Journal of
Clinical Medicine

MDPI

Article

Utilizing ChatGPT for Curriculum Learning in Developing a Clinical Grade Pneumothorax Detection Model: A Multisite Validation Study

Joseph Chang [1,2], Kuan-Jung Lee [2], Ti-Hao Wang [2,3,4,*] and Chung-Ming Chen [1,*]

[1] Department of Biomedical Engineering, College of Medicine and College of Engineering, National Taiwan University, No. 1, Sec. 1, Jen-Ai Road, Taipei 100, Taiwan
[2] EverFortune.AI Co., Ltd., Taichung 403, Taiwan
[3] Department of Medicine, China Medical University, Taichung 404, Taiwan
[4] Department of Radiation Oncology, China Medical University Hospital, Taichung 404, Taiwan
* Correspondence: thothwang@gmail.com or 030860@tool.caaumed.org.tw (T.-H.W.); chung@ntu.edu.tw (C.-M.C.)

Abstract: Background: Pneumothorax detection is often challenging, particularly when radiographic features are subtle. This study introduces a deep learning model that integrates curriculum learning and ChatGPT to enhance the detection of pneumothorax in chest X-rays. **Methods**: The model training began with large, easily detectable pneumothoraces, gradually incorporating smaller, more complex cases to prevent performance plateauing. The training dataset comprised 6445 anonymized radiographs, validated across multiple sites, and further tested for generalizability in diverse clinical subgroups. Performance metrics were analyzed using descriptive statistics. **Results**: The model achieved a sensitivity of 0.97 and a specificity of 0.97, with an area under the curve (AUC) of 0.98, demonstrating a performance comparable to that of many FDA-approved devices. **Conclusions:** This study suggests that a structured approach to training deep learning models, through curriculum learning and enhanced data extraction via natural language processing, can facilitate and improve the training of AI models for pneumothorax detection.

Keywords: artificial intelligence; deep learning; curriculum learning; pneumothorax

check for updates

Citation: Chang, J.; Lee, K.-J.; Wang, T.-H.; Chen, C.-M. Utilizing ChatGPT for Curriculum Learning in Developing a Clinical Grade Pneumothorax Detection Model: A Multisite Validation Study. *J. Clin. Med.* **2024**, *13*, 4042. https://doi.org/10.3390/jcm13144042

Academic Editor: Enrico M. Clini

Received: 31 May 2024
Revised: 5 July 2024
Accepted: 7 July 2024
Published: 10 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Pneumothorax is a thoracic disease defined as the presence of air in the pleural space between the lungs and the chest wall. This condition has a significant incidence in both primary spontaneous pneumothorax and secondary spontaneous pneumothorax; its detection is critical, especially given that the high risk and recurrence in the secondary form is associated with underlying lung diseases, complicating both diagnosis and management [1,2]. Secondary spontaneous pneumothorax, in particular, is a potentially life-threatening condition that demands immediate clinical attention, as recommended by the American College of Chest Physicians (ACCP) [3–5]. Chest X-ray (CXR) imaging remains the primary diagnostic tool for identifying pneumothorax, offering quick assessment. However, the interpretation of these images can be challenging; subtle signs of pneumothorax may be missed, with a reported miss rate of up to 20% for occult pneumothoraces [6]. Variabilities in thoracic disease presentations and patient postures can distort CXR images, complicating the detection process further [7–10].

In recent years, advances in deep learning (DL) have shown promising improvements in the automated detection of pneumothorax or other thoracic conditions, with studies leveraging large, open-source datasets like ChestX-ray14 (NIH) and CheXpert that demonstrate promising performances [11–13]. However, the effectiveness of these advances can be compromised by a failure to differentiate between simple and hard samples during

training, a distinction crucial for achieving robust, generalizable models. This issue is compounded by the common practice of internal validation, which may not adequately represent real-world clinical settings and could lead to potential overfitting and limited generalizability [14–16].

To address these challenges, this study proposes a novel integration of ChatGPT with a curriculum learning framework to refine the training of DL model for pneumothorax detection. Recent studies have demonstrated that ChatGPT can effectively interpret and handle clinical data to support medical education and clinical practice, demonstrating its ability in managing complex medical terminologies [17–21]. Utilizing ChatGPT's capabilities, we extracted and curated "hard samples" from radiology reports. These hard samples were based on pneumothorax size, a critical factor according to the ACCP guidelines for assessing severity and treatment pathways [22,23]. This curated approach enhances the training samples based on difficulty and informativeness, which is aligned with clinical realities where smaller pneumothoraces, although harder to detect, carry significant clinical risks if missed [24].

The utilization of curriculum learning (CL), a proven educational methodology, in our model's training regimen allows for a systematic increase in diagnostic challenges, starting with larger, easily detectable pneumothoraces and gradually incorporating smaller, more subtle cases [25–27]. This progression mirrors the clinical urgency to detect all pneumothoraces, irrespective of size, and promotes a more natural, human-like learning process in training the AI algorithm. Furthermore, the integration of ChatGPT in extracting pneumothorax size from unstructured radiology text addresses the potential to enrich datasets for DL training, providing a novel approach to creating more robust and clinically relevant models.

This study aimed to develop an AI algorithm for pneumothorax detection by leveraging the data extraction capabilities of ChatGPT and employing strategic curriculum learning. The results were then validated on external multisite and multireader data to ensure the performance of the DL model for potential use in clinical practice.

2. Materials and Methods

2.1. Patient Selection and Classification

This retrospective study used data that were fully de-identified, anonymized and accessed under IRB CMUH113-REC2-049 with waived consent. For the development of the AI algorithm, a dataset comprising 6445 anonymized chest radiographs was utilized. These images were consecutively collected over a span of time from 2006 to 2018. Accompanying these radiographs, corresponding radiology reports were processed to extract clinically relevant information utilizing the advanced natural language processing capabilities of ChatGPT4, which was created by OpenAI. The primary task assigned to ChatGPT was to determine the presence of pneumothorax from the text of the radiology reports. Additionally, for cases where pneumothorax was identified, ChatGPT was further tasked with classifying the condition as either "large" or "small" based on descriptors found within the reports, as shown in Figure 1 below. This classification acts as a preliminary result to help differentiate simple and hard samples in an unsupervised manner to help structure the subsequent training phases of the AI model, aligning with the curriculum learning framework that progresses from more straightforward to more challenging cases.

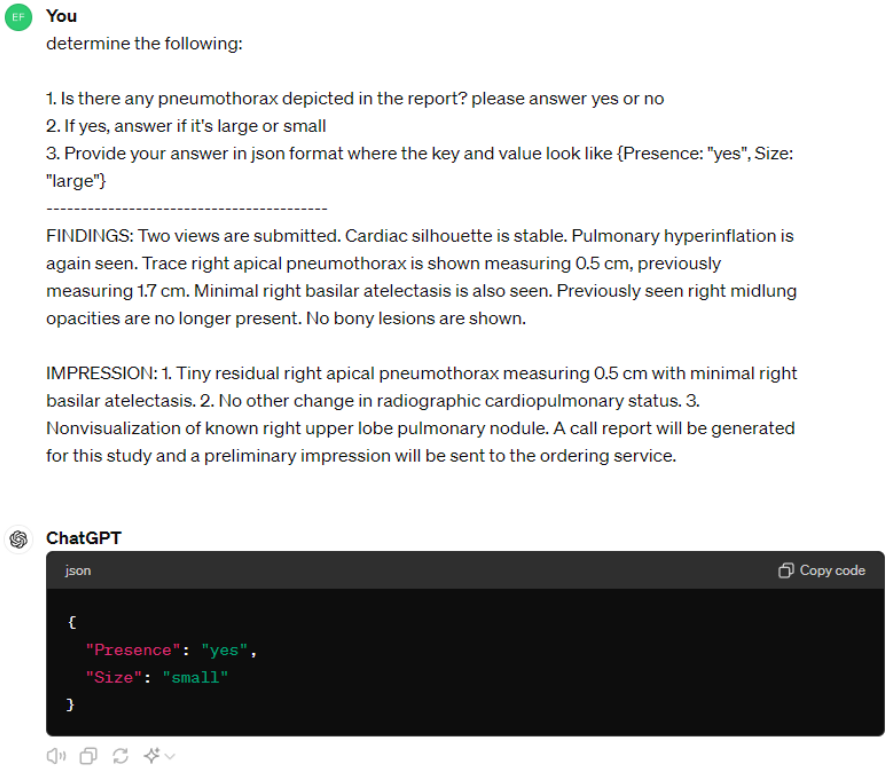

Figure 1. Example of obtaining pneumothorax information with ChatGPT.

2.2. Deep Learning Model Building

In this study, we utilized EfficientNetB3 as the backbone architecture for our deep learning model. EfficientNetB3 offers a balanced structure that enhances performance without significantly increasing computational demand, making it suitable for medical image analysis [28,29]. Its architecture, which scales uniformly with depth, width, and resolution, is ideal for detailed image features, such as those required for detecting varying sizes of pneumothorax [30,31]. The EfficientNetB3 model was optimized using the Adam optimizer, with an initial learning rate of 0.001, which was reduced by a factor of 10 whenever the validation loss plateaued for over 10 epochs. The model was trained using a binary cross-entropy loss function. To reduce overfitting and ensure generalization across diverse radiographic data, we implemented dropout at a rate of 0.5 and L2 regularization. Data augmentation techniques, including random rotations (up to 30 degrees) and horizontal flips, were applied to increase the diversity of training images. Training was conducted over 400 epochs with a batch size of 32, utilizing early stopping based on validation loss to prevent overtraining.

The training process was structured around a CL framework designed to adaptively introduce complexity, where the presence of pneumothorax and sizes were based solely on the results from ChatGPT. Initially, the model was trained exclusively on datasets labeled as "large" pneumothorax for cases where the condition was present; these are generally easier to identify due to their distinct visual characteristics. This initial phase was set up such that positive cases in each batch were composed of larger pneumothorax cases to establish a baseline proficiency in recognizing clearer and more distinguishable patterns. As the model's performance on these cases improved, we gradually introduced small pneumothoraces into the batches. The introduction was controlled and incremental, starting with a 5% inclusion of small pneumothorax cases; this percentage was increased by 5% each time the model achieved a predetermined performance threshold on the validation set. The progression from large to mixed batches was governed by an adaptive feedback mechanism, wherein the model's performance on periodically evaluated validation sets determined the timing and rate of transition. To align with the US Food and Drug Administration (FDA)

recommendations regarding approved device performances in developing a clinical-grade model, the transition criteria were designed such that the model was required to achieve and maintain an area under the receiver's operating characteristic curve (AUC) greater than 95%, as well as both sensitivity and specificity greater than 80%, before each transition of introducing additional small pneumothorax cases [32].

This curriculum-based approach is inspired by previous studies, where incremental learning has shown to improve the efficiency of model training processes by initially focusing on less complex examples [33–35]. In our implementation, this principle was applied by using a dynamic adjustment strategy, where the ratio of large to small pneumothorax cases in training batches was continually adapted based on the model's current learning state [36–38]. This approach ensures that the model is not overwhelmed by the complexity of smaller, subtler pneumothorax cases, thereby enhancing its ability to develop a comprehensive analytic capability over time. The overall CL framework design is shown in Figure 2.

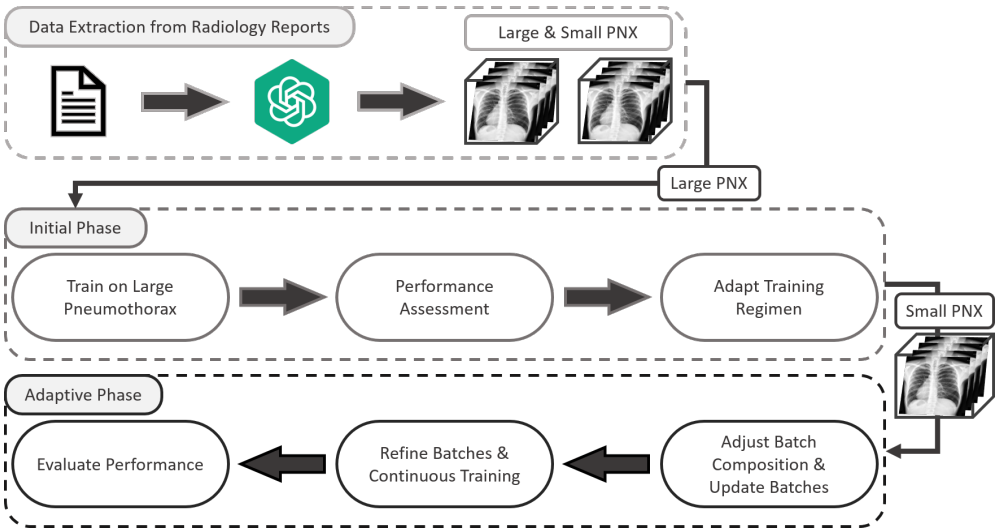

Figure 2. The curriculum training framework for pneumothorax (PNX) detection.

2.3. External Multisite Validation Data Collection

A total of 300 anonymized chest radiographs were consecutively collected between 2016 and 2021 from 3 clinical sites in the U.S. and from 1 clinical site in Taiwan (OUS). The dataset was generated from 9 different manufacturers of radiologic data sources, including Samsung Electronics, Shimadzu, Toshiba, GE Healthcare, Swissray, Philips Medical System, Konica Minolta, Canon Inc., Fujifilm Corporation, etc. The radiographs were collected following the inclusion and exclusion criteria, with stratification depending on whether the patient received thoracentesis, as shown in Figure 3.

In this study, all patients were required to be at least 18 years old, applicable to both males and females. For chest X-ray images, the inclusion criteria emphasized clarity and diagnostic value, requiring images to feature an unobscured chin, arms not covering the chest's lateral walls, symmetrically placed sternoclavicular joints relative to the spinal column, horizontal clavicles with no more than ten ribs visible above the diaphragm, and a diaphragm intersecting between the 5th and 7th anterior ribs. Additionally, clear visibility of ribs and the thoracic cage over the heart area, along with distinctly visible vascular lung markings, were mandatory. Any images that met the following criteria were excluded: obscured chins, arms obscuring the chest's lateral walls, asymmetrically placed sternoclavicular joints, non-horizontal clavicles, more than ten visible ribs above the diaphragm, inappropriate diaphragm intersection, or compromised visibility of ribs and thoracic cage. These criteria ensured that the X-rays used were of high quality and suitable for accurate analysis.

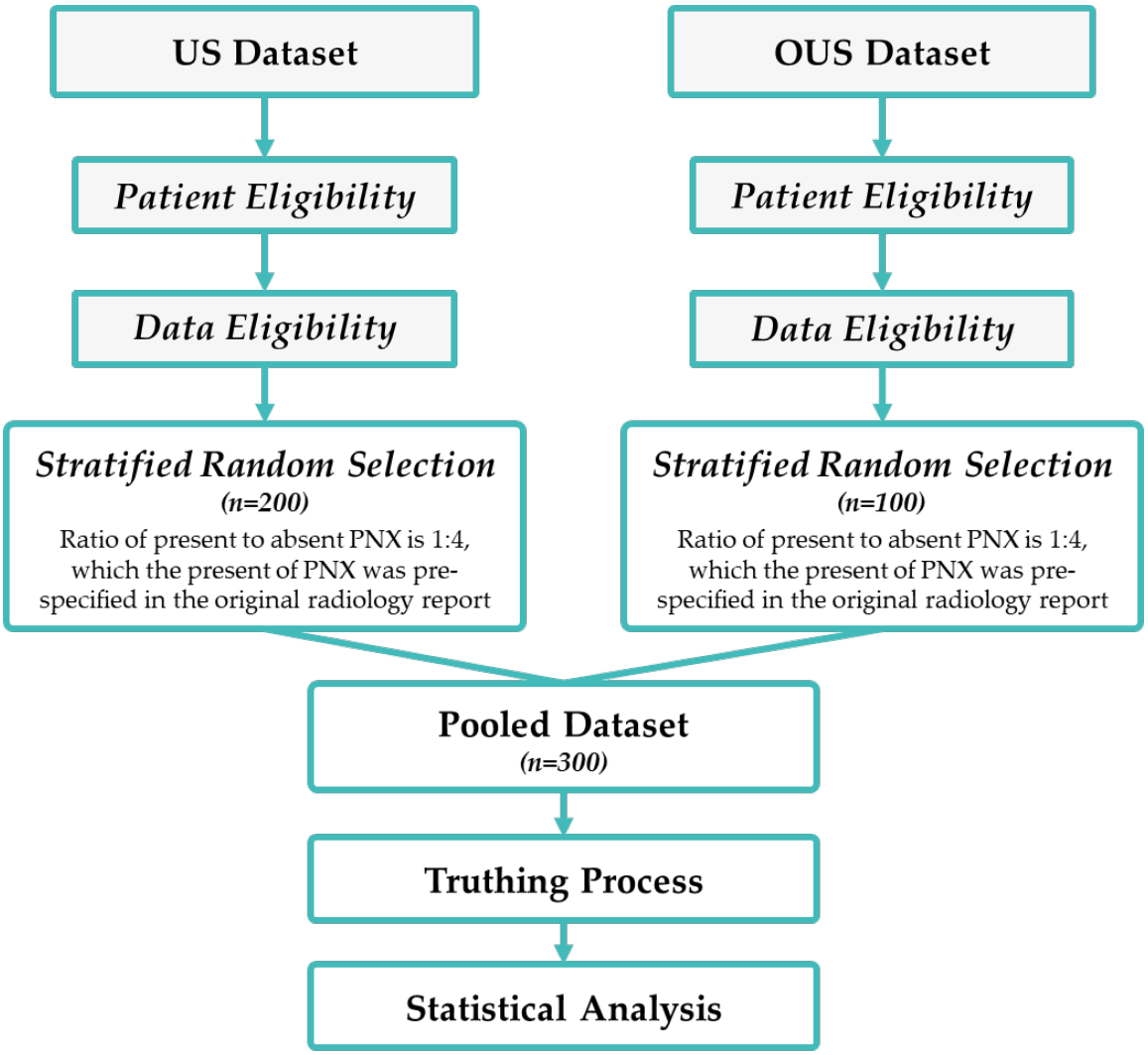

Figure 3. Flowchart of the validation workflow.

2.4. Ground Truth Definition

The ground truthing of this assessment study included three U.S. board-certified ex-pert radiologists reviewing the radiographs and assigning whether a pneumothorax was present in each image. These radiologists all had the U.S. American Board of Radiology (ABR) board certification in Diagnostic Radiology, with more than 10 years of experience in assessing chest X-rays and conducting a high volume (greater than 75 cases per week) of CXR assessments. For each case, each radiologist was asked to provide the following information: whether pneumothorax is present or absent; the size (large $\geq$ 3 cm and small $<$ 3 cm), which was measured from the thoracic cupola to the lung apex and defined by the ACCP guidelines; the location, including right upper lobe (RUL), right middle lobe (RML), right lower lobe (RLL), left upper lobe (LUL), and left lower lobe (LLL); and any additional comments the radiologist would like to provide about the case. The presence, size, and location of pneumothorax were determined based on the majority agreement of the three U.S. board-certified expert radiologists who reviewed the radiographs independently, and this information was further defined as the ground truth (GT).

2.5. Statistical Analysis

Our study utilized Python 3.6 and R 4.0.2 for data processing and statistical analyses. The chi-square test assessed the independence of categorical variables, examining whether significant associations existed between them. To validate the AI algorithm's sensitivity,

specificity, and AUC, One-Sample Z tests were employed, comparing the algorithm's results against the GT. We calculated 95% confidence intervals (95% C.I.) for sensitivity, specificity, and AUC to comprehensively evaluate the algorithm's ability to detect pneumothorax. DeLong's method was applied to compute the 95% CIs for the AUC, a technique specifically suited for correlated ROC curves. This method corrects for the correlation between AUC estimates, ensuring a rigorous assessment of diagnostic accuracy. For sensitivity and specificity, Wilson's score method was used to calculate 95% CIs, offering precise intervals for binomial proportions, which is especially valuable in scenarios with small sample sizes or extreme event rates. This study used an upper-tail test with a significance level (Type I error, α) of 2.5% to test our hypotheses. This framework supported the AI algorithm's generalizability across various subpopulations, including different genders (male, female), age groups (18–34, 35–49, 50–64, above 65 years), and data sources (U.S., Taiwan). We also examined performance variations among different manufacturers (Samsung Electronics (Suwon-si, Republic of Korea), Shimadzu (Kyoto, Japan), Toshiba (Tokyo, Japan), etc.), and analyzed sensitivity based on pneumothorax size (small, large) and anatomical location (RUL, RML, RLL, LUL, LLL). Potential confounders, such as image quality issues and unrelated radiologic findings, were also considered to determine their impact on the AI's performance in detecting pneumothorax.

3. Results

3.1. Patient Characteristics

During the training process, ChatGPT processed each radiograph's corresponding radiology report, and out of the 6445 chest radiographs, a total of 2554 radiographs were identified to contain pneumothoraces. Among these cases, ChatGPT classified 773 as "small" pneumothoraces and 1781 as "large" pneumothoraces. This classification enabled the structuring of the training phases of the AI model, aligning with the curriculum learning framework to ensure that the model was exposed to varying complexities in a more meaningful order.

To validate the model's generalizability, external multisite data of a total of 300 chest X-ray images that met the inclusion criteria were consecutively collected for the study. Among them, 162 (54.0%) were male and 138 (46.0%) were female. The mean (standard deviation, SD) age was 49.5 (16.3) years. The case distribution was as detailed in Table 1 based on the presence of pneumothorax or not.

Table 1. Basic characteristics for 300 external validation datasets.

	Pneumothorax Absent (*n* = 232)	**Pneumothorax Present (*n* = 68)**	***p*-Value**
Gender			0.0607
Female	114	24	
Male	118	44	
Age			<0.0001
18–34 y/o	37	25	
35–49 y/o	68	13	
50–64 y/o	99	15	
above 65 y/o	28	15	
Data Source			0.4072
U.S.	158	42	
Taiwan	74	26	
X-ray Manufacturer			<0.0001
Samsung Electronics	137	10	
Shimadzu	30	15	
Toshiba	44	11	
Others *	21	32	

* Other manufacturers include GE Healthcare (*n* = 16), Swissray (*n* = 14), Philips Medical System (*n* = 9), Konica Minolta (*n* = 9), Canon Inc. (*n* = 3), and Fujifilm Corporation (*n* = 2).

3.2. Evaluation of AI Performance

The performance metrics used were sensitivity, specificity, and the AUC. In order to assess whether the model performance has potential for future clinical use, we set the performance goals to match closely with the ones used in many US FDA-approved devices, where the sensitivity/specificity should reach over 0.8 and the AUC should be greater or equal to 0.95 [32]. As shown in Table 2, the One-Sample Z tests showed that the sensitivity and specificity both exceeded 0.8; the AUC also exceeded 0.95. The sensitivity of the AI algorithm was 0.97, with a 95% CI of [0.90, 0.99]; the specificity was 0.97, with a 95% CI of [0.94, 0.99]; and the AUC was 0.98, with a 95% DeLong's CI of [0.96, 1.00]. Overall, the agreement between the AI algorithm and the GT met the performance goal of exceeding 0.8 in both sensitivity and specificity and of exceeding 0.95 in AUC, compared with the GT.

Table 2. Performance of the AI algorithm.

Metrics	Performance	95% C.I.
Sensitivity	0.97	(0.90, 0.99)
Specificity	0.97	(0.94, 0.99)
AUC	0.98	(0.96, 1.00)

The ROC curve (Figure 4) offers a graphical representation of the AI algorithm's performance, illustrating how the AUC encompasses the entire two-dimensional space under the ROC curve at all classification thresholds, spanning from (0, 0) to (1, 1), with a value of 0.98. This high AUC value for the 300 evaluated cases underscores the algorithm's strong classification capability.

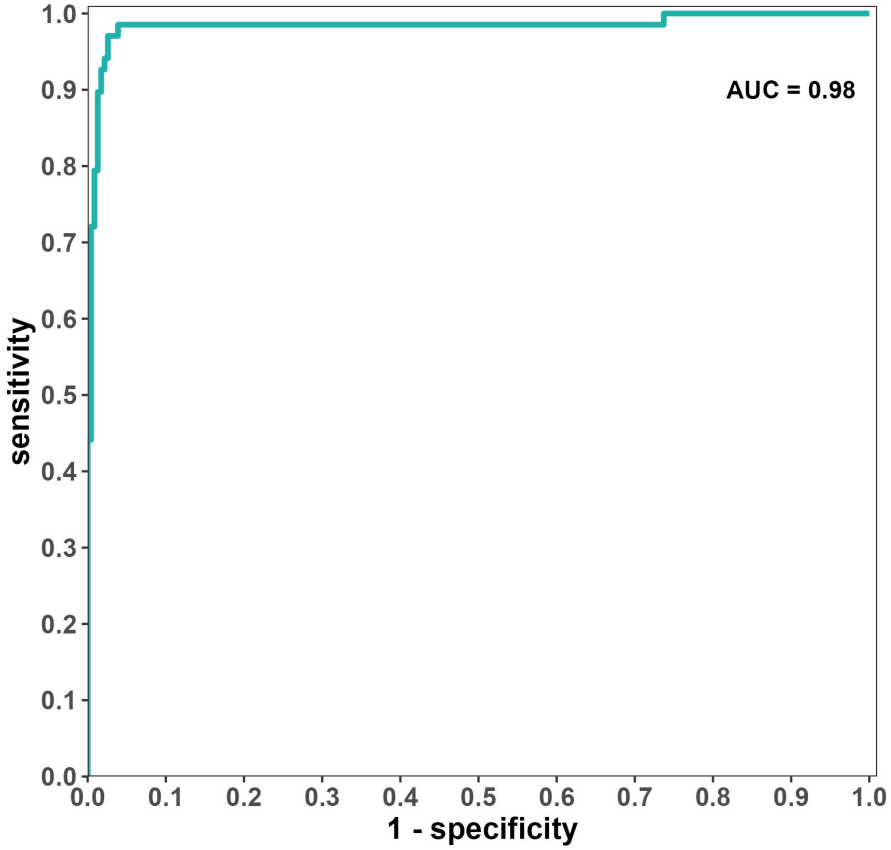

Figure 4. ROC curve of the AI algorithm.

3.3. Subgroup Analysis Results

We conducted subgroup analyses to evaluate the AI algorithm's performance across various subpopulations (Tables 3 and 4). This examination against the GT for specific groups ensures the AI system's reliability and effectiveness across a wide range of patient demographics.

Table 3. Performance of the AI algorithm in each subpopulation.

Characteristics	Sensitivity (Wilson's CI)	Specificity (Wilson's CI)	AUC (DeLong's CI)
Gender			
Female	1.00 (0.86, 1.00)	0.97 (0.93, 0.99)	0.99 (0.98, 1.00)
Male	0.95 (0.85, 0.99)	0.97 (0.92, 0.99)	0.98 (0.95, 1.00)
Age			
18–34 y/o	0.96 (0.80, 1.00)	0.97 (0.86, 1.00)	0.97 (0.92, 1.00)
35–49 y/o	1.00 (0.77, 1.00)	0.97 (0.90, 0.99)	1.00 (0.99, 1.00)
50–64 y/o	0.93 (0.70, 1.00)	0.97 (0.91, 0.99)	0.99 (0.98, 1.00)
above 65 y/o	1.00 (0.80, 1.00)	0.96 (0.82, 1.00)	0.98 (0.93, 1.00)
Data Source			
U.S.	0.95 (0.84, 0.99)	0.96 (0.92, 0.98)	0.97 (0.93, 1.00)
Taiwan	1.00 (0.87, 1.00)	0.99 (0.93, 1.00)	1.00 (0.99, 1.00)
Manufacturer			
Samsung Electronics	1.00 (0.72, 1.00)	0.99 (0.95, 1.00)	1.00 (1.00, 1.00)
Shimadzu	1.00 (0.80, 1.00)	1.00 (0.89, 1.00)	1.00 (1.00, 1.00)
Toshiba	1.00 (0.74, 1.00)	0.98 (0.88, 1.00)	1.00 (0.99, 1.00)
Others *	0.94 (0.80, 0.98)	0.91 (0.80, 0.92)	0.90 (0.81, 1.00)

* Other manufacturers including GE Healthcare, Swissray, Philips Medical System, Konica Minolta, Canon Inc and Fujifilm Corporation.

Table 4. Sensitivity of the AI algorithm for each pneumothorax size and location.

Characteristics	Sensitivity (Wilson's CI)
Pneumothorax Size	
Small	0.97 (0.87, 1.00)
Large	0.97 (0.83, 1.00)
Pneumothorax Location	
RUL	1.00 (0.90, 1.00)
RML	1.00 (0.78, 1.00)
RLL	1.00 (0.83, 1.00)
LUL	0.97 (0.83, 1.00)
LLL	0.96 (0.84, 1.00)

This study's results show that all subpopulations achieved a high performance in detecting pneumothorax, with sensitivity ranging from 0.93 to 1.00 and specificity between 0.91 and 1.00. These results highlight the AI algorithm's robustness in evaluating pneumothorax across various patient demographics and clinical conditions, effectively identifying different sizes of pneumothorax.

3.4. Ablation Study Results

The CL framework was designed to incrementally introduce images in a more gradual order, in this case, by starting with easier samples such as larger pneumothorax and adaptively moving into more complex subtle cases. The external multisite validation results, as shown in Figure 5, demonstrate that the CL framework achieved a superior performance compared to the traditional training method. Specifically, CL achieved a higher overall AUC of 0.98, whereas the traditional method, without the graduated complexity of curriculum learning, reached an AUC of 0.92 on the external validation data. Initially, models trained with only large pneumothorax exhibited a slower improvement in performance in the

early epochs as the model began with simpler, more obvious examples, allowing for the consolidation of foundational knowledge before progressively introducing more complex examples. In contrast, the traditional training method exposed the model to the full diversity of the dataset from the beginning, including both large and small pneumothorax cases, resulting in faster initial improvements. However, after the initial phase, the model with curriculum learning showed a sharper increase in performance as more complex data were introduced. This is indicative of the model's enhanced tuning to foundational patterns, enabling it to generalize more effectively from increasingly difficult data. Conversely, the traditional method, per curve, tended to plateau earlier as the model potentially began to overfit or exhaust the learning from the diverse training set.

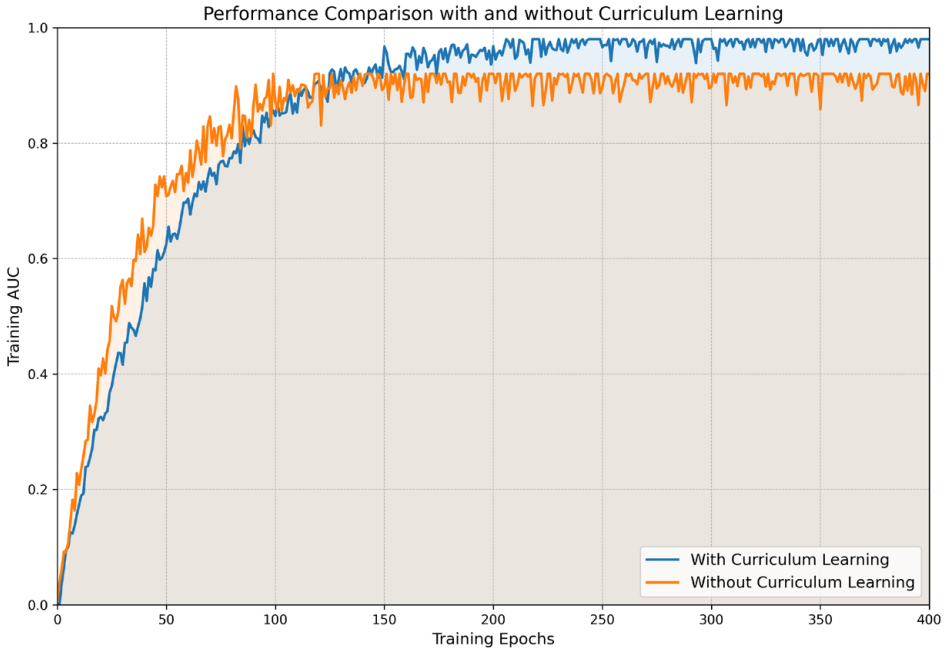

Figure 5. Performance comparison with and without curriculum learning.

3.5. Comparison of CADt Algorithms

As shown in Table 5, we conducted a comparative analysis of our algorithm versus existing US FDA-approved devices, specifically designed for pneumothorax detection. Key performance indicators included sensitivity, specificity, and the AUC. Our primary objectives were to achieve a minimum sensitivity and specificity of 0.8 and an AUC of 0.95. The results of our proposed algorithm achieved a sensitivity of 0.97, a specificity of 0.97, and an AUC of 0.98, thereby meeting and exceeding the benchmarks set by leading FDA-approved solutions for pneumothorax detection [32].

Table 5. Performance comparison against FDA 510(k)-approved CAD(t) AI algorithms [32].

Devices	Sensitivity	Specificity	AUROC
HealthPNX	0.93	0.92	0.98
Critical Care Suite	0.84	0.93	0.96
Red Dot	0.94	0.87	0.97
AIMI-Triage CXR	0.92	0.90	0.96
Ours	0.97	0.97	0.98

4. Discussion

This study demonstrates the utilization of ChatGPT with a curriculum learning framework, enhancing the potential of the development of a more robust pneumothorax detection model for potential clinical use. Our multisite validation demonstrated that leveraging

advanced natural language processing tools could help build a more clinically tailored CL framework to handle varied complexities, such as different sizes of pneumothoraces. The results reveal that the CL framework can outperform traditional training methods. Systematically introducing more subtle and clinically challenging cases into the training process allowed the model not only to achieve higher performance but also to ensure robustness and generalizability across diverse subgroups. This improvement is particularly notable in the detection of small pneumothoraces, which are often more challenging to diagnose due to their subtle radiographic features.

The incremental introduction of complexity in the training set proved crucial in enhancing the model's diagnostic capabilities. Starting with simpler, large pneumothorax cases allowed the model to establish a stronger foundation model. This strategic approach was followed by the introduction of smaller, more complex pneumothorax cases, which challenged the model's learning depth and adaptability. Such a structured approach prevented the early plateauing of model performance, a common issue in traditional training methods where the model is exposed to a heterogeneous mix of cases from the beginning. The model's performance in detecting both large and small pneumothoraces on external multisite and multireader ground truth results suggests its potential for real-world clinical application. In clinical practice, the ability to accurately identify small pneumothoraces is crucial, as these cases are more likely to be overlooked yet carry significant risks if undetected.

Previous studies have also shown promising results by integrating LLMs such as ChatGPT to help streamline the analysis of clinical reports. For example, Doshi et al. [39] demonstrated that ChatGPT and similar LLMs are adept at simplifying radiology reports, which can help enhance patient understanding. Similarly, a study by Tan et al. [40] demonstrated the effectiveness of LLMs in inferring cancer disease responses from radiology reports. The results of this study further confirm the potential of LLMs to process and analyze large clinical datasets with minimal human intervention and extend the use case and scope of curriculum learning to include more varied and complex cases.

This study has certain limitations and potential areas of future work. One area of enhancement would be the integration of a multimodal deep learning approach. Our study primarily leveraged text-based radiology reports; however, incorporating radiology images alongside reports could further enhance the model's performance by capturing more comprehensive and nuanced features. Another consideration is the diversity of the sample complexity. While this study considered pneumothorax size as a primary factor in defining sample complexity, it is also worth exploring other features beyond size that may complicate diagnosis. Features such as atypical presentations, overlapping pathologies, or subtle radiographic signs could be extracted from radiology reports to provide a more holistic training. Additionally, incorporating difficult negative cases, such as conditions that mimic pneumothorax into the training process, could also enhance the model's discriminative capabilities, reducing false positives. In terms of limitations, the model's performance was dependent on the quality and comprehensiveness of the data extracted by ChatGPT, indicating that the model's effectiveness was as good as the data it was trained on. Future studies could explore the integration of additional data sources and perhaps a more extensive range of text-based clinical documentation to further enhance the model's training. Additionally, ongoing refinements in the CL algorithm could provide more dynamic adjustments to training complexities based on real-time performance metrics.

Furthermore, the integration of curriculum learning with the capabilities of ChatGPT in extracting and curating clinical information offers a novel approach in training models with data that are not only more insightful but also more clinically relevant. This model's performance in the multisite study suggests that such an approach could be replicated and scaled to other tasks within the field.

5. Conclusions

In conclusion, in this study, we developed a novel approach to integrating ChatGPT with a CL framework by creating a pneumothorax detection model. By methodically enhancing the training process with the inclusion of clinical information extracted from radiology reports, the model was able to achieve an overall AUC of 0.98, thus surpassing traditional training methods. The multisite external validation confirms the model's effectiveness across different subgroups, demonstrating its robustness and its potential for future clinical use. Lastly, the findings suggest that a robust AI model such as this one could potentially enhance clinical practice by improving the detection of pneumothorax, particularly smaller instances, to reduce the risk associated with delayed or missed diagnosis.

Author Contributions: Conceptualization, J.C. and T.-H.W.; methodology, J.C., K.-J.L. and T.-H.W.; software, J.C. and K.-J.L.; validation, K.-J.L. and T.-H.W.; formal analysis, J.C. and K.-J.L.; data curation, J.C. and K.-J.L.; writing—original draft preparation, J.C.; writing—review and editing, J.C., K.-J.L., T.-H.W. and C.-M.C.; visualization, J.C. and K.-J.L.; supervision, T.-H.W. and C.-M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The protocol of this retrospective study was reviewed and approved by the Institutional Review Board of China Medical University Hospital (IRB CMUH113-REC2-049) (approval date 6 March 2024).

Informed Consent Statement: The requirement for informed consent was waived due to the retrospective nature of this study.

Data Availability Statement: The datasets analyzed in the current study are not publicly available due to ethical restrictions and the proprietary nature of the study.

Acknowledgments: This study was partially supported by China Medical University Hospital (CMUH) Grant DMR-111-076. We also thank EverFortune.AI Co., Ltd., Taichung, Taiwan, for their generous support and invaluable guidance in the realm of deep learning architectures, as well as for facilitating the hardware support for this study.

Conflicts of Interest: J.C, K.-J.L and T.W. are employees of EverFortune.AI Co., Ltd. The authors declare that despite this affiliation, the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Melton, L.J.; Hepper, N.G.G.; Offord, K.P. Incidence of spontaneous pneumothorax in Olmsted County, Minnesota: 1950 to 1974. *Am. Rev. Respir. Dis.* **1979**, *120*, 1379–1382. [PubMed]
2. Sadikot, R.T.; Greene, T.; Meadows, K.; Arnold, A.G. Recurrence of primary spontaneous pneumothorax. *Thorax* **1997**, *52*, 805–809. [CrossRef] [PubMed]
3. Baumann, M.H.; Strange, C.; Heffner, J.E.; Light, R.; Kirby, T.J.; Klein, J.; Luketich, J.D.; Panacek, E.A.; Sahn, S.A. Management of spontaneous pneumothorax: An American College of Chest Physicians Delphi consensus statement. *Chest* **2001**, *119*, 590–602. [CrossRef] [PubMed]
4. Odaka, M.; Akiba, T.; Mori, S.; Asano, H.; Yamashita, M.; Kamiya, N.; Morikawa, T. Thoracoscopic surgery for refractory cases of secondary spontaneous pneumothorax. *Asian J. Endosc. Surg.* **2013**, *6*, 104–109. [CrossRef]
5. Jeon, H.; Kim, Y.D.; Choi, S.; Park, J. When is the optimal timing of the surgical treatment for secondary spontaneous pneumothorax? *Thorac. Cardiovasc. Surg.* **2015**, *65*, 050–055. [CrossRef] [PubMed]
6. Brar, M.S.; Bains, I.; Brunet, G.; Nicolaou, S.; Ball, C.G.; Kirkpatrick, A.W. Occult Pneumothoraces Truly Occult or Simply Missed: Redux. *J. Trauma Acute Care Surg.* **2010**, *69*, 1335–1337. [CrossRef] [PubMed]
7. Lichtenstein, D.; Mezière, G.; Lascols, N.; Biderman, P.; Courret, J.P.; Gepner, A.; Goldstein, I.; Tenoudji-Cohen, M. Ultrasound Diagnosis of Occult Pneumothorax. *Crit. Care Med.* **2005**, *33*, 1231–1238. [CrossRef] [PubMed]
8. Matsumoto, S.; Kishikawa, M.; Hayakawa, K.; Narumi, A.; Matsunami, K.; Kitano, M. A Method to Detect Occult Pneumothorax with Chest Radiography. *Ann. Emerg. Med.* **2011**, *57*, 378–381. [CrossRef] [PubMed]
9. Quekel, L.G.; Kessels, A.G.; Goei, R.; van Engelshoven, J.M. Miss Rate of Lung Cancer on the Chest Radiograph in Clinical Practice. *Chest* **1999**, *115*, 720–724. [CrossRef]

10. Matsumoto, S.; Sekine, K.; Funabiki, T.; Orita, T.; Shimizu, M.; Hayashida, K.; Kazamaki, T.; Suzuki, T.; Kishikawa, M.; Yamazaki, M.; et al. Diagnostic Accuracy of Oblique Chest Radiograph for Occult Pneumothorax: Comparison with Ultrasonography. *World J. Emerg. Surg.* **2016**, *11*, 1. [CrossRef]
11. Thian, Y.; Ng, D.; Hallinan, J.; Jagmohan, P.; Sia, S.Y.; Tan, C.; Ting, Y.; Kei, P.L.; Pulickal, G.; Tiong, V.T.Y.; et al. Deep Learning Systems for Pneumothorax Detection on Chest Radiographs: A Multicenter External Validation Study. *Radiol. Artif. Intell.* **2021**, *3*, e200190. [CrossRef] [PubMed]
12. Hallinan, J.; Feng, M.; Ng, D.; Sia, S.Y.; Tiong, V.T.Y.; Jagmohan, P.; Makmur, A.; Thian, Y. Detection of Pneumothorax with Deep Learning Models: Learning From Radiologist Labels vs Natural Language Processing Model Generated Labels. *Acad. Radiol.* **2021**, *29*, 1350–1358. [CrossRef]
13. Rajpurkar, P.; Irvin, J.; Ball, R.L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; et al. Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists. *PLoS Med.* **2018**, *15*, e1002686. [CrossRef]
14. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
15. Kim, M.; Yun, J.; Cho, Y.; Shin, K.; Jang, R.; Bae, H.; Kim, N. Deep Learning in Medical Imaging. *Neurospine* **2019**, *16*, 657–668. [CrossRef]
16. Lee, J.-G.; Jun, S.; Cho, Y.; Lee, H.; Kim, G.; Seo, J.; Kim, N. Deep Learning in Medical Imaging: General Overview. *Korean J. Radiol.* **2017**, *18*, 570–584. [CrossRef] [PubMed]
17. Tsang, R. Practical Applications of ChatGPT in Undergraduate Medical Education. *J. Med. Educ. Curric. Dev.* **2023**, *10*. [CrossRef]
18. Liu, J.; Wang, C.; Liu, S. Utility of ChatGPT in Clinical Practice. *J. Med. Internet Res.* **2023**, *25*, e48568. [CrossRef]
19. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in Medicine: An Overview of Its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595. [CrossRef] [PubMed]
20. Zhou, Z. Evaluation of ChatGPT's Capabilities in Medical Report Generation. *Cureus* **2023**, *15*, e37589. [CrossRef]
21. Alhaidry, H.; Fatani, B.; Alrayes, J.O.; Almana, A.M.; Alfhaed, N.K. ChatGPT in Dentistry: A Comprehensive Review. *Cureus* **2023**, *15*, e38317. [CrossRef]
22. Nikolic, M.; Lok, L.; Mattishent, K.; Barth, S.; Yung, B.; Cummings, N.; Shulgina, L.; Wade, D.; Shittu, M.; Vali, Y.; et al. Multi-centre prospective comparison of the BTS and ACCP guidelines to determine size in primary spontaneous pneumothorax. *Thorax* **2013**, *68*, A44. [CrossRef]
23. Kelly, A.M.; Druda, D. Comparison of Size Classification of Primary Spontaneous Pneumothorax by Three International Guidelines: A Case for International Consensus? *Respir. Med.* **2008**, *102*, 1830–1832. [CrossRef] [PubMed]
24. Costa, E.; Chaves, C.N.; Gomes, S.; Beraldo, M.A.; Volpe, M.; Tucci, M.; Schettino, I.A.; Bohm, S.H.; Carvalho, C.R.; Tanaka, H.; et al. Real-time Detection of Pneumothorax Using Electrical Impedance Tomography. *Crit. Care Med.* **2008**, *36*, 1230–1238. [CrossRef] [PubMed]
25. Ghebrechristos, H.; Alaghband, G. Deep Curriculum Learning Optimization. *SN Comput. Sci.* **2020**, *1*, 245. [CrossRef]
26. Wang, Y.; Yue, Y.; Lu, R.; Liu, T.-D.; Zhong, Z.; Song, S.; Huang, G. EfficientTrain: Exploring Generalized Curriculum Learning for Training Visual Backbones. *arXiv* **2022**, arXiv:2211.09703.
27. Simionescu, C. Deep Learning Training Procedure Augmentations. *arXiv* **2022**, arXiv:2211.14395.
28. Dudeja, T.; Dubey, S.; Bhatt, A. Ensembled EfficientNetB3 Architecture for Multi-class Classification of Tumours in MRI Images. *Intell. Decis. Technol.* **2023**, *17*, 395–414. [CrossRef]
29. Sheela, R.K.; Nagaraju, Y.; Sahu, D.A. Histopathological Image Classification of Breast Cancer Using EfficientNet. In Proceedings of the 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 27–29 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–8. [CrossRef]
30. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946.
31. Tadepalli, Y.; Kollati, M.; Kuraparthi, S.; Kora, P. EfficientNet-B0 Based Monocular Dense-Depth Map Estimation. *Trait. Du Signal* **2021**, *38*, 1485. [CrossRef]
32. Yuba, M.; Iwasaki, K. Systematic Analysis of the Test Design and Performance of AI/ML-Based Medical Devices Approved for Triage/Detection/Diagnosis in the USA and Japan. *Sci. Rep.* **2022**, *12*, 16874. [CrossRef]
33. Hua, Y.; Li, Y.F.; Qi, G.; Wu, W.; Zhang, J.; Qi, D. Less Is More: Data-Efficient Complex Question Answering Over Knowledge Bases. *J. Web Semant.* **2020**, *65*, 100612. [CrossRef]
34. Guo, S.; Huang, W.; Zhang, H.; Zhuang, C.; Dong, D.; Scott, M.R.; Huang, D. CurriculumNet: Weakly Supervised Learning from Large-Scale Web Images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 135–150. [CrossRef]
35. Abbe, E.; Cornacchia, E.; Lotfi, A. Provable Advantage of Curriculum Learning on Parity Targets with Mixed Inputs. *Adv. Neural Inf. Process. Syst.* **2024**, *36*.
36. Wang, Y.; Gan, W.; Yang, J.; Wu, W.; Yan, J. Dynamic Curriculum Learning for Imbalanced Data Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2019, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5017–5026. Available online: https://arxiv.org/abs/1901.06783 (accessed on 6 March 2024).
37. Gong, Y.; Liu, C.; Yuan, J.; Yang, F.; Cai, X.; Wan, G.; Chen, J.; Niu, R.; Wang, H. Density-Based Dynamic Curriculum Learning for Intent Detection. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Gold Coast, Australia, 1–5 November 2021; pp. 3034–3037. [CrossRef]

38. Lin, H.; Zhang, H.; Ma, Y.; He, T.; Zhang, Z.; Zha, S.; Li, M. Dynamic Mini-Batch SGD for Elastic Distributed Training: Learning in the Limbo of Resources. *arXiv* **2019**, arXiv:1904.12043.
39. Doshi, R.; Amin, K.; Khosla, P.; Bajaj, S.; Chheang, S.; Forman, H.P. Utilizing Large Language Models to Simplify Radiology Reports: A Comparative Analysis of ChatGPT-3.5, ChatGPT-4.0, Google Bard, and Microsoft Bing. *medRxiv* **2023**. [CrossRef]
40. Tan, R.S.; Lin, Q.; Low, G.H.; Lin, R.; Goh, T.C.; Chang, C.C.; Lee, F.F.; Chan, W.Y.; Tan, W.C.; Tey, H.J.; et al. Inferring Cancer Disease Response from Radiology Reports Using Large Language Models with Data Augmentation and Prompting. *J. Am. Med. Inform. Assoc.* **2023**, *30*, 1657–1664. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.