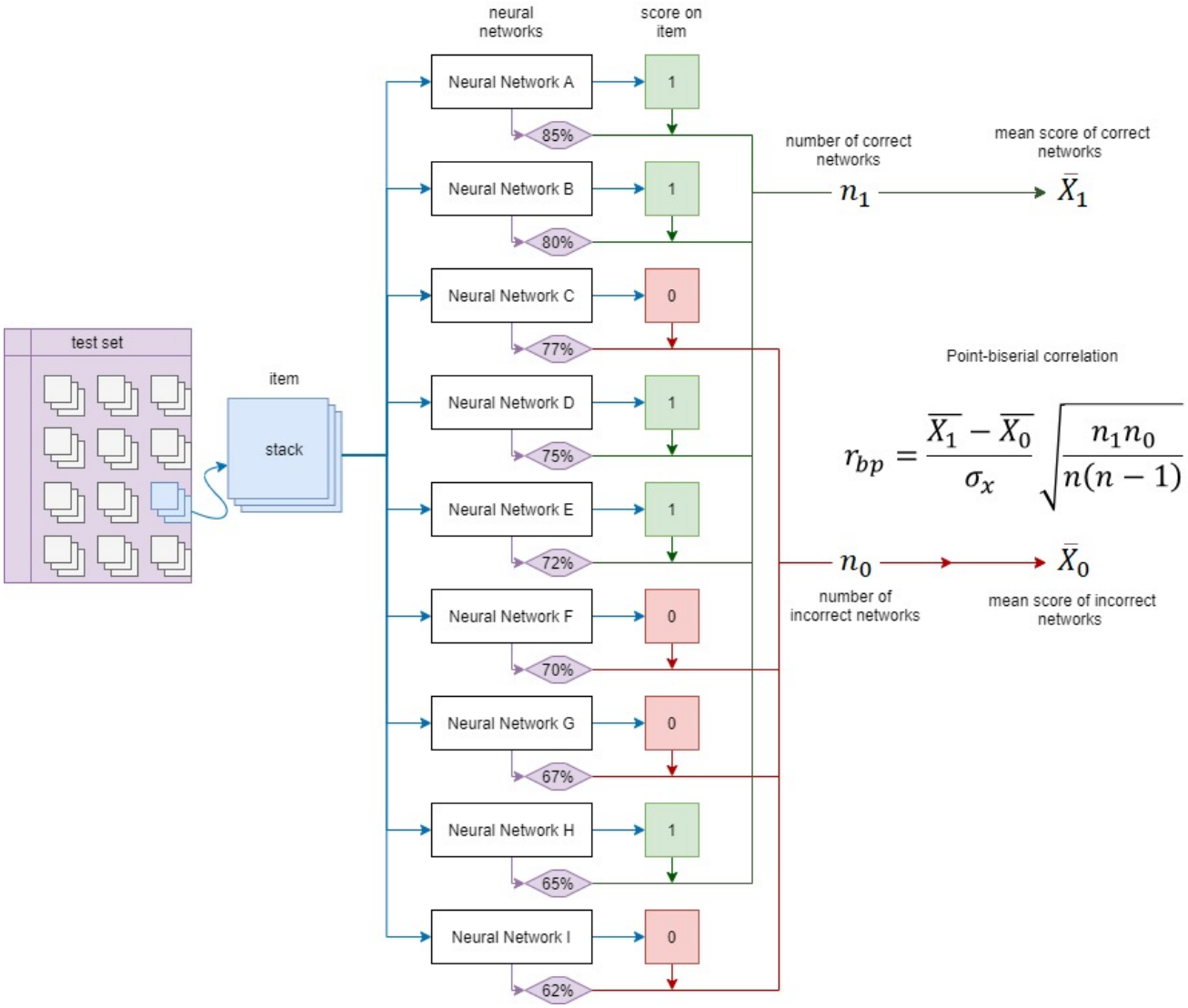

Figure 3.3: Flow chart of the composition of the Point-biserial correlation. A stack from the test set is presented to all neural networks. Their prediction is transformed to a score on the item. The number of correct networks and the average performance of the networks answering correct is calculated. The same is done for all networks answering incorrect. Together with the standard deviation of the performance of all networks, the point-biserial correlation can be calculated.

neural networks. The point-biserial correlation gives a value ranging form -1 to +1, where +1 indicates strong discrimination and 0 means weak discrimination. A value of -1 indicates that the detection of the nodule in the stack was random. The point-biserial correlation is calculated with the scipy.stats.pointbiserialr() function (41).

3.3.1 Point Biserial Correlation

For the purpose of this research a measure of difficulty is determined for each stack. The trained neural networks take a slice of 512 by 512 pixels as input. The output is a prediction map of 16 by 16 squares, which gives a the probability of presence of a nodule. Since the neural network takes 1 slice as an input, the calculation of the Point Biserial correlation for a stack of 10 slices can be conducted in several ways. Ideally, the point-biserial correlation can be related to the subtlety score given by the radiologist. 3 methods of how the point-biserial correlation can be calculated will be outlined.

For the point-biserial correlation a binary score needs to be provided. The stacks in the test set all have one nodule, but some nodules are less than 10 slices thick, resulting in stacks that have slices without a visible or annotated nodule. In each method the true positives, true negatives, false positives and false negatives are regarded to determine the binary score.

Method 1 In this method a score for each slice is determined from which the point-biserial correlation value is calculated for each slice. The point-biserial correlation of the stack is then determined by taking the mean value of the point-biserial correlation over all 10 slices in the stack. By looking at individual slices, one discriminates slices with an annotation of a nodule (a positive slice) and slices without an annotation of a nodule (a negative slice). As mentioned, the network gives a prediction map onto which a threshold is applied. The threshold is scaled to individual performance of each Net in the Difficulty-Net ensemble. A binary mask is created which can be compared to the label of the slice.

If the slice is positive and the nodule is detected, a score of 1 will be granted. This means that there is one true positive in the slice, and all false positives are ignored. In case of a positive slice and the nodule is not detected a score of 0 will be given, this means the scan is false negative.

For negative slices the score is more difficult to determine. Ultimately, if the slice is negative, the binary mask is also completely negative, by 256 squares correctly. Since the false positives are ignored in the positive slices, a score of 1 will be granted when 95% of the squares are assigned negative. A score of 0 will be given, when less than 95% of the squares are assigned negative.

The binary scores are saved in a matrix with the number of slices as rows and the neural networks from the Difficulty-Net ensemble as columns. The matrix is used to calculate the point-biserial correlation, illustrated in figure 3.3. The difficulty score of the stack is the mean value of the point-biserial correlation of each slice in the stack. This method uses the assumption that every slice in the stack contributes equally to the difficulty score.

Method 2 In this method the binary score of the stack is determined by the binary score of the individual slices. First a binary score is assigned to each of the slices in the stack as described in method 1. If the slice is positive and the nodule is detected, a score of 1 will be granted. In case of a positive slice and the nodule is not detected a score of 0 will be given. A score of 1 will be granted when 95% of the squares are assigned negative. A score of 0 will be given, when less than 95% of the squares are assigned negative.

If 5 or more slices in the stack have a binary score of 1, the binary score of the stack will be 1. If less than 5 slices in the stack have a binary score of 1, the binary score of the stack will be 0. The binary scores of the stacks is saved in the matrix and and a point-biserial correlation is calculated for each stack.

3.4 Evaluation

To investigate the relation ship of the point-biserial correlation with the level of difficulty of a stack, a evaluation method is developed. The LIDC/IDRI data set is provided with meta data and each radiologist gives a subtlety score of their annotated nodule, described in section 2.4.1.

The subtlety scores of the stacks in the test set are shown in figure 3.4, where it can be seen that most of the nodules in the test set are score a 3 ("Fairly Subtle") or higher. Only ten nodules score have a subtlety score of 1 ("Subtle").

3.4.1 Ordinal regression

For each of the methods described in 3.3.1 ordinal categorical logistic regression analysis is applied with the use of IBM SPSS statistics (IBM Corp.). We define here the subtlety categories as dependent variable and the point-biserial correlation as an independent variable. With the use of the Goodness-of-Fit test, the -2 log likelihood and test of parallel lines, the statistical

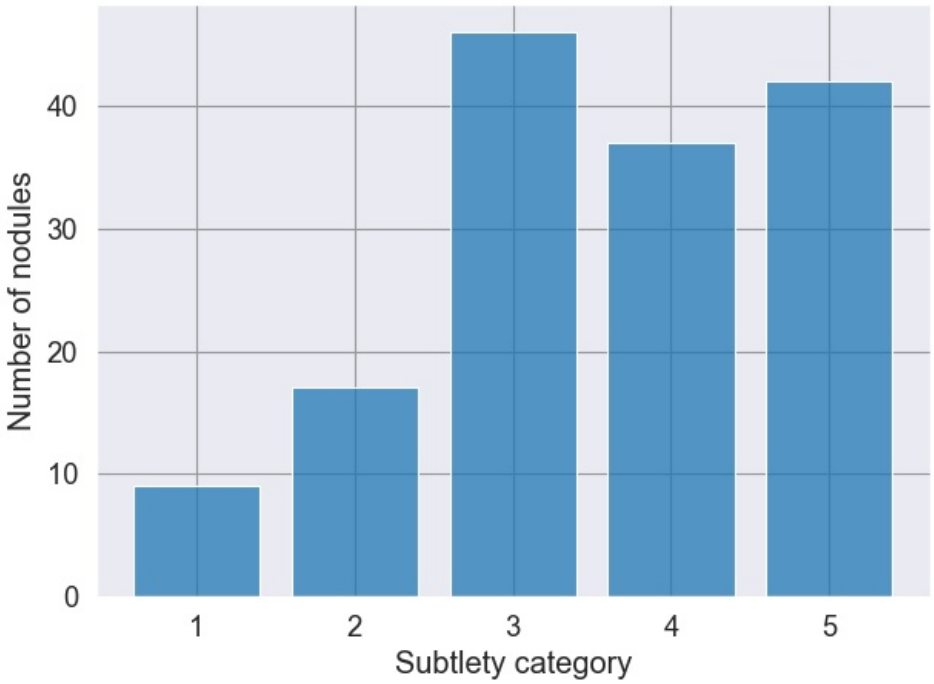

Figure 3.4: Histogram of the distribution of subtlety scores in the test set.

significance of the analysis will be tested. The logit values are transformed to the odds ratio, which gives insights into the relationship between the point-biserial correlation and the subtlety scores.

4 Results

4.1 lung nodule detection network

The performance of the neural networks in the Difficulty-Net ensemble are presented in a Receiver Operating Characteristic (ROC) curve in figure 4.1 and in table 4.1. Variation in the performance of the neural networks can be seen in the ROC curves. The True Positive rate and False Positive rate differ per network. Table 4.1. shows the Area Under Curve (AUC) of each of the network, which is regarded as a measure for the performance of the network. Increasing size of the train set was expected to have an effect on the performance of the networks on the test set. From table 4.1 and figure 4.1 the performance of the network does not seem to increase or decrease by the size of the train set. The least performing Net from the ensemble is NetB 90/10, where an AUC of 0.573 on the test set was found. A good performing network in the Difficulty-Net ensemble is NetF 70/30. Here, 70 % consists data equal to the train set used for NetA 100. Another 30 percent of excluded data is added, as described in 3.1.2. To verify if this performance is caused by the additional 30% of data or by random initialization, NetF 70/30 is trained 3 times. In figure 4.2 the ROC curve is presented, where it can be seen that the performance of the repetitions is lower than than the initial performance of NetF 70/30, used in the Difficulty-Net ensemble. The average performance with a standard deviation of 0.033 indicates that NetF 70/30 obtained an AUC of 0.704 because of the favorable random initialisation.

For this research, a variation in performance is desired to be able to perform Item analysis. The performance of the networks presented in table 4.1 shows an average performance of 0.667 with a standard deviation of 0.043. With the knowledge of the checks conducted with NetF 70/30, it is suspected that the performance of the network is limited. It was decided to attempt to prove the research objectives with the 10 neural networks in the Difficulty-Net ensemble, because addition of Networks with similar performance would not influence the variation in performance. Further elaboration on the limited performance is given in discussion section 5.

Difficulty-Net	Train set size (slices)	AUC
NetA 100	6590	0.671
NetB 90/10	7320	0.572
NetC 85/15	7750	0.665
NetD 80/20	8230	0.603
NetE 75/25	8780	0.672
NetF 70/30	9410	0.704
NetG 65/35	10130	0.702
NetH 60/40	10980	0.673
NetI 50/50	13180	0.698
NetJ 45/55	14640	0.707
Average	9701	0.667 ± 0.043

Table 4.1: Performance on the test set expressed in Area Under the Curve (AUC) of the neural networks in the Difficulty-Net ensemble with train set sizes.

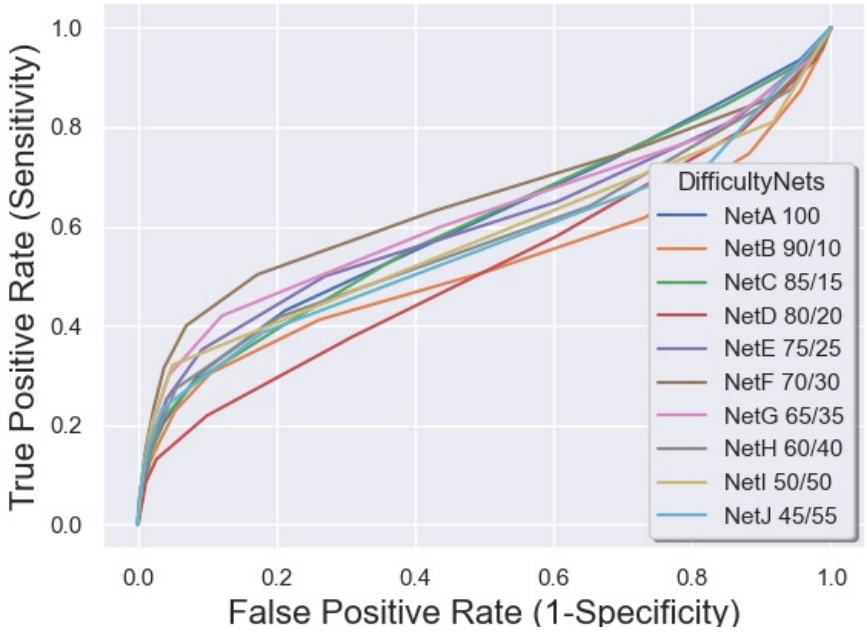

Figure 4.1: ROC curves of the neural networks in the Difficulty-Net ensemble.

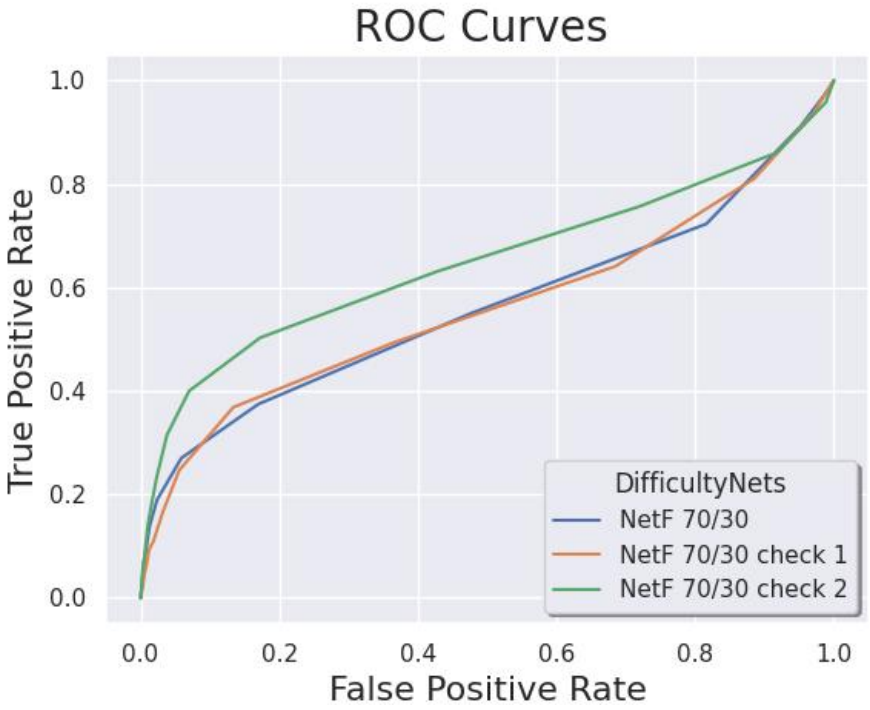

Figure 4.2: ROC curves of NetF 70/30, NetF 70/30 check 1 and NetF 70/30 check 3.

Difficulty-Net	Train set size (slices)	AUC
NetF 70/30	9410	0.704
NetF 70/30 check 1	9410	0.639
NetF 70/30 check 2	9410	0.629
Average	9410	0.658 ± 0.033

Table 4.2: Performance on the test set expressed in AUC of NetF 70/30 trained 3 times with identical train set.

4.2 Item analysis

point-biserial correlation

The neural networks from the Difficulty-Net ensemble are used to perform Item analysis on the stacks in the test set. A value for the point-biserial correlation is calculated for each stack. Figure 3.3 describes how the performance of the neural networks and the score on the stacks is used to determine a value for the point-biserial correlation. The score on the stacks is determined by the predictions of the networks in the Difficulty-Net ensemble. To determine the score with the predictions, four methods have been used. These four methods result each in four distributions of the point-biserial correlation over the test set. The two methods described in 3.3.1 are presented in this section, while two other methods and results can be found in the appendix 7.2. The values of the point-biserial correlation are presented in a histogram to show how the values are distributed. The value of the point-biserial correlation ranges from -1 to +1. The values of the point-biserial correlation of the complete test set are shown in a histogram, where the distribution of the value can be assessed. For each method described in 3.3.1 a histogram is given.

Method 1 In this method the point-biserial correlation of the stack is calculated by taking the mean of the point-biserial correlation of the slices, as described in section 3.3.1. In figure 4.3 a distribution of the point-biserial correlation of the stacks can be found. Each stack has been given a subtlety score by the annotated radiologist. To investigate the relationship between the point-biserial correlation of the stack and the subtlety score of the stack. Figure 4.4 shows raincloud plots for each subtlety score. The raincloud plots show a distribution with below a strip plot where the individual data points are indicated. Over the strip plot a boxplot is placed to show the median, and quartiles of the distribution. For subtlety category 1, a distribution that is more orientated towards positive values of the point-biserial correlation. For category 2, 3 and 4, values of the point-biserial correlation are distributed over total range. For subtlety score 5 the distribution is more focused on negative values for the point-biserial correlation. It is expected that an ordinal categorical logistic regression model can be fitted onto the data to investigate the statistical significance of the relationship between the point-biserial correlation of the stacks and its subtlety score.

Method 2 The values of the point-biserial correlation calculated by method 3, described in section 3.3.1, are presented the histogram of figure 4.5. Negative values, as well as positive values of the point-biserial correlation are found for the stacks. In figure 4.6 it can be seen that negative values for point-biserial correlation are more associated with subtlety score 5 and positive values are associated with subtlety score 1. For subtlety score 3, both positive and negative values of the point-biserial correlation.

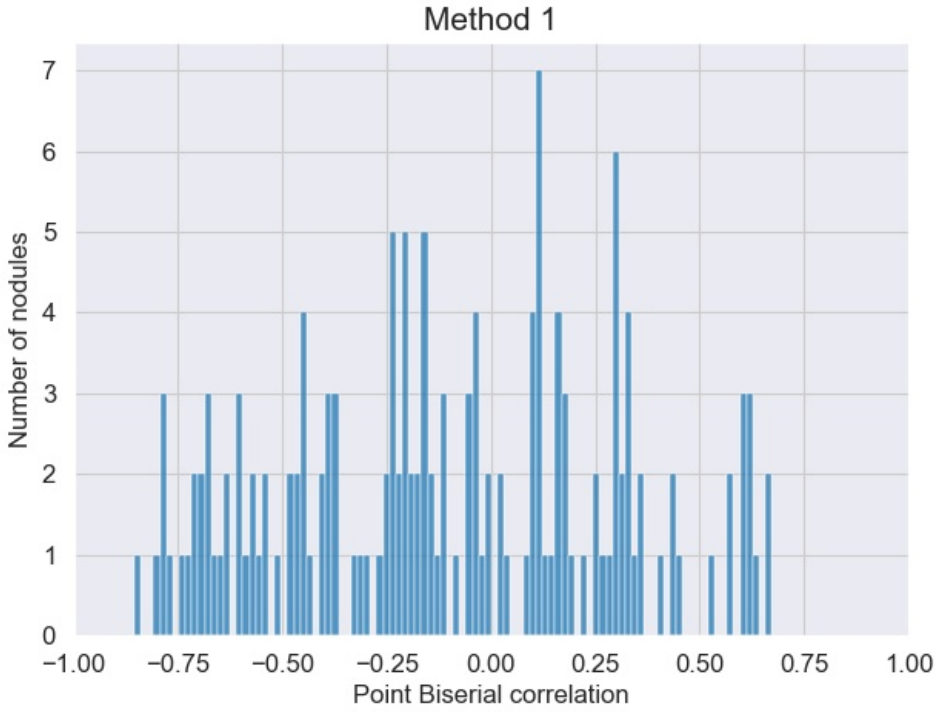

Figure 4.3: Histogram showing the distribution of the point-biserial correlation of stacks, determined by method 1.

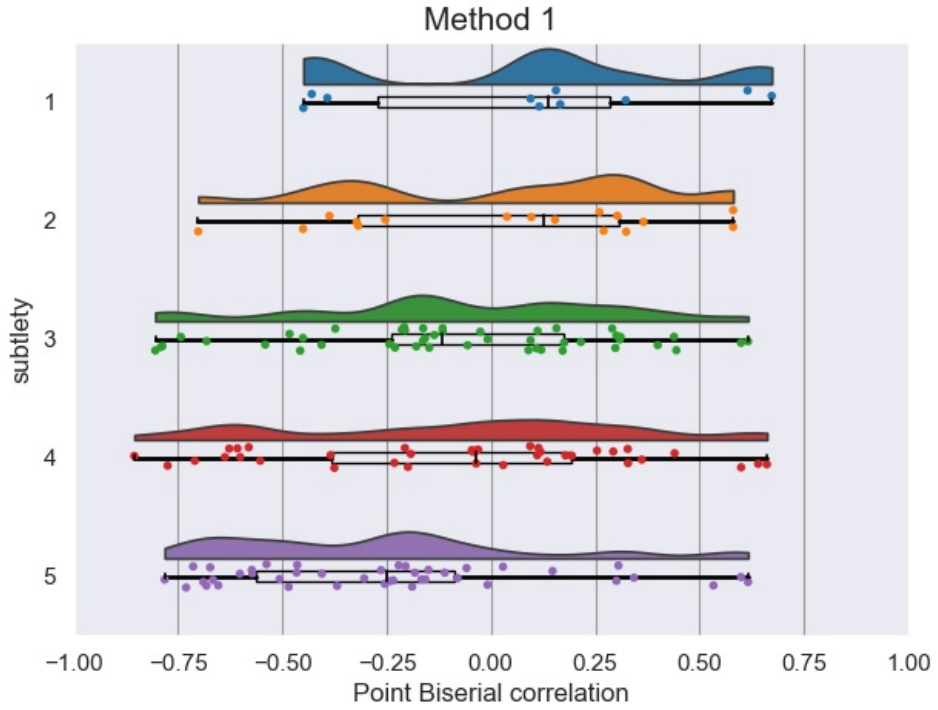

Figure 4.4: Raincloud plots of the distribution of the point-biserial correlation acquired with method 1 for each subtlety score. The top of the plot shows the distribution over the range oft the point-biserial correlation. Below the distribution a strip plot displays the individual data points. A boxplot is added to show the median and quartiles of the distribution.

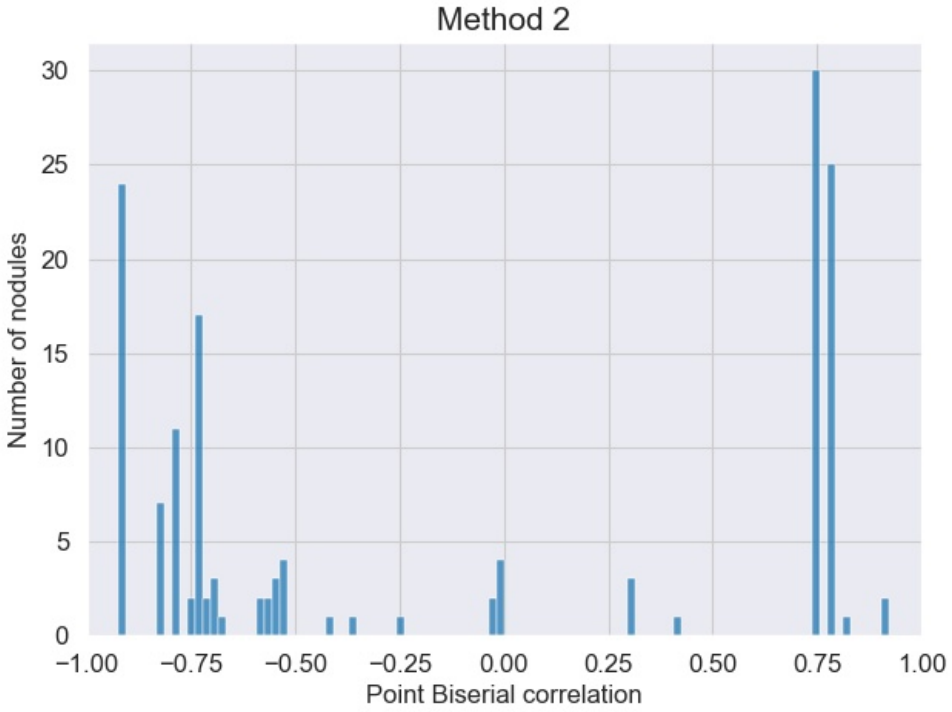

Figure 4.5: Histogram showing the distribution of the point-biserial correlation of stacks, determined by method 2.

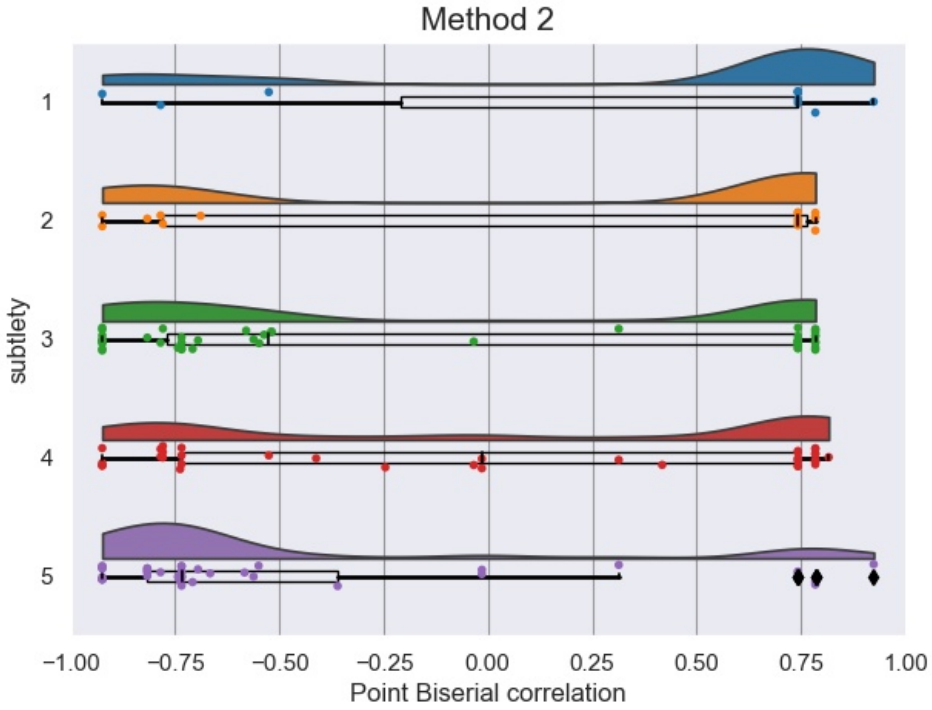

Figure 4.6: Raincloud plots of the distribution of the point-biserial correlation acquired with method 2 for each subtlety score. The top of the plot shows the distribution over the range oft the point-biserial correlation. Below the distribution a strip plot displays the individual data points. A boxplot is added to show the median and quartiles of the distribution.

4.3 Ordinal categorical logistic Regression

A ordinal categorical logistic regression model was fitted to the data acquired with method 1 and method 2, described in section 3.3.1 to research the hypothesis regarding the relationship between of the point-biserial correlation of the stack and the subtlety score. The ordinal categorical regression analysis was performed with the use of IBM SPSS statistics (IBM Corp.), where the PLUM procedure is carried out according a tutorial (Ord). Firstly, the interpretation of the model fitted to the data acquired by method 1 is given, followed by the model fitted to the data obtained by method 2 as described in section 3.3.1.

4.3.1 Regression Model for data acquired with Method 1

In 2.4.1 4 assumptions are stated, to which the data must comply. The first two assumptions are satisfied as the subtlety category is naturally ordered and the point-biserial correlation is continuous. The third assumption is relevant for two or more independent variables, which is not the case for the data. The fourth assumption is that the model gives a better fit to the data than the ordinal proportional odds. The test of parallel lines, in table 4.3, compares the ordinal logistic model with the proportional odds. Assumption 4 is rejected for a statistical better fit. From the test of parallel lines it can be concluded that assumption four is also satisfied.

In table 4.4 it can be see that the -2 Log Likelihood with the χ^2-test gives an improvement on the baseline model and therefore, the ordinal logistic model is more effective than the null model. In 4.5 the statistical test of individual predictors can be found, where the statistical significance of the individual regression coefficients is tested with the Wald chi-square statistic. The Point Biseral correlation is a significant predictor of the subtlety. The Goodness-of-Fit statistics in table 4.4 assess the fit of the model against the actual outcomes (the subtlety category). The Pearson test yields a $\chi^2(587)$ of 585.977 and is not statistically significant, suggesting that the model is a good fit to the data. The additional descriptive measures R^2 indices, Cox and Snell, McFadden and Nagelkerke are given in the caption of table 4.4. They represent the proportion of the variation in the dependent variable that can be explained by predictors in the model (30). They do not give information about the explained variance and also do not correspond to predictive efficiency. Therefore, they are merely given as an addition to the overall evaluation of the model and the Goodness-of-Fit test statistics.

The ordinal categorical logistic regression calculates a value for the logistic odds or logit of an event outcome from the predictor. The logit is the natural log of the odds, and therefore can be transformed to the odds as explained in section 2.4.1. The odds ratio shows the factor of increase in the dependent variable per unit increase of the independent variable. In SPSS this calculation is performed manually with the syntax code given in appendix 7.3.3. From table 4.5 we can see that the calculated odds ratio $e^{\beta} = 0.319$, indicating a descending predicted probability. The odds of being in higher subtlety category decreases from 1.0 to 0.319, with every unit increase of the point-biserial correlation.

Model	-2 Log Likelihood	χ^2	df	p
Null Hypothesis	435.090			
General	433.638	1.451	3	0.694

Table 4.3: Test of parallel lines for method 1 to compare the model of to the proportional odds.

Test	χ^2	df	p
-2 Log Likelihood	9.174	1	0.002
Pearson	585.977	587	0.504

Table 4.4: Model fit Goodness-of-Fit tests of Pearson Deviance and the -2 Log Likelihood of the ordinal regression analysis for data acquired by method 1. Pseudo R-squared values; Cox and Snell R^2=0.059, Nagelkerke R^2=0.062, McFadden R^2=0.021.

Predictor	β	$SE\ \beta$	Wald	p	e^{β}	Lower	Upper
Subtlety 1	-2.611	0.328	63.327	<0.001	0.073	0.039	0.140
Subtlety 2	-1.511	0.217	48.260	<0.001	0.221	0.144	0.338
Subtlety 3	0.016	0.170	0.009	0.926	1.016	0.728	1.418
Subtlety 4	1.107	1.93	32.846	<0.001	3.026	2.072	4.420
point-biserial correlation	-1.142	0.380	9.021	0.003	**0.319**	0.151	0.672

Table 4.5: Ordinal logistic regression analysis on the calculated point-biserial correlation by method 1 and subtlety score.

4.3.2 Regression Model for data acquired with Method 2

In table 4.6, 4.7 and 4.8 the results of the ordinal categorical regression model on the data acquired with the second method, described in section 3.3.1 are summarized. The test of parallel lines shows that assumption 4 is satisfied in table 4.6. In table 4.7 it can be seen that the -2 Log Likelihood with the χ^2-test gives an improvement on the baseline model and therefore, the ordinal logistic model is more effective than the null model. Like the model described in 4.3.1, the point-biserial correlation is a significant predictor of the subtlety, according to the Wald chi-square statistic in table 4.8. In table 4.7 the Pearson test yields a $\chi^2(115)$ of 101.418 and is not statistically significant, suggesting that the model is a good fit to the data. The additional descriptive measures R^2 indices, Cox and Snell, McFadden and Nagelkerke are also given in the caption of table 4.4 as a supplement to the Goodness-of-Fit test. The logit and the odds ratio e^{β} are given in table 4.8, where indication for descending probability is found. The odds of being in higher subtlety category decreases from 1.0 to 0.529, with every unit increase of the point-biserial correlation.

Model	-2 Log Likelihood	χ^2	df	p
Null Hypothesis	168.279			
General	162.752	5.527	3	0.137

Table 4.6: Test of parallel lines for method 2 to compare the model of to the proportional odds.

Test	χ^2	df	p
-2 Log Likelihood	9.909	1	0.002
Pearson	101.418	115	0.813

Table 4.7: Goodness-of-Fit tests of Pearson Deviance and the -2 Log Likelihood of the ordinal regression analysis for data acquired by method 2. Pseudo R-squared values; Cox and Snell R^2=0.064, Nagelkerke R^2=0.068, McFadden R^2=0.023.

Predictor	β	$SE\ \beta$	Wald	p	e^{β}	Lower	Upper
Subtlety 1	-2.658	0.329	65.109	<0.001	0.070	0.037	0.134
Subtlety 2	-1.602	0.222	52.209	<0.001	0.202	0.131	0.311
Subtlety 3	-0.047	0.169	0.076	0.783	0.955	0.686	1.329
Subtlety 4	1.063	0.190	31.166	<0.001	2.896	1.994	4.206
point-biserial Correlation	-0.636	0.204	9.703	0.002	**0.529**	0.355	0.790

Table 4.8: Ordinal logistic regression analysis on the calculated point-biserial by method 2 correlation and subtlety score.

5 Discussion

5.1 Interpretation of the results

5.1.1 Lung nodule detection network

Training a lung nodule detection network is an essential task to determine an objective difficulty score as a property of a lung nodule case. Many scientist and engineers have achieved to write lung nodule detection algorithms with extremely high performances, and research is continuing to improve these algorithms. Lung nodule detection from a low dose CT scan is a complex task, because of the varying sizes and density of nodules and the imbalance of voxels with nodules and without nodules. To detect nodules with algorithms, multiple preprocessing steps and two-step network architectures are applied to achieve a high sensitivity and specificity. This research did not aim for high sensitivity and specificity, but for a variation in the performances of multiple networks. Therefore, only basic preprocessing was applied to the data set. To aim for a simple model to detect nodules, a one-step detection algorithm was selected. A one-step model has the advantage that only one network needs to be trained, it is fast to train in the network and hyper parameter optimization is easily performed. The disadvantage is that the performance of a one-step network is lower compared to algorithms that are currently published. The performance of the initial model could have been higher if the model would have been based on a two-step process, where first nodule candidates are selected followed by a False Positive reduction. Additional preprocessing steps, for instance lung segmentation or removal of the vessels, could have improved the average performance of the Nets in the ensemble.

5.1.2 The Difficulty-Net ensemble

Besides the desire of a simple model, the item analysis is performed with the use of the point-biserial correlation, utilizing varying performances as a continuous value. The range in performance of the network was aimed to be wide, resulting in good performing models and bad performing models. In section 4.1 the performance of the networks in the Difficulty-Net ensemble are presented in table 4.1 and plotted in an ROC curve in figure 4.1. It was expected that the variation in performances would result from increasing size in train set. The performance ranges between 0.572 and 0.706, but does not increase or decrease, with the size of the test set. To check the influence of the increasing test set, NetF 70/30 was trained with the same parameters and same train set 2 times. Figure 4.2 shows the ROC curves, where the repetitions are similar in performance, but lower than the network used in the ensemble. The variation in performance is likely to be caused by the random initialization of the network. Increasing the size of the train does not have the desired effect on achieving an ensemble of networks with varying performances. Moreover, the neural networks in the Difficulty-Net ensemble are almost identical, which has consequences for the validity and significance of the the Item analysis. To prove the concept presented in this research, it was chosen to keep the architecture and parameters of the networks the same. Item analysis with networks constructed of different architectures and parameters could impede the interpretation and explanation of the results. Nevertheless, for future research it is suggested to use a different methods for controlling the performance of neural networks.

The Difficulty-Net ensemble is applied for item analysis where it is assumed that all individual subjects are independent. Since the Nets have the same architecture and parameters, the same classifying method and are trained on similar training data, the networks in the ensemble are not independent. To achieve more independency for the individuals in the ensemble, other machine learning and deep learning algorithms could be added. As mentioned in section 2.2,

support vector machines, linear classifiers, decision trees and other methods could be added to the ensemble to achieve more independency. By varying the approaches of the individuals in the ensemble, the quality of item analysis might improve.

5.1.3 Item analysis

This research has used item analysis to define a measure of difficulty of a radiological case. The statistical analysis applied originates from educational sciences where the quality of questions on an exam or test are evaluated. Item analysis can use multiple methods to assess the goodness of a question, but in this research, it was chosen to investigate whether the discriminating power, calculated with the point-biserial correlation, can give an indication of the level of difficulty of a radiological case. It should be noted that the discriminating power of a radiological case has an indirect relation to difficulty. Difficulty is subjective property, making the scale or measure of difficulty unique for every person. The discriminating power is determined by taking the performance of a group of individuals into consideration. The discriminating power of a question can therefore be an indicator of difficulty for the overall group or for the average performing individual. The size of the group of subjects and the variance in the performance of the subjects influences the results of the item analysis. Literature does not describe a minimal number of subjects to perform item analysis, but as a rule of thumb, 30 individuals are recommended. Literature about educational sciences describe a minimum of 30 subjects to calculate a significant value for the point-biserial correlation.

The proposed method utilizes merely 10 neural networks as subjects to calculate a value for the point-biserial correlation of a stack. This choice is sustained by the fact that training a single neural network takes time and energy. The purpose of this research is to explore whether the method is feasible and to prove the concept of creating a measure of difficulty with the use of neural networks. Moreover, the performance of the networks in the ensemble is limited due to the reasons described in section 5.1.1 The point-biserial correlation takes the variation of the performance of the networks into account by using the standard deviation of the performance of the ensemble. The value of the point-biserial correlation might describe the discriminating power of the stack more accurately in case of more neural networks. It is expected that the point-biserial correlation would be impacted significantly if multiple better performing networks and worse performing networks would be added. For this project it was chosen to keep the architecture and parameters of the networks the same. Future research could result in an ensemble of more networks, with better performance and worse performance.

The resulting measure of difficulty determined with the presented method is scaled to the performance of the individual subjects used, in this case, the neural networks. The neural networks have a relatively low performance with an average AUC of 0.667, compared to neural networks commonly described in literature. The level of difficulty is determined for all the networks in the ensemble or for the average performing network in the ensemble. To test the validity of the method and the measure of the difficulty, the results are compared to the categorical difficulty given by 4 experienced radiologists. Since the performance of the networks is not comparable to the performance of 4 experienced radiologist, the comparison is not impartial. The annotations of the radiologists are used as the ground truth, meaning that if one network in the ensemble can detect the nodules in all stacks, comparison to radiologists would be more equitable. Comparison of the level of difficulty determined by networks with an average performance, like that of experienced radiologists, would give more insight about the validity of the method presented in this project.

The determined level of difficulty could also be compared to a different group of individuals. When the stacks would be presented to a group of individuals that perform similar to the neural networks, a more equitable comparison is accomplished. A user study is suggested in section

6.2 to explore the legitimacy of the measure of difficulty determined in this research and the soundness of the method.

5.1.4 Methods for score

A dichotomous score or binary score is needed to calculate a value for the point-biserial correlation. It should be emphasized that the Difficulty-Net ensemble is used to predict the presence of a lung nodule in a slice, meaning that the 3D information of the stack is not considered. To transform the prediction per slice of the networks to a binary score for the calculation of the point-biserial correlation, two methods are described in this report. Appendix 7.2 elaborates on 2 additional methods to determine the binary score for the calculation of the point-biserial correlation. A stack contains slices with nodules (positive slices), which can be found in the center of the stack, and slices without nodules (negative slices), the peripheral slices of the stack. The network is required to perform two tasks; 1. To detect the nodule in the positive slices and 2. To not detect nodules in the negative slices. For positive slices, a score of 1 is granted if the nodule is detected, where the number of false positives is not considered. Negative slices are granted a score of 1 if 95% percent of the slice is predicted negative of nodules. Here, the confidence interval is applied, because in very few cases the networks were able to completely classify a negative slice correct. Appendix 7.3.1 shows the confusion matrices of the neural networks in the ensemble where it can be seen, that some networks perform better on positive slices and some networks perform better on negative slices. In method 1 the binary score is determined for each slice resulting in a point-biserial correlation for each slice. Assuming every slice contributes equally to the level of difficulty of the stack, the average value of the point-biserial correlations of the slices is taken. Averaging the value of the point-biserial correlation removes the extreme values and can therefore give misleading results. Figure 7.3 shows the distribution of the standard deviation of each stack. The highest density is found at a standard deviation of 0.56, indicating high variance between the point-biserial correlation of the slices within the stacks. Filtering the extreme values of individual slices can lead to misinterpretation of the relation of the point-biserial correlation and difficulty. Therefore, it is not preferable to take mean value of the point-biserial correlation of the slices in the stacks.

Method 2 bases the binary score of the stack on the binary score of the individual slices in the stack. The stack is granted with a score of 1 when 50% or more slices in the have a score of 1. The effect of a lower or higher threshold on the results of the point-biserial correlation, but also other methods to transform the prediction to a binary score could be explored in future work. The framework presented in this research uses a point-biserial correlation, which requires a binary value. Additionally, it would be interesting to search for other statistical methods that can handle the predictions of the networks directly, instead of transforming it to a binary value.

5.1.5 Ordinal regression

In section 4.3 an ordinal regression model is used to investigate the relationship of the point-biserial correlation and the subjective category of subtlety of nodules. Analysis is performed on the data acquired by method 1 and method 2. According to the Goodness of fit and overall model fit, the model for both methods are statistically significant. One should be aware of the following when drawing conclusions. In table 4.5 and table 4.8 the parameter estimates are summarized. The model predicts intercepts for subtlety category 1, 2, and 4 are significant. The intercept 3 is not statistically significant. Looking back in figure 4.4 for method 1, subtlety score 2, 3 and 4 have values over the complete range of the point-biserial correlation. Figure 4.6 for method 2 show for subtlety scores 2, 3 and 4 values of the point-biserial correlation around -0.75 and 0.75. The distributions over the point-biserial correlation are relatively similar to one another. Therefore, it makes sense that the ordinal logistic regression did not lead to a significant predictor for subtlety category 3. Based on the ordinal logistic regression analysis the two

methods cannot be compared to one another, as similar results are found. Section 3.3.1 and 5.1.4 describe and discuss method 1 where the mean value of the point biserial correlation is taken. Taking the mean over correlation values is regarded as controversial. Therefore, the results of the ordinal logistic regression analysis for method 2 are considered promising, compared to the results of method 1. When looking at the subtlety scores for each of the stacks, there is an imbalance in the number of data points per subtlety category. Figure 3.4 shows there are only ten nodules in category one while up to 30 nodules in category 3, 4 and 5. More homogeneous data set, where the number of nodules per category are similar, might have lead to a better fit of the ordinal logistic regression model.

5.2 Limitations

Data

The preprocessing steps that were undertaken result in some limitations to the research. The normalization of the voxel size to 1x1x1 mm needs interpolation. Scans with a transversal slice thickness of more than 3 mm were deemed to be excluded, because the interpolation would result in a manifold of transversal slices containing the same information. The manual exclusion could have been avoided if the LUng Nodule Analysis 2016 (LUNA16) dataset, an extract of the LIDC/IDRI dataset would have been used. The LUNA16 dataset excludes scans with nodules smaller then 3 mm and scans with a slice thickness larger than 3 mm. The data set also excluded nodules annotated by 1 or 2 radiologists, which were considered irrelevant findings (33). For the method of determining a level of difficulty of lung nodule cases, it should be noted that the subtlety score of nodules is coherent with the number radiologist annotating nodules (refer to 2D histogram!!!). To validate the level of difficulty, comparison with the opinion of all 4 radiologists is preferred. Therefore, the choice for the LIDC/IDRI dataset prevailed. The aim of the research is to define a measure of difficulty of a lung nodule case. The measure of difficulty for a complete CT scan, with the use of deep learning, gives rise to multiple problems. Firstly, complete CT scans cannot be processed by neural networks, as this is computationally too heavy with the contemporary computational power. Secondly, considering a thoracic CT scan, different locations within a scan can have a varying level of difficulty. It is unknown how the general difficulty of a scan is influenced by specific locations. To illustrate, healthy tissue in a CT scan also possesses a level of difficulty with respect to pathologies. artefacts visible in specific locations of the scan can influence the overall difficulty of a lung nodule in a CT scan. that can be visible section 1.2.1 describes that healthy cases are often regarded as difficult for inexperienced radiology residents. To maintain a Cartesian approach, the level of difficulty of a nodule in a stack of 10 slices was determined. To use the presented approach in the future for a computer assisted learning system, it was decided to determine the level of difficulty of 3D data instead of 2D slices. Detection of lung nodules from 2D slices is challenging for human subjects, as the 3D information of a lung CT scan is used to discriminate lung nodules from vessels. The chosen neural network architecture takes 2D transversal slices as input leading to prediction of lung nodule location in 2D. Thus, the method presented in this framework indirectly determines the level of difficulty of a lung nodule in a stack of 10 slices, by assessing it slice by slice. For the development of a supervised computer aided detection system the data is labelled. In this research, any annotation out of 4 experienced radiologists are assumed to be the ground truth. The From 157 patients the true diagnosis is recorded. How the true diagnosis relates to radiological errors in terms of lung nodule detection is not investigated in the scope of this research.

lung nodule detection network

Most lung nodule detection networks use a two-step architecture to achieve high sensitivity and specificity. In section 5.1.1 the choice of a one step process is explained. In the preprocessed data, the lung nodules are small, resulting in a disbalance of data with a nodule and

data without a nodule. The one-step network architecture that was selected for the task of lung nodule detection is not adequate enough to achieve performance similar to two step processes. Most likely, the neural network is not able to sufficiently learn the features of small nodules. To achieve this, a more advanced and complex one-step network should be used to be able to detect lung nodules.

Difficulty-Net ensemble
As mentioned in section 5.1.1, the neural networks in the Difficulty-Net ensemble ought to be independent. Since they share their architecture, parameters and they are trained on the same dataset, their performance is not varying as was desired. The concept of item analysis with the use of machine learning or deep learning has not been performed. To set up a baseline experiment for a measure of difficulty, determined with the use of deep learning, the architecture and parameters should not be varied. to be able to set a baseline experiment. Future work could use different machine learning and deep learning methods to see its effects on the item analysis and the measure of difficulty.

Item analysis
The item analysis conducted in this research, uses the point-biserial correlation. A correlation value between a continuous value (performance of neural networks) and a binary value (score on stack). The output of a neural network consists of a prediction on the location of the nodule in the slice. To calculate a point-biserial correlation, the prediction must be transformed into a binary score. This transformation is based on formulated restrictions, which could be reformulated in multiple ways. The transformation takes the prediction of each slice separately and determines a binary score for each slice. This might influence the outcome of the value of the point-biserial correlation.

Validation
The subtlety scores provided by the LIDC/IDRI dataset are used to validate the method. Each radiologist annotating a nodule is asked to give a subtlety score for the nodule. The annotations of the nodules are clustered to its corresponding nodule. The set of 151 stacks contains nodules that are annotated by all radiologists, but also by solely by one radiologist. For this research only one subtlety score per nodule has been taken into consideration. In a later stage, the other subtlety scores were analysed and compared with one another. The mean subtlety scores and the standard deviation are shown in the appendix in figure 7.8 and figure 7.9. The distributions are similar, and it shows for most cases a standard deviation of 0, indicating perfect agreement between the radiologists. Few cases have a standard deviation higher than 1. In future research the point-biserial correlation should be compared to the annotation of all annotating radiologists.

6 Conclusion

6.1 Conclusions

The method presented in this research aims to define a measure of difficulty of lung nodule cases as a property of the scan. A method as such can be useful for image synthesis for computer-assisted learning in radiology education to optimize the learning process and to fit personalized training to knowledge gaps. The research has been guided by two research objectives. The first research objective is; How can item analysis with the use of deep learning determine a level of difficulty for lung nodule cases? was approached in the following way. By training 10 one-step deep neural networks to detect lung nodules in slices of lung CT scans, the networks predict the location of the nodules. Two methods were used to transform predictions of the networks are transformed to binary scores for Item analysis. For each item in the data set of 151 stacks, consisting of 10 transversal CT slices, a value of the point-biserial correlation was calculated. The relationship between the point-biserial correlation and the subtlety scores is the focus of the second research objective: How does a measure of difficulty, as a property of the scan, relate to the subjective difficulty of a lung nodule case? Ordinal categorical logistic regression is applied to analyze the relationship between the point-biserial correlation and the subtlety scores. From the results in chapter 4 it can be concluded that positive values of the point-biserial correlation are related to “subtle” nodules, while negative values of the point-biserial correlation can be related to “obvious” nodules. Given the above, a measure of difficulty is defined, with the use of item analysis and deep learning. As the method applied item analysis with supervised neural networks, the difficulty is scaled to the annotations of 4 experienced radiologists. Therefore, the measure of difficulty of the lung nodule case is not a property of the scan. Nevertheless, the measure of difficulty has the potential to be applied to image synthesis for the design of a computer-assisted learning system for radiology education.

6.2 Recommendations and Future Work

The architecture described in 3.2.1, gives a 16 by 16 pixels prediction map. The output is down sampled and does not give the exact bounding box of the nodule, which also would indicate the size, but the location of a 32 by 32 pixel square on the slice. A second network could be added to the network pipeline that reduces the false positives. The network selects a 32 by 32 pixel square with a potential nodule and classifies it as a True Positive or False Positive. This might improve the performance of the networks, but also offers another opportunity to vary the performance of the Nets in the ensemble. In this research, it was attempted to vary the performance of the Nets by adding more data to the train set. The same could be done for the false positive reducing network. The combination of the Net in the Difficulty-Net ensemble and the false positive reducing network could also create variation in the performance.

The method presented in this research uses neural networks to determine a point-biserial correlation which can be related to a measure of difficulty of a stack of lung CT slices. The values of the point-biserial correlation are dependent on the performance of the neural networks and therefore scaled and limited to the performance of the neural networks. In the introduction, it was mentioned that the level of difficulty is useful for radiology education. Yet, a level of difficulty is subjective and a scale of difficulty might vary for groups with a specific level of knowledge and skill. The measure of difficulty might be different for first-year medical students and that of experienced radiologists. In other words, can one define a scale of difficulty that includes experienced radiologists and inexperienced students? In this framework, the difficulty is based upon neural networks and their performance on a lung nodule detection task. The relation to difficulty was investigated by validation with the subtlety score of an experienced radiologist. To see how the difficulty measure developed in this research relates to for

example medical students, a validation study is suggested. With a simple graphical user interface, the stacks can be presented to a group of medical students. They are asked to detect the nodule and indicate the square in which the nodule is present. Item analysis can be performed retrospectively and the values of the point-biserial correlation can be compared to the values determined in this framework. This will add to the validity of the developed method to determine a measure of difficulty and can give insights into the differences in interpretation between medical students and the used neural networks in the method.

For future work, a measure of difficulty of radiology cases can be particularly useful for radiology education. Computer-assisted learning has been proven to be a good addition to traditional radiology education. The development of generative adversarial networks has made it possible to generate medical images from existing data sets, that can be used for radiology education. When the level of difficulty or complexity of the generated slices can be guided and controlled, the data can be adapted to the individual's level. This offers the opportunity for computer-assisted learning systems to be fully personalized and fitting to the student's knowledge gap. Implementation of the presented framework to the workflow of GANs should be investigated and validated. By implementing the presented framework to the workflow of image synthesis with GANs, computer-assisted learning can be lifted to the next level regarding personalized education.

7 Appendix 1

7.1 Descriptive values of data set

The thickness of all nodules in the data set are displayed in a histogram in figure 7.1. The median of the thickness is 6 slices per nodule. The upper quartile of of this data is at 11 transversal slices per nodule, meaning that 75% of the nodules is smaller than 10 slices. The median is at 8 transversal slices per nodule.

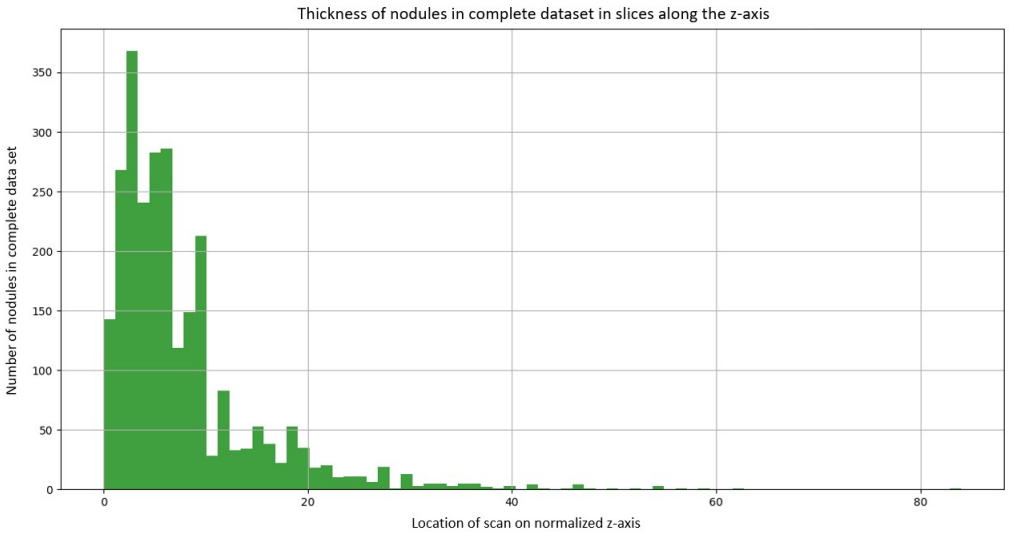

Figure 7.1: Thickness of the nodules in number of transversal slices in the total data set.

Figure 7.2 shows the location of all nodules in the data set on a normalized height scale. Not every scan has the same number of slices. To be able to compare the location of the nodules, the number of slices were normalized, were 0 indicates the bottom-most transversal slice and 100 indicates the upper transversal slice. According to the histogram, nodules can be found in each location along the cephalad direction.

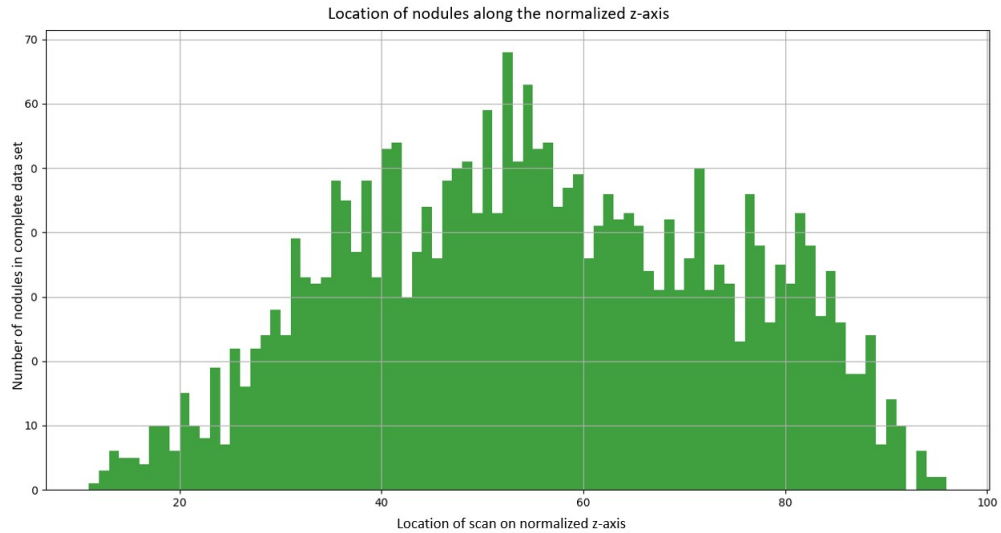

Figure 7.2: Location of nodules on a normalized height, where 0 indicates the bottom-most transversal slices of the scan and 100 indicates the upper transversal slice of the scan.

Table 7.1 gives an overview of the number of patients, stacks and slices within the train and test set.

	Ratio	Patients	Stacks	Slices
Total	100		810	8100
Train set	80%		659	6590
Test set	20%		151	1510

Table 7.1: Overview of the number of patients stacks and slices for the train and test set.

7.2 Additional Methods Binary score and results

In method 1 described in 3.3.1, the mean value of the point-biserial correlation over the whole stack is taken. Figure 7.3 shows how the standard deviation of each stack is distributed.

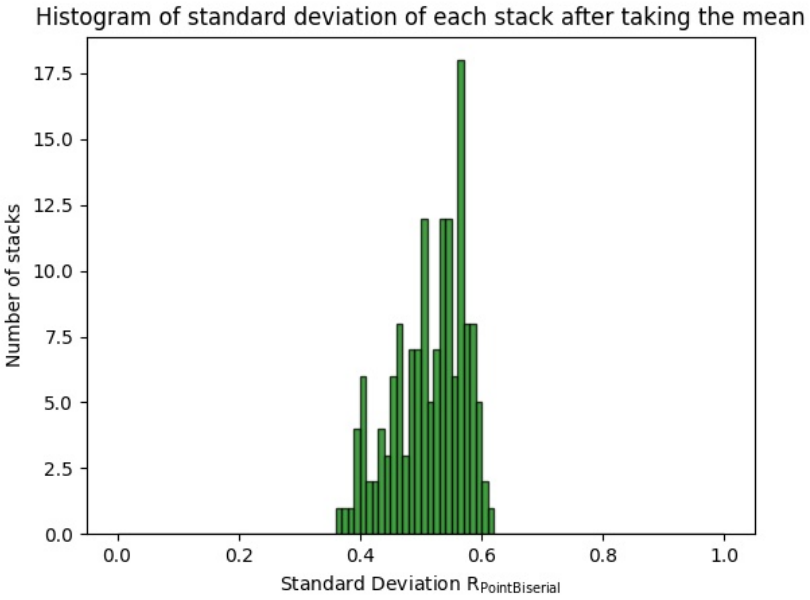

Figure 7.3: Distribution showing the standard deviation of the point-biserial correlation calculated by method 1 of each stack.

To transform the predictions of the neural networks to a binary score serving the calculation of a point-biserial correlation, 2 other methods were explored. These transformations showed less promising results, and were therefore not further investigated. The two additional methods 3 and 4 are described below and the results can be found in 7.3.3.

Method 3 Method 3 first determines a binary score of the whole stack and calculates the point-biserial correlation. The assumption is here is that if the network was able to detect the nodule correctly in any slice of the stack, a score of 1 will be granted for the whole stack. If the nodule was not detected in any of the slices of the stack, a score of 0 was given. The binary scores are stored in a matrix, where the number of of rows is equal to the number of stacks, and the number of columns is equal to the number of nodules. The point-biserial correlation is calculated for each stack.

Method 4 Similar to method 3, the binary score is determined by the binary score of the individual slices. If the slice is positive and the nodule is detected, a score of 1 will be granted. In case of a positive slice and the nodule is not detected a score of 0 will be given. A score of 1 will be granted when 95% of the squares are assigned negative. A score of 0 will be given, when less than 95% of the squares are assigned negative.

The score of the whole stack is determined by the score of two consecutive slices. If two consecutive slices both have a score of 1, the score of the stack will be one. If no consecutive slices both have a score of 1, the score of the stack will be zero.

7.3 Difficulty-Net results

7.3.1 Tables of confusion

Table 7.2 till table 7.11 show the confusion matrices of the networks giving an indication of the performance of the network.

NetA 100	Predicted Positive	Predicted Negative
Actual Positive	246	609
Actual Negative	31051	354654

Table 7.2: Confusion matrix of NetA 100.

NetB 90/10	Predicted Positive	Predicted Negative
Actual Positive	855	0
Actual Negative	385699	6

Table 7.3: Confusion matrix of NetB 90/10.

NetC 85/15	Predicted Positive	Predicted Negative
Actual Positive	217	638
Actual Negative	20983	364722

Table 7.4: Confusion matrix of NetC 85/15.

NetD 80/20	Predicted Positive	Predicted Negative
Actual Positive	637	218
Actual Negative	315746	69959

Table 7.5: Confusion matrix of NetD 80/20.

NetE 75/25	Predicted Positive	Predicted Negative
Actual Positive	100	755
Actual Negative	4819	380886

Table 7.6: Confusion matrix of NetE 75/25.

NetF 70/30	Predicted Positive	Predicted Negative
Actual Positive	40	815
Actual Negative	1036	384669

Table 7.7: Confusion matrix of NetF 70/30.

NetE 65/35	Predicted Positive	Predicted Negative
Actual Positive	13	842
Actual Negative	513	385192

Table 7.8: Confusion matrix of NetG 65/35.

NetH 65/35	Predicted Positive	Predicted Negative
Actual Positive	22	833
Actual Negative	614	385091

Table 7.9: Confusion matrix of NetH 60/40.

NetI 50/50	Predicted Positive	Predicted Negative
Actual Positive	5	850
Actual Negative	160	385545

Table 7.10: Confusion matrix of NetI 50/50.

NetJ 45/55	Predicted Positive	Predicted Negative
Actual Positive	14	841
Actual Negative	189	385516

Table 7.11: Confusion matrix of NetJ 45/55.

7.3.2 Additional results point-biserial correlation

Method 3
The following method first determines a score of the whole stack, based on the predictions of the network on the stack. The score and the performance of the Nets are used to calculate a point-biserial correlation of the stack. The transformation of the predictions of the Difficulty-Net is described in 3.3.1. In figure 7.4 the distribution of the values of the point-biserial correlation can be found. All calculated values for the point-biserial correlation are negative. The raincloud plot therefore shows the highest densities for negative values of the point-biserial correlation. a negative values indicate that good performing networks did not correctly classify the stack whilst bad performing networks did classify correct by chance. The distributions of the point-biserial correlation for each subtlety score look very similar, therefore ordinal logistic regression analysis was not applied.

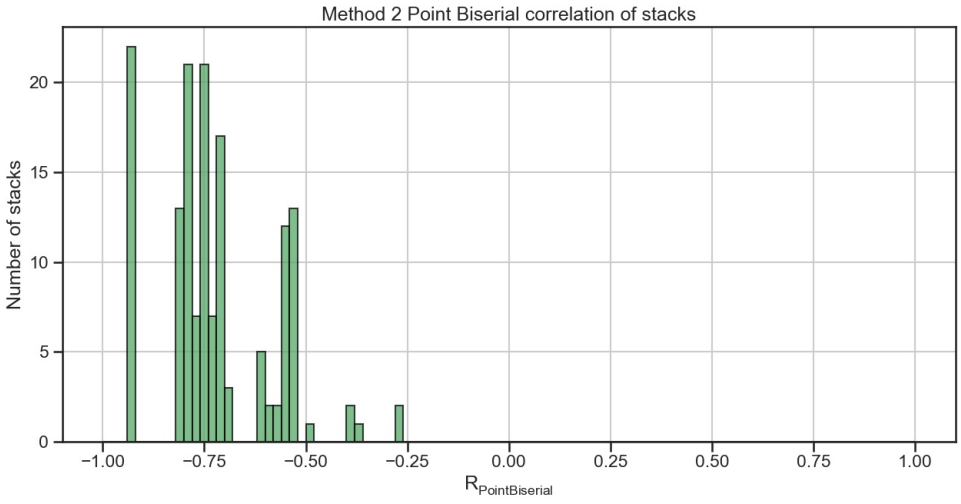

Figure 7.4: Distribution of the point-biserial correlation, determined by method 2, of the stacks in the test set.

Method 4 Figure 7.6 shows the histogram of the point-biserial correlation calculated by method 4 described in 3.3.1. The histogram shows that many stacks have a point-biserial correlation of 0. There some stacks that have a positive or negative point-biserial correlation. From the raincloud plots of figure 7.7, one can see that subtlety score 2 and 3 have mostly values of 0, with some outliers, when looking at the boxplot. For every subtlety score the median can be

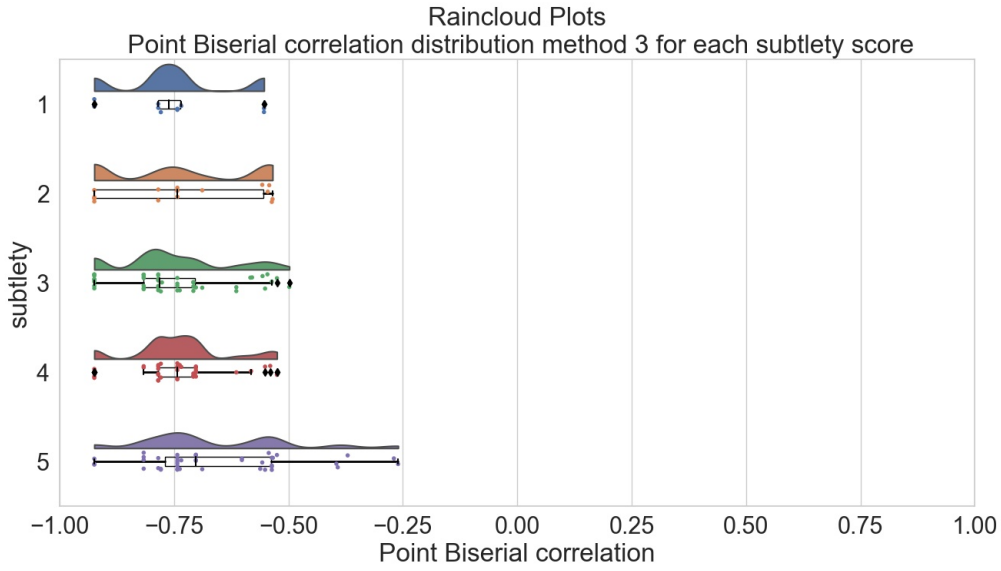

Figure 7.5: Raincloud plots of the point-biserial correlation for each subtlety score, determined by method 3.

found around 0. There is not much variation in the distribution of the point-biserial correlation for each subtlety score. For this reason, no ordinal logistic regression was performed.

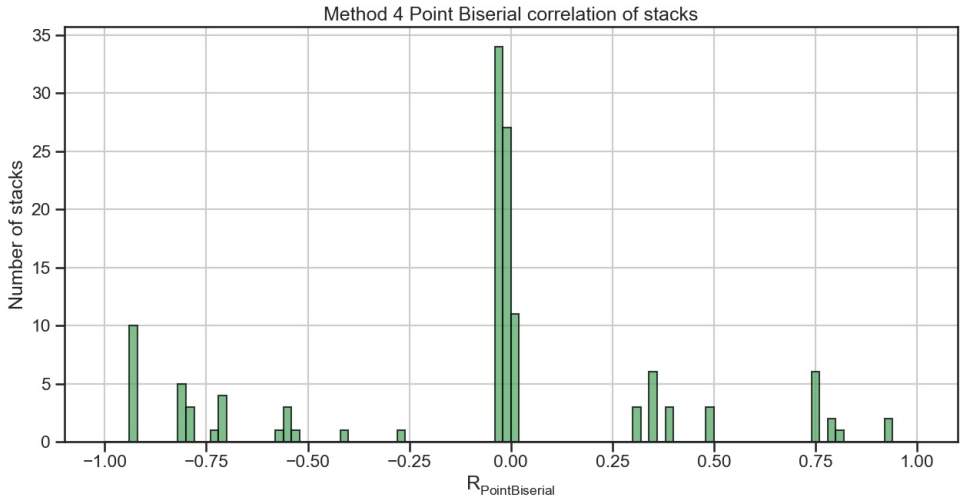

Figure 7.6: Distribution of the point-biserial correlation, determined by method 4, of the stacks in the test set.

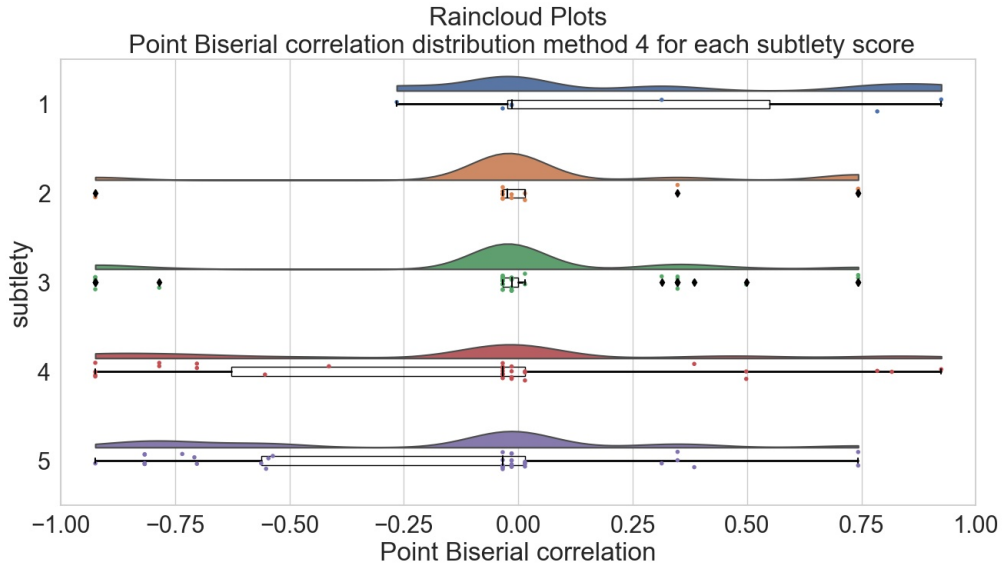

Figure 7.7: Raincloud plots of the point-biserial correlation for each subtlety score, determined by method 4.

7.3.3 Ordinal Regression method

Below, one can find the syntax used to calculate the odds ratio from the logits in SPSS.

```
COMPUTE Exp_B = EXP(Estimate).
COMPUTE Lower = EXP(LowerBound).
COMPUTE Upper = EXP(UpperBound).
FORMATS Exp_B Lower Upper (F8.3).
EXECUTE.
```

7.3.4 Subtlety scores

Figure 7.8 show the mean values of the subtlety scores of all annotating radiologists. The standard deviation of the subtlety score for each nodule is shown in figure 7.9.

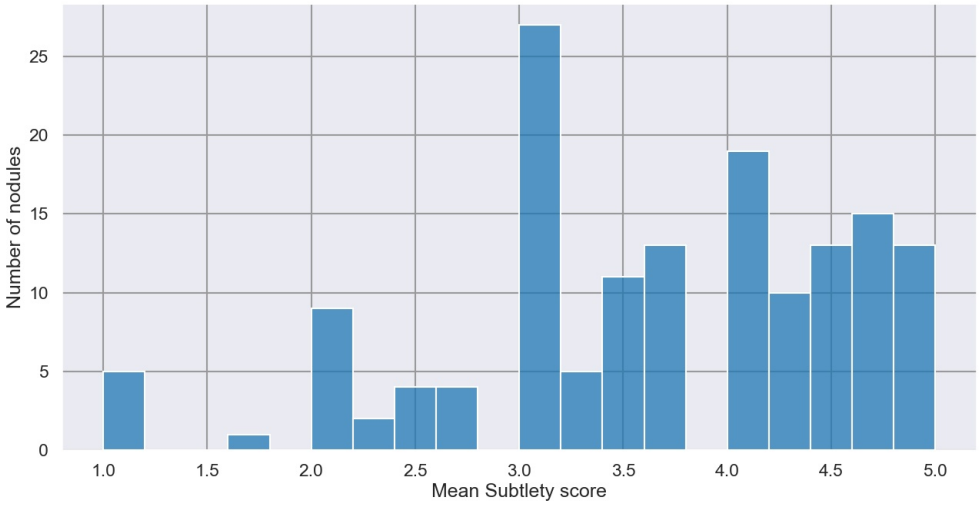

Figure 7.8: Mean values of subtlety scores of all annotating radiologists.

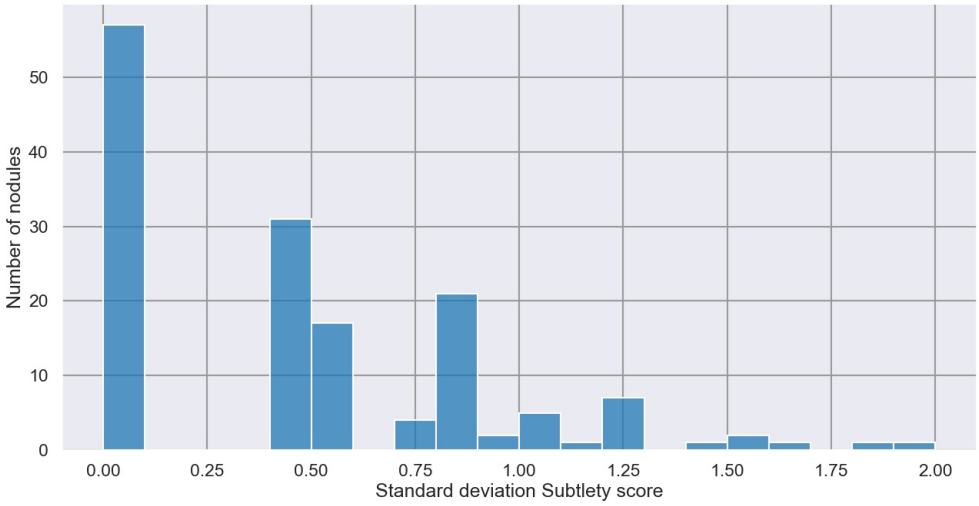

Figure 7.9: Standard deviation of mean values of subtlety scores of all annotating radiologists.

Bibliography

[mer] (), Difficulty, *Merriam-Webster*. Accessed June 6, 2022 [Online].
https://www.merriam-webster.com/dictionary/difficulty

[Ord] (), Ordinal regression using SPSS statistics (cont...), *Knowing what to Interpret from an Ordinal Regression | Laerd Statistics*.
https://statistics.laerd.com/spss-tutorials/ordinal-regression-using-spss-statistics-3.php

[wor] (), The top 10 causes of death.
https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[4] (2011), Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening, **vol. 365**, no.5, pp. 395–409, doi:10.1056/NEJMoa1102873, pMID: 21714641.
https://doi.org/10.1056/NEJMoa1102873

[5] Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng (2015), TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org.
https://www.tensorflow.org/

[6] Agresti, A. (2010), *Analysis of Ordinal Categorical Data*, Wiley.

[7] Armato III, S. G., G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman et al. (2011), The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans, **vol. 38**, no.2, pp. 915–931.

[8] Blandin Knight, S., P. A. Crosbie, H. Balata, J. Chudziak, T. Hussell and C. Dive (2017), Progress and prospects of early detection in lung cancer, **vol. 7**, no.9, p. 170070.

[9] Bruno, M. A., E. A. Walker and H. H. Abujudeh (2015), Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction, **vol. 35**, no.6, pp. 1668–1676.

[10] Cao, H., H. Liu, E. Song, G. Ma, X. Xu, R. Jin, T. Liu and C.-C. Hung (2020), A two-stage convolutional neural networks for lung nodule detection, **vol. 24**, no.7, pp. 2006–2015.

[11] Chollet, F. et al. (2015), Keras.
https://github.com/fchollet/keras

[12] George, J., S. Skaria, V. Varun et al. (2018), Using YOLO based deep learning network for real time detection and localization of lung nodules from low dose CT scans, in *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, International Society for Optics and Photonics, p. 105751I.

[13] Van der Gijp, A., M. Van der Schaaf, I. Van der Schaaf, J. Huige, C. Ravesloot, J. Van Schaik and T. J. Ten Cate (2014), Interpretation of radiological images: towards a framework of knowledge and skills, **vol. 19**, no.4, pp. 565–580.

[14] Glass, G. and K. Hopkins (1996), Statistical methods in education and psychology, **vol. 41**, no.12.

[15] Gu, Y., J. Chi, J. Liu, L. Yang, B. Zhang, D. Yu, Y. Zhao and X. Lu (2021), A survey of computer-aided diagnosis of lung nodules from CT scans using deep learning,

Computers in Biology and Medicine, p. 104806.

[16] Haladyna, T. and S. Downing (2006), *Handbook of Test Development (Educational Psychology Handbook)*, Routledge, 1 edition.

[17] Hancock, M. C. and J. F. Magnan (2016), Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods, **vol. 3**, no.4, pp. 1 – 15, doi:10.1117/1.JMI.3.4.044504. https://doi.org/10.1117/1.JMI.3.4.044504

[18] Hansell, D. M., A. A. Bankier, H. MacMahon, T. C. McLoud, N. L. Muller and J. Remy (2008), Fleischner Society: glossary of terms for thoracic imaging, **vol. 246**, no.3, pp. 697–722.

[19] Huang, X., J. Shan and V. Vaidya (2017), Lung nodule detection in CT using 3D convolutional neural networks, in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, IEEE, pp. 379–383.

[IBM Corp.] IBM Corp. (), IBM SPSS Statistics for Windows. https://hadoop.apache.org

[21] Jenuwine, N. M., S. N. Mahesh, J. D. Furst and D. S. Raicu (2018), Lung nodule detection from CT scans using 3D convolutional neural networks without candidate selection, in *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, International Society for Optics and Photonics, p. 1057539.

[22] John, L. J. (2013), A review of computer assisted learning in medical undergraduates, **vol. 4**, no.2, p. 86.

[23] Khosravan, N. and U. Bagci (2018), S4ND: Single-shot single-scale lung nodule detection, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp. 794–802.

[24] Kim, Y. W. and L. T. Mansfield (2014), Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors, **vol. 202**, no.3, pp. 465–470.

[25] Liu, C., S.-C. Hu, C. Wang, K. Lafata and F.-F. Yin (2020), Automatic detection of pulmonary nodules on CT images with YOLOv3: development and evaluation using simulated and patient data, **vol. 10**, no.10, p. 1917.

[26] Loverdos, K., A. Fotiadis, C. Kontogianni, M. Iliopoulou and M. Gaga (2019), Lung nodules: A comprehensive review on current approach and management, **vol. 14**, no.4, p. 226.

[27] Mastouri, R., N. Khlifa, H. Neji and S. Hantous-Zannad (2020), Deep learning-based CAD schemes for the detection and classification of lung nodules from CT images: A survey, **vol. 28**, no.4, pp. 591–617.

[28] Matlock-Hetzel, S. (1997), Basic Concepts in Item and Test Analysis.

[29] Nakrani, M. G., G. S. Sable and U. B. Shinde (2021), A Comprehensive Review on Deep Learning Based Lung Nodule Detection in Computed Tomography Images, *Intelligent System Design*, pp. 107–116.

[30] Peng, C.-Y. J., K. L. Lee and G. M. Ingersoll (2002), An Introduction to Logistic Regression Analysis and Reporting, **vol. 96**, no.1, pp. 3–14, doi:10.1080/00220670209598786. https://doi.org/10.1080/00220670209598786

[31] Redmon, J. and A. Farhadi (2018), Yolov3: An incremental improvement, *arXiv preprint arXiv:1804.02767*.

[32] Schittek, M., N. Mattheos, H. Lyon and R. Attström (2001), Computer assisted learning. A review, **vol. 5**, no.3, pp. 93–100.

[33] Setio, A. A. A., A. Traverso, T. de Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, R. van der Gugten, P. A. Heng, B. Jansen, M. M. de Kaste, V. Kotov, J. Y.-H. Lin, J. T. Manders, A. Sóñora-Mengana, J. C. García-Naranjo, E. Papavasileiou, M. Prokop, M. Saletta, C. M. Schaefer-Prokop, E. T. Scholten, L. Scholten, M. M. Snoeren, E. L. Torres, J. Vandemeulebroucke, N. Walasek, G. C. Zuidhof, B. van Ginneken and C. Jacobs (2017), Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge, *Medical Image Analysis*, **vol. 42**, pp. 1–13, ISSN 1361-8415, doi:https://doi.org/10.1016/j.media.2017.06.015. https://www.sciencedirect.com/science/article/pii/S1361841517301020

[34] Shaikh, F., F. Inayat, O. Awan, M. D. Santos, A. M. Choudhry, A. Waheed, D. Kajal and S. Tuli (2017), Computer-assisted learning applications in health educational informatics: a review, **vol. 9**, no.8.

[35] Skandarani, Y., P.-M. Jodoin and A. Lalande (2021), Gans for medical image synthesis: An empirical study, *arXiv preprint arXiv:2105.05318*.

[36] Software, N. S. (2021), Point-Biserial and Biserial Correlations.

[37] Spinnato, P. (2021), Low-dose computed tomography screening proposal for the "Big-3 Diseases": Lung cancer, chronic obstructive pulmonary disease, and cardiovascular disease, **vol. 28**, no.1, pp. 46–48.

[38] Sreekumar, A., K. R. Nair, S. Sudheer, H. G. Nayar and J. J. Nair (2020), Malignant Lung Nodule Detection using Deep Learning, in *2020 International Conference on Communication and Signal Processing (ICCSP)*, IEEE, pp. 0209–0212.

[39] Stewart, B. W., P. Kleihues et al. (2003), World cancer report.

[40] Ullah, M. I. and S. K. Kuri (2020), Lung nodule Detection and Classification using Deep Neural Network, in *2020 IEEE Region 10 Symposium (TENSYMP)*, IEEE, pp. 1062–1065.

[41] Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors (2020), SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nature Methods*, **vol. 17**, pp. 261–272, doi:10.1038/s41592-019-0686-2.

[42] Waite, S., J. Scott, B. Gale, T. Fuchs, S. Kolla and D. Reede (2017), Interpretive error in radiology, **vol. 208**, no.4, pp. 739–749.

[43] Wang, X., K. Mao, L. Wang, P. Yang, D. Lu and P. He (2019), An appraisal of lung nodules automatic classification algorithms for CT images, **vol. 19**, no.1, p. 194.

[44] Xie, H., D. Yang, N. Sun, Z. Chen and Y. Zhang (2019), Automated pulmonary nodule detection in CT images using deep convolutional neural networks, *Pattern Recognition*, **vol. 85**, pp. 109–119.

[45] Zhu, W., C. Liu, W. Fan and X. Xie (2018), Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification, in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, pp. 673–681.