

OPPONENT'S EVALUATION OF THE MASTER'S THESIS

Student: Mohiuddin Abdul Kader

Opponent: Assoc. Prof. Martin Kotyrba,
Ph.D.

Study program: **Information Technologies**
Specialization: **Software Engineering**
Academic year: **2024/2025**

Master's Thesis topic: **AI for Medical Diagnostics**

The thesis demonstrates a thorough elaboration of the application of deep learning models for automated medical diagnosis using chest radiographs. The work focuses on multi-label classification of 14 thoracic diseases using the NIH Chest X-ray14 dataset, a significant benchmark in the field. The theoretical part provides a comprehensive overview of traditional AI approaches in diagnosis, the rise of machine learning and deep learning in healthcare, explainability (XAI), generative models, and existing deep learning techniques in medical image analysis, including specific applications in classification, segmentation, and detection.

The validity of the approach is supported by the use of established state-of-the-art convolutional neural network (CNN) architectures (ResNet-50, DenseNet-121, EfficientNet-B0), standard evaluation metrics suitable for multi-label classification (Accuracy, Precision, Recall, F1 Score, AUC), and recognized techniques like transfer learning, data augmentation, and consideration of ensembling and threshold tuning. The methodology aligns with common practices in deep learning for medical imaging. The development of an interactive Colab interface also demonstrates a practical aspect of validation.

- The thesis explicitly identifies and attempts to resolve several difficulties, particularly those encountered when applying deep learning to real-world medical imaging data:
- Challenges in AI-Based Diagnostics: Generalization and real-world deployment. Data quality and bias. Regulatory and ethical concerns. Interpretability and trust in "black-box" models.
- Specific Dataset Challenges: The large scale but complex structure of the NIH Chest X-ray14 dataset. Data imbalance and annotation issues, where some diseases are much rarer than others, leading to biased predictions. Label noise derived from NLP parsing of radiology reports.
- Model Performance Issues: Consistently low recall (sensitivity) for rare or subtle conditions such as Pneumonia, Fibrosis, and Hernia, despite strong overall classification capabilities. Limitations due to thresholding at a fixed value. The finding that simply increasing image resolution did not universally improve performance.

Practical Implementation: Development of a Google Colab interface for real-time inference demonstrates practical application. The level of concepts is appropriate for a Master's thesis in this field. It involves understanding and applying state-of-the-art deep learning models and best practices in medical image analysis, addressing challenges specific to real-world datasets like class imbalance and the need for interpretability.

The level of the thesis topic is high and highly relevant to current research in AI for healthcare. Automated medical diagnostics using deep learning on chest X-rays is a prominent and impactful application area. The chosen dataset, NIH Chest X-ray14, is a widely recognized and challenging benchmark.

Errors and mistakes:

Language and Formatting: There are some minor grammatical errors and awkward phrasing (e.g., section titles like "For 128x128 Pixelof images RESNET"), more mistakes in contents. Capitalization is inconsistent in some headings. Descriptions for Figure 6 and Figure 7 seem swapped, and Figure 5 is referenced in text before Figure 4 but listed after it. The List of Figures has an incorrect order for Figures 4, 6, and 5 compared to their appearance in the text according me and figure 5 is missing in the list of figures.

Validation/Test Set: The experimental setup states that due to constraints, the validation and test sets were combined into a single validation set. This is a methodological limitation as the final performance is reported on a set used for monitoring training progress and early stopping, not a truly unseen test set. This limits the confidence in the model's generalization ability to completely new data. **Thresholding:** The results are primarily reported based on a fixed threshold of 0.5, despite the discussion highlighting that the disparity between high AUC and low recall suggests the need for per-class threshold optimization. While threshold tuning is mentioned as a recommendation, reporting performance after tuning would have provided a better indication of the models' potential clinical utility. **Data Augmentation:** Mentions planned Gaussian noise augmentation but states the final implementation only relied on contrast and brightness jittering. While not necessarily an error, this could be a point for clarification. **Citation Inconsistencies:** Some in-text citations, particularly early on, use numbers that seem inconsistent with the primary citation style and bibliography format, potentially remnants of different formatting or referencing styles.

5. Questions relating to the defence of the thesis.

1. Given the consistently low recall for Pneumonia and Consolidation across all tested models, what specific reasons (e.g., visual characteristics of these diseases on X-rays, severity of class imbalance, label noise) do you believe contributed most significantly to this limitation, and how would you prioritize future efforts to improve detection for these conditions?
2. You discussed the potential benefits of per-class threshold calibration to address the discrepancy between high AUC and low recall. Can you show or discuss the expected improvement in recall for key diseases (like Pneumonia or Fibrosis) if thresholds were optimized on the validation set, even if precision might decrease?
3. The experimental setup mentions combining the validation and test sets due to constraints. How might the reported performance metrics, particularly concerning the model's ability to generalize to completely unseen data, be different if a separate, independent test set had been used? What steps could be taken in future work to ensure more robust external validation?

•
The presentation of results using a default threshold, despite acknowledging its limitations and potential for optimization, doesn't fully showcase the potential of the models. There are minor language and formatting inconsistencies. Despite the experimental limitation regarding the test set, the depth of the technical work, the relevance of the problem, the systematic analysis of model performance and its limitations, and the clear articulation of future research directions make this a solid and valuable Master's thesis. It demonstrates the student's ability to apply advanced AI techniques to a complex, real-world medical problem, critically evaluate the results, and identify paths forward.

Evaluation of the thesis:

Overall evaluation of the thesis:

The Opponent shall grant a mark according to the ECTS classification scale:

A – Excellent, B – Very Good, C – Good, D – Satisfactory, E – Sufficient, F – Insufficient

An “F” grade also means "I do not recommend the thesis for defence."

I recommend this thesis to be defended and suggest the following evaluation:

B - Very Good

In the case of an evaluation grade of “F – Insufficient”, please supply the main shortages and reasons for this assessment.

Date: 04.06.2025

Thesis Opponent's Signature: