



# *Automatic assignment of diagnosis codes to free-form text medical notes*

Stefan Strydom

Supervisor: Prof Brink van der Merwe

23 October 2020



UNIVERSITEIT  
STELLENBOSCH  
UNIVERSITY



# Presentation outline

---



## Background

Electronic Health Records  
Clinical coding



## Dataset used in this research



## Problem statement



## Evaluation metrics



## Results

# Electronic Health Records (EHRs)

- *“Real-time, patient-centered record that makes information available instantly and securely to authorized users”* The United States Office of the National Coordinator for Health Information Technology (ONC)
- Ensures the availability of comprehensive patient information at the point of care, including patient demographics, medical history, family medical history, diagnoses, treatments, test results and medications prescribed
- Significant amount of information stored as free-form text notes



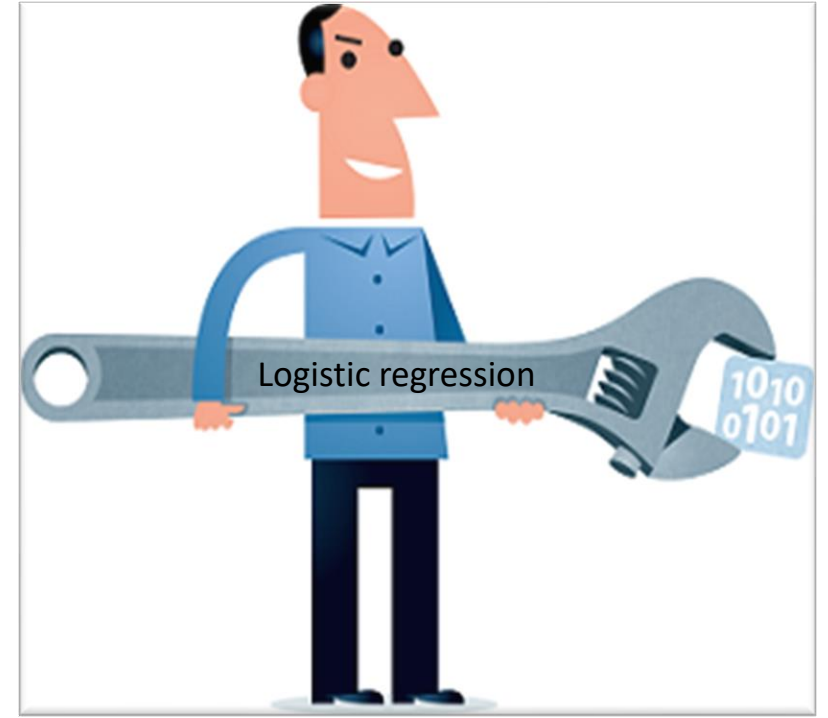




Large volumes of data generated...



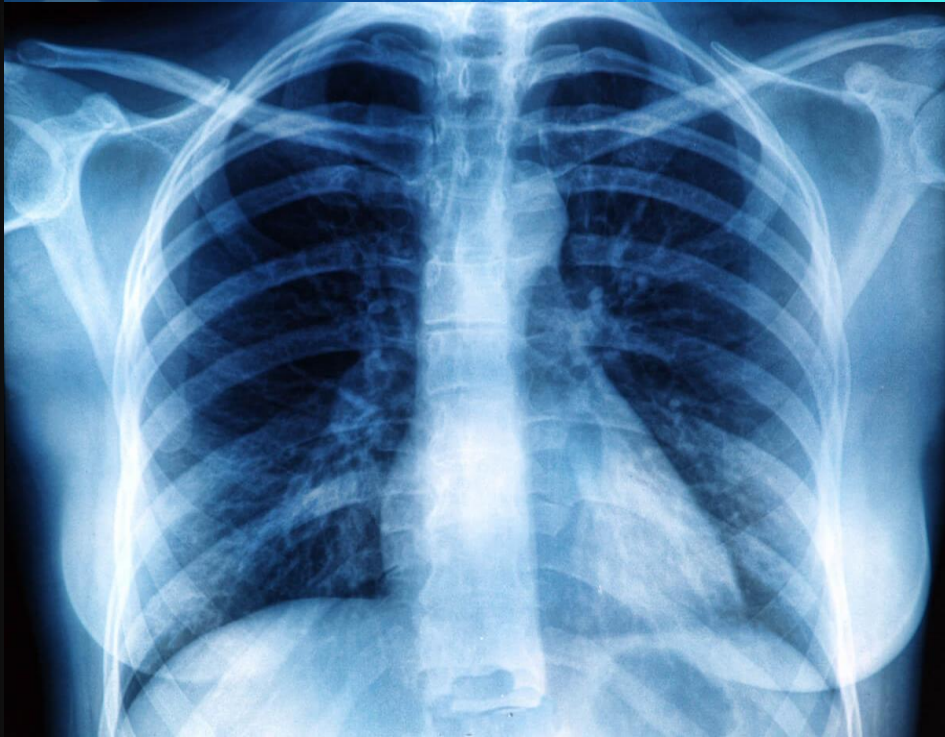
... but lack of standardisation  
and interoperability...



... and a healthcare analytics  
toolbox not fit for the problem

# EHRs and ML

---

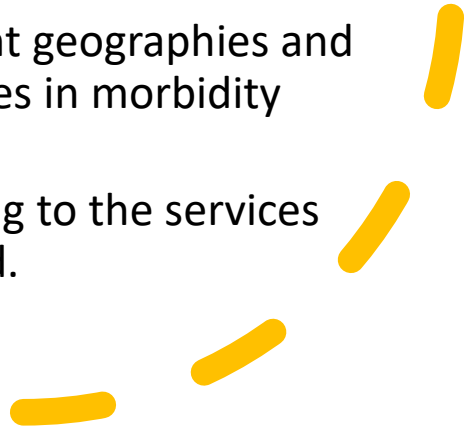


# EHRs and ML

---

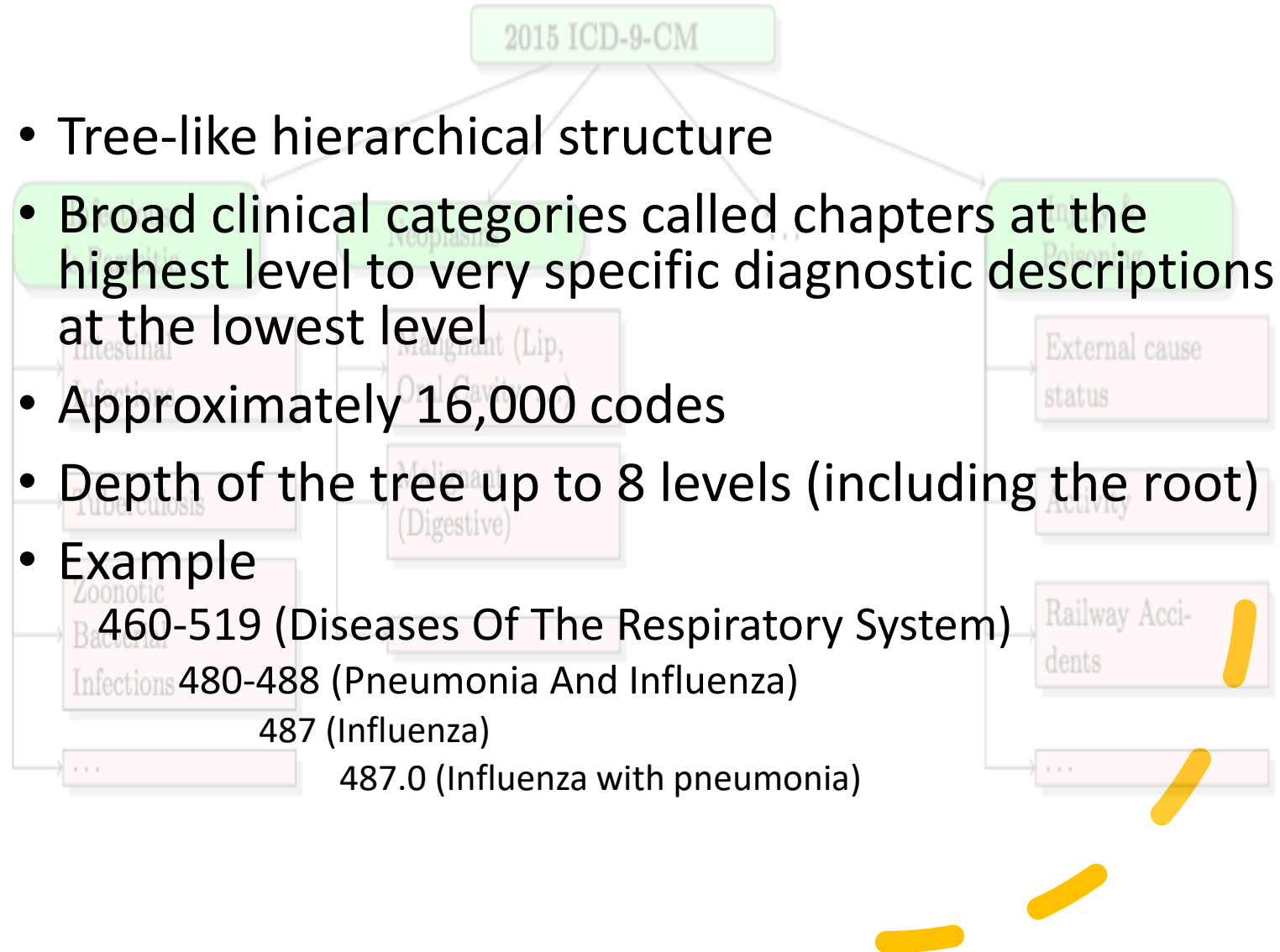
- The application of end-to-end deep learning approaches is starting to change this in NLP and computer vision
  - Recent studies have shown that models learning from longitudinal patient data recorded in EHRs outperform traditional approaches such as SCORE, PROCAM and Framingham to predict cardiovascular events by significant margins<sup>1</sup>
  - Machine learning has also been used to predict which patients are most likely to develop complications during or after hospital stays<sup>2,3</sup>
  - Machine learning applied to EHRs have also been successfully applied to case-finding<sup>4</sup>, to predict patient outcomes<sup>5</sup> and to predict mortality risk<sup>6</sup>
  - **My research focuses on assigning diagnosis codes to free-form text clinical notes**
-

# International Classification of Diseases (ICD)

- Disease classification system maintained by the WHO
  - ICD-10 most widely used globally, including South Africa
  - According to WHO, 70% of the world's total healthcare expenditures are allocated using the ICD system, either through reimbursement or budget allocation
  - Downstream uses of ICD-coded data include:
    - Describing the morbidity levels in different populations for planning and equitable distribution of healthcare resources according to the needs of populations
    - Comparing morbidity levels between different geographies
    - Comparing morbidity levels and case complexity between different hospitals and other healthcare facilities and practitioners
    - Comparing health outcomes between different geographies and different hospitals after allowing for differences in morbidity ("case-mix adjustment")
    - Reimbursing healthcare practitioners according to the services provided or the complexity of patients treated.
- 

# ICD-9 Structure

- Tree-like hierarchical structure
- Broad clinical categories called chapters at the highest level to very specific diagnostic descriptions at the lowest level
- Approximately 16,000 codes
- Depth of the tree up to 8 levels (including the root)
- Example
  - 460-519 (Diseases Of The Respiratory System)
  - 480-488 (Pneumonia And Influenza)
  - 487 (Influenza)
  - 487.0 (Influenza with pneumonia)





# MIMIC dataset

- Database of de-identified medical records from intensive care unit (ICU) stays at the Beth Israel Medical Centre<sup>7</sup>
- Versions II and III available; this talk focuses on MIMIC II
- 22,815 records with non-empty free-form text clinical notes
- Median note length = 1,322 tokens (10<sup>th</sup> percentile = 430; 90<sup>th</sup> percentile = 2,279)
- Each note is associated with one or more ICD-9 codes (mean = 9.45 codes per document)
- 5,031 ICD-9 codes present
- If the assigned codes are augmented to include their ancestors in the ICD-9 tree, then 7,042 codes are present



# Problem statement and features



## Problem statement:

*To automatically assign ICD-9 codes to free-form clinical text notes*



## Problem features:

Multi-label text classification problem

The input documents are long

Consists of non-standard (medical domain-specific) terminology and abbreviations

Relatively small dataset

High-dimensional, hierarchical imbalanced label space



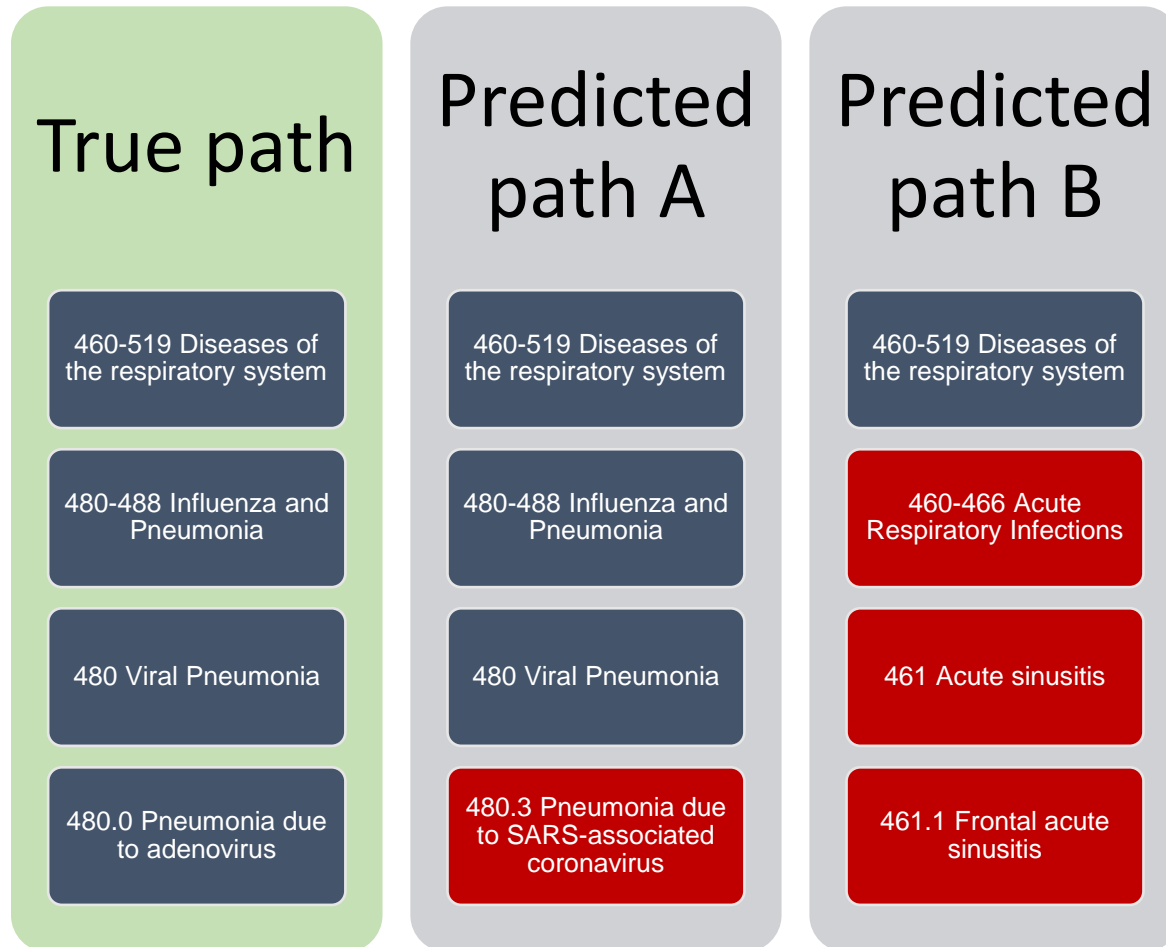
# Related work

---

- Several papers exist, but comparing results are difficult
- Deep-learning based papers employ different word embedding approaches and sizes, pre-processing techniques, etc., but improvement to the state-of-the-art are always ascribed to an architectural tweak
- Quality issues with some papers – test set peaking surprisingly common
- Hierarchical nature of the label space is not considered in model evaluation
- First aim of my research is to implement existing approaches systematically to determine which features are most important for this type of problem

# Hierarchical classification metrics

## Motivating example



## Flat evaluation metrics

Set of ground truths  $T = \{480.0\}$

Set of predicted codes  $\hat{P}_1 = \{480.3\}$

Set of predicted codes  $\hat{P}_2 = \{461.1\}$

$$recall = \frac{\hat{P} \cap T}{T}$$

$$recall_1 = \frac{0}{1} = 0$$

$$recall_2 = \frac{0}{1} = 0$$

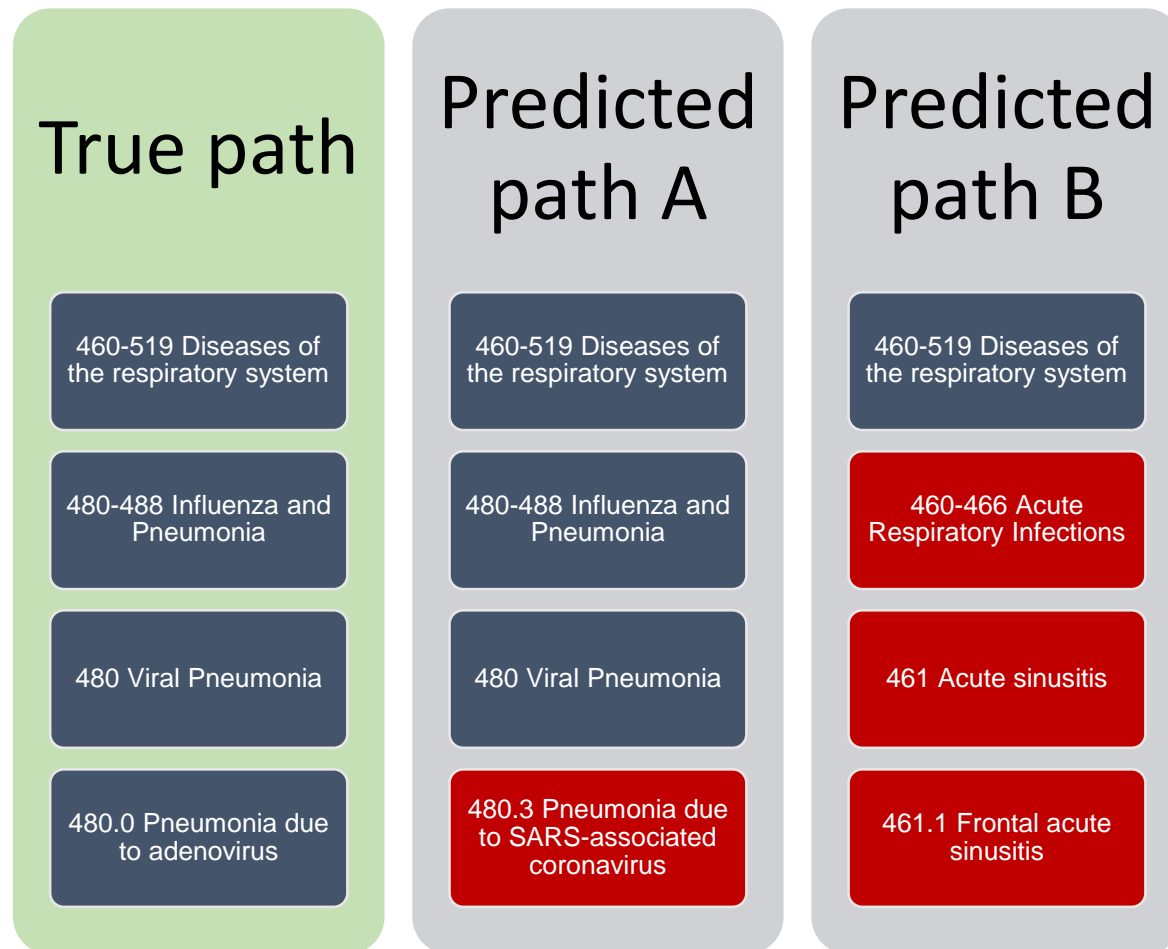
$$precision = \frac{\hat{P} \cap T}{\hat{P}}$$

$$precision_1 = \frac{0}{1} = 0$$

$$precision_2 = \frac{0}{1} = 0$$

# Hierarchical classification metrics

## Motivating example



## Hierarchical evaluation metrics

Now let  $hT$  and  $\widehat{hP}$  include the ancestors of the true positive and predicted codes respectively<sup>8</sup>:

$$hT = \{460-519, 480-488, 480, 480.0\}$$

$$\widehat{hP}_1 = \{460-519, 480-488, 480, 480.3\}$$

$$\widehat{hP}_2 = \{460-519, 460-466, 461, 461.1\}$$

$$recall = \frac{\widehat{hP} \cap hT}{T}$$

$$recall_1 = \frac{3}{4} = 0.75$$

$$recall_2 = \frac{1}{4} = 0.25$$

$$precision = \frac{\widehat{hP} \cap T}{P}$$

$$precision_1 = \frac{3}{4} = 0.75$$

$$precision_2 = \frac{1}{4} = 0.25$$



# A few interesting research findings



CNNs outperform RNNs



Attention mechanism provides performance boost to RNNs and CNNs



Gradient boosting machines, trained on bag-of-words features and binary relevance method, outperform baseline neural nets



Significant boost when using pretrained word vectors for RNNs and CNNs when the models are large



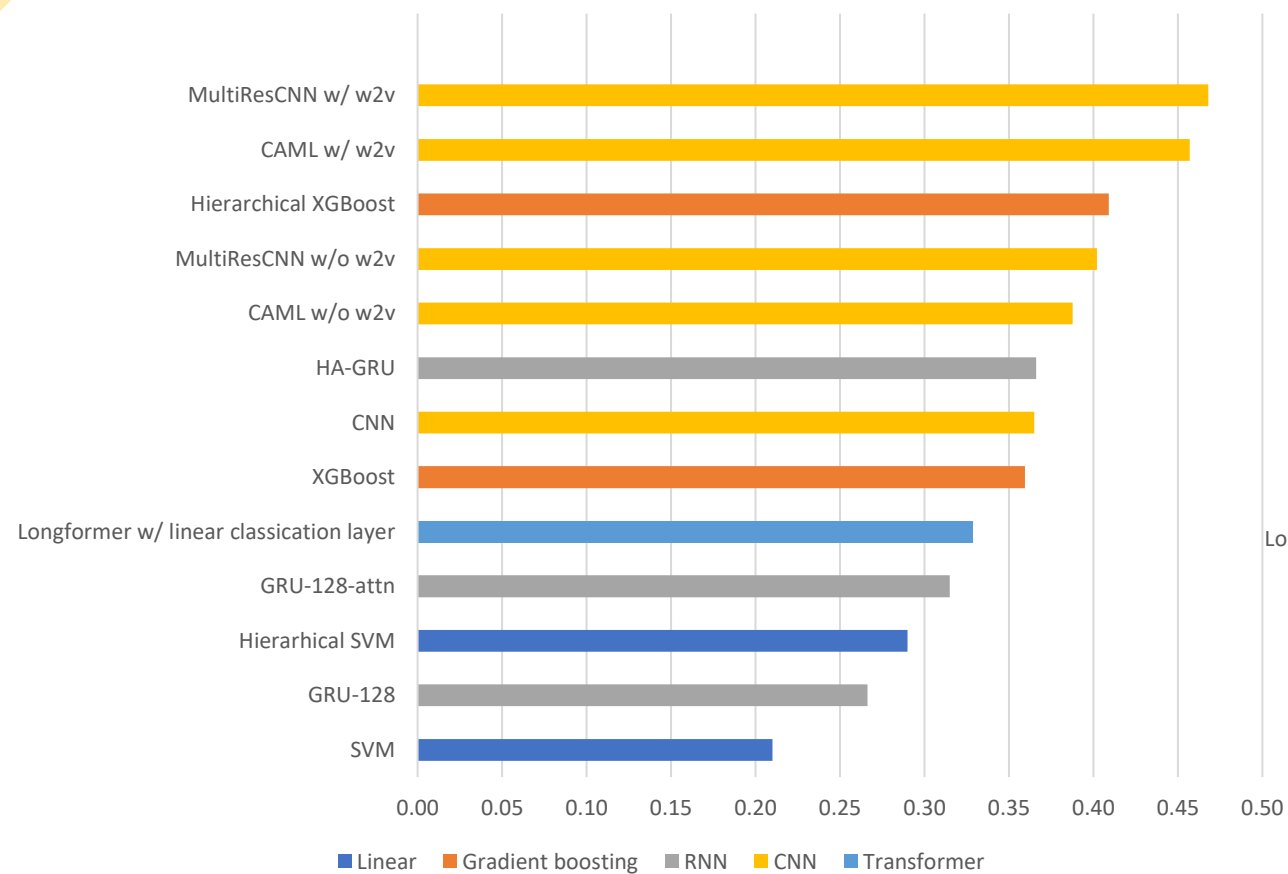
Neural nets doesn't automatically learn the relationship between ancestors and children in the ICD tree



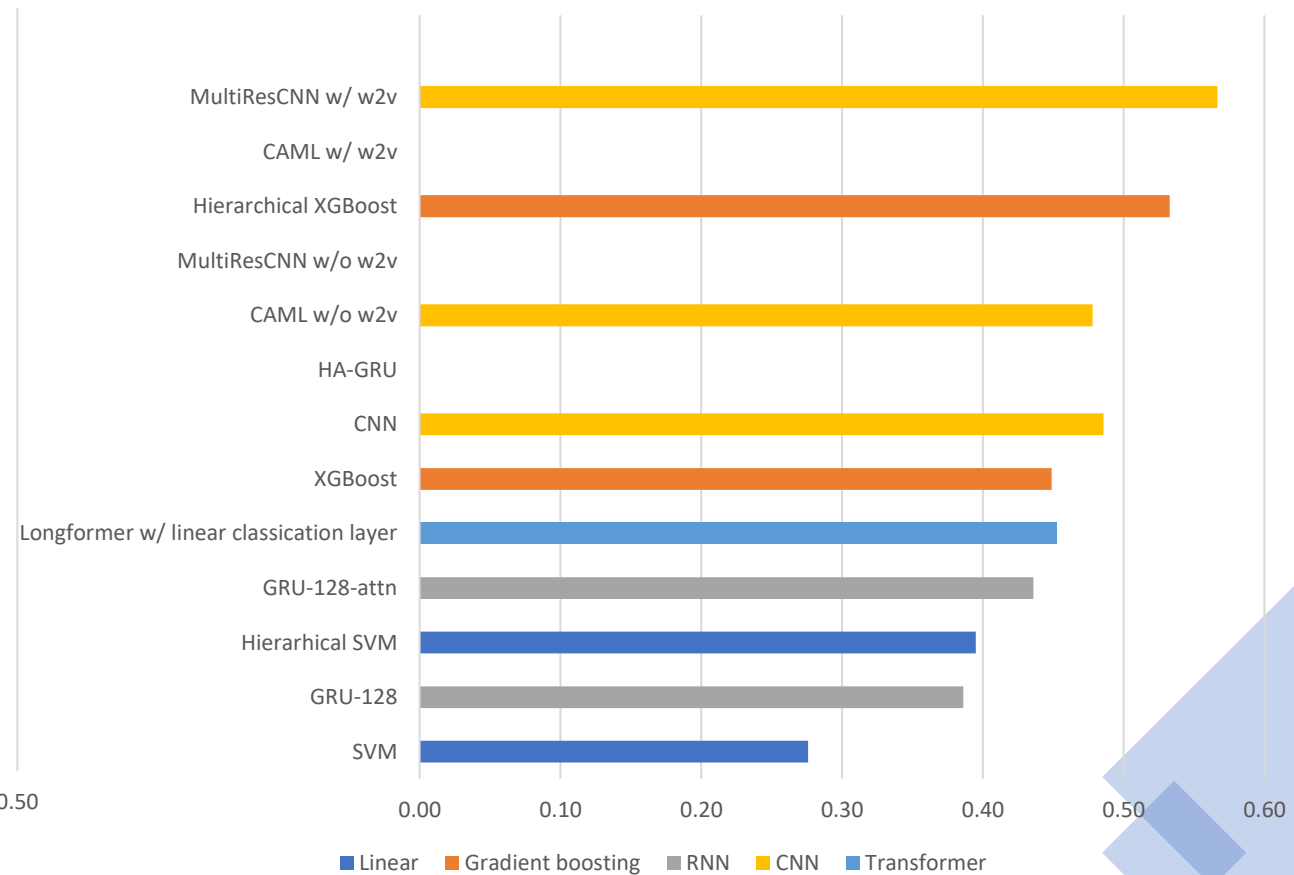
The models are prone to overfitting the data

# Summary of results

- Micro F1



- Hierarchical Micro F1





# Problem that I'm grappling with

---

How to force the neural nets to learn  
the hierarchy



*Thank you*

---

Stefan Strydom

stefan.strydom87@gmail.com





# References

1. Korsakov, Igor & Gusev, Aleksander & Kuznetsova, T. & Gavrillov, Denis & Novitskiy, Roman. (2019). P1923Deep and machine learning models to improve risk prediction of cardiovascular disease using data extraction from electronic health records. *European Heart Journal*. 40. 10.1093/eurheartj/ehz748.0670.
2. Li, Benjamin & Oh, Jeeheh & Young, Vincent & Rao, Krishna & Wiens, Jenna. (2019). Using Machine Learning and the Electronic Health Record to Predict Complicated *Clostridium difficile* Infection. *Open Forum Infectious Diseases*. 6. 10.1093/ofid/ofz186.
3. Wong, Andrew & Young, Albert & Liang, April & Gonzales, Ralph & Douglas, Vanja & Hadley, Dexter. (2018). Development and Validation of an Electronic Health Record–Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment. *JAMA Network Open*. 1. e181018. 10.1001/jamanetworkopen.2018.1018.
4. Tedeschi, Sara & Cai, Tianrun & He, Zeling & Ahuja, Yuri & Hong, Chuan & Yates, Katherine & Dahal, Kumar & Xu, Chang & Lyu, Houchen & Yoshida, Kazuki & Solomon, Daniel & Cai, Tianxi & Liao, Katherine. (2020). Classifying Pseudogout using Machine Learning Approaches with Electronic Health Record Data. *Arthritis Care & Research*. 10.1002/acr.24132.
5. Wong, Jenna & Horwitz, Mara & Zhou, Li & Toh, Sengwee. (2018). Using Machine Learning to Identify Health Outcomes from Electronic Health Record Data. *Current Epidemiology Reports*. 5. 10.1007/s40471-018-0165-9.
6. Slattery, Susan & Knight, Daniel & Weese-Mayer, Debra & Grobman, William & Downey, Doug & Murthy, Karna. (2019). Machine Learning Mortality-Classification in Clinical Documentation with Increased Accuracy in Visual-Based Analyses. *Acta Paediatrica*. 109. 10.1111/apa.15109.
7. Johnson, Alistair & Pollard, Tom & Shen, Lu & Lehman, Li-wei & Feng, Mengling & Ghassemi, Mohammad & Moody, Benjamin & Szolovits, Peter & Celi, Leo & Mark, Roger. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*. 3. 160035. 10.1038/sdata.2016.35.
8. Silla, Carlos & Freitas, Alex. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*. 22. 31-72. 10.1007/s10618-010-0175-9.