Topzorg voor uw levenskwaliteit zgt

UNIVERSITY OF TWENTE.

Comorbidity identification in clinical documents with weak supervision.

Sylvain Brouwer
Maurice van Keulen
Jeroen Geerdink
Johannes H. Hegeman

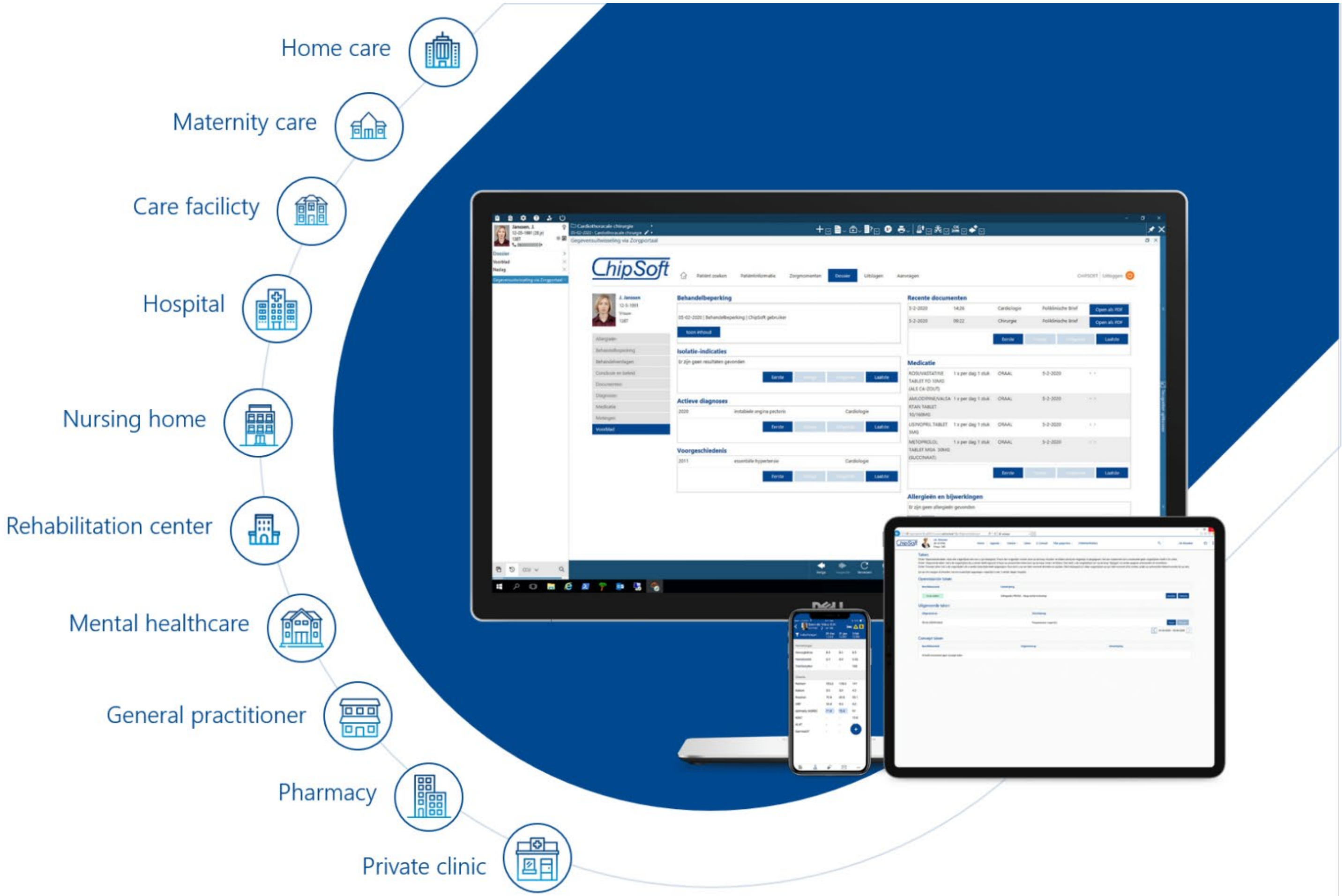
/// Comorbidity: Definition

Presence of additional chronic diseases concurrently with an index disease in one individual.[1]

[1] Valderas et al. (2009) Defining Comorbidity: Implications for Understanding Health and Health Services

/// Current Situation

- Electronic Health Record: HiX
- 80% of data is unstructured
  - Images
  - Documents
- 20% of data is structured
  - Lab measurements
  - Medication lists
  - List of diagnoses



/// Motivation

Topzorg voor uw levenskwaliteit

UNIVERSITY OF TWENTE.

Clinical Practice

- Clinicians would like a comprehensive overview of patient comorbidity.
- Comorbidities are buried in texts, not available immediately.

Research

- Comorbidities are important inputs for research and predictive models.
- Manual extraction of comorbidities from the EHR is a time-consuming task for large patient cohorts.

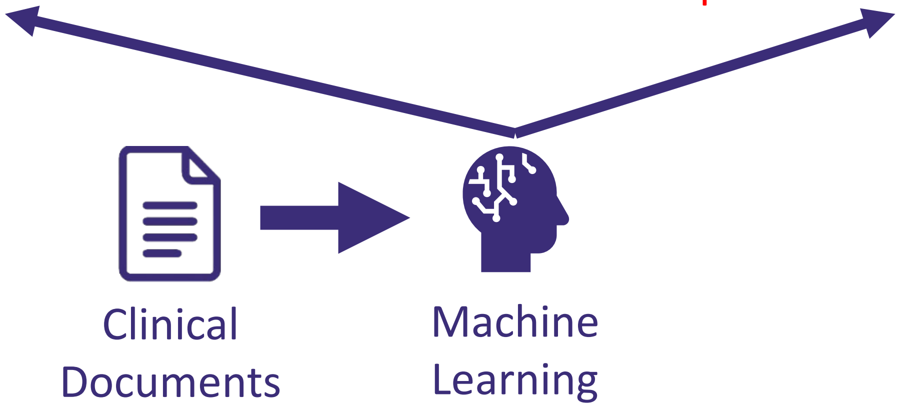
/// Motivation

Clinical Practice

- Clinicians would like a comprehensive overview of patient comorbidity.
- Comorbidities are buried in texts, not available immediately.
- Complete the overview.

Research

- Comorbidities are important inputs for research and predictive models.
- Manual extraction of comorbidities from the EHR is a time-consuming task for large patient cohorts.
- Replace manual annotation.



/// Research Question 1

Topzorg voor uw levenskwaliteit zgt

UNIVERSITY OF TWENTE.

Q1: How can we design a machine learning approach or artifact for obtaining relevant comorbidities from clinical notes?

Topzorg voor uw levenskwaliteit

UNIVERSITY OF TWENTE.

Methodology

/// How do we frame our problem?

complaint:
potential collum fracture r after fall

anamnesis:
heteroanamnesis due to dementia.
patient fell out of bed this morning, was no longer able to mobilize afterwards.

medical history:
hypertension, osteoporosis, dvt
2010 – claudicatio intermittens
2002 – knee fracture

lab: …

conclusion/therapy: …

How do we define identification / extraction?

What medical conditions are relevant?

/// How do we frame our problem?

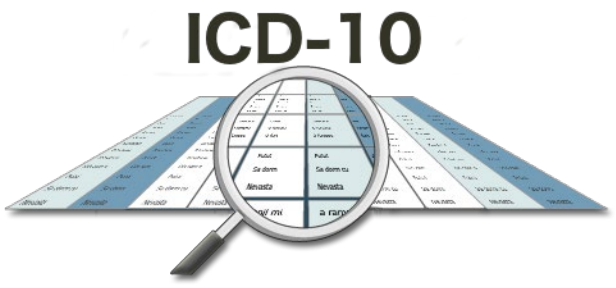
complaint:
potential collum fracture r after fall

anamnesis:
heteroanamnesis due to **dementia**.
patient fell out of bed this morning, was no longer able to mobilize afterwards.

medical history:
hypertension, **osteoporosis**, **dvt**
2010 – **claudicatio intermittens**
2002 – **knee fracture**

lab: …

conclusion/therapy: …



F03	Unspecified dementia
I10	Essential (primary) hypertension
M81	Osteoporosis without pathological fracture
I82	Other venous embolism and thrombosis

…

/// Relevant Conditions: Charlson Index

Weight	Condition
1	Peripheral vascular disease
	Dementia
	Myocardial infarction
	Chronic pulmonary disease
	Mild liver disease
	Congestive heart failure
	Peptic ulcer disease
	Cerebrovascular disease
	Diabetes, without chronic complications
	Rheumatic disease
2	Hemiplegia
	Renal disease
	Malignancy, except skin neoplasms
	Diabetes, with chronic complications
3	Moderate/severe liver disease
6	Metastatic solid tumor
	AIDS/HIV

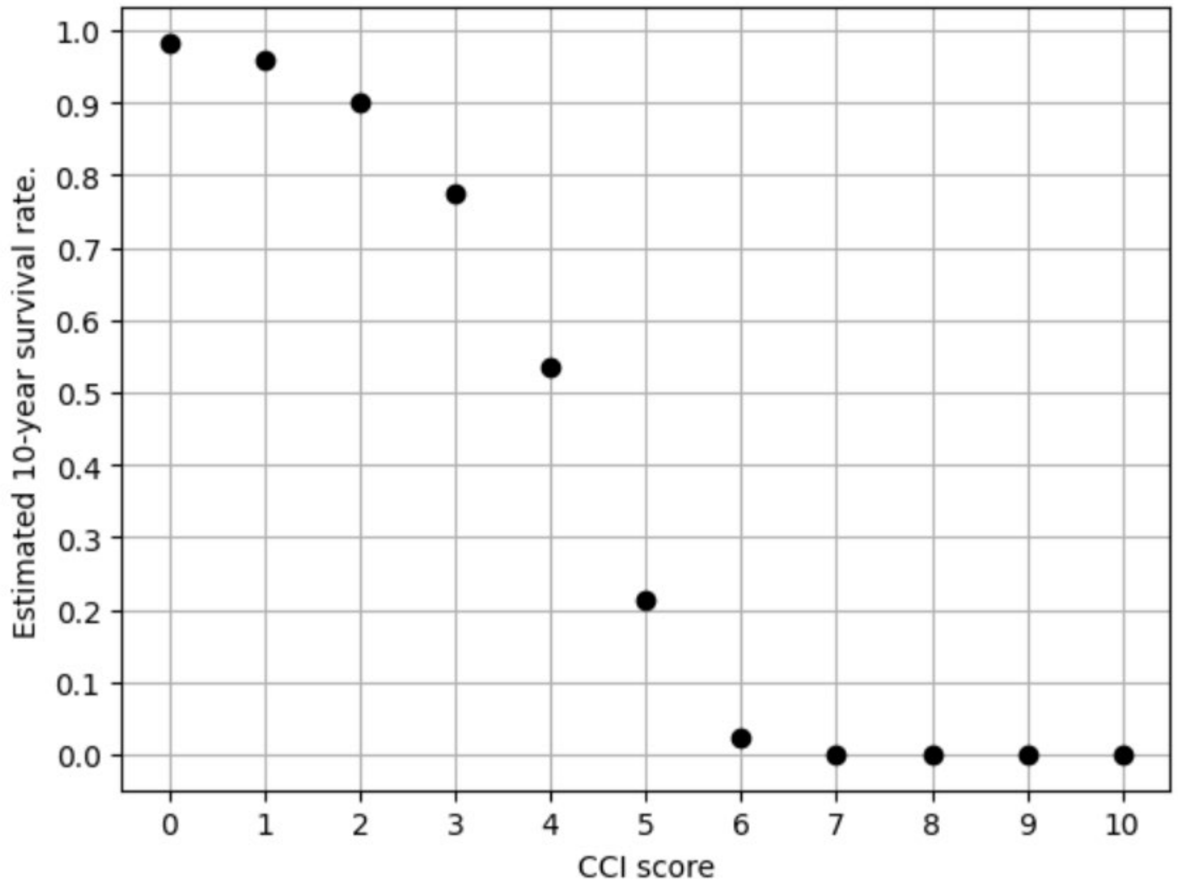


FIGURE 3.1: Estimated 10-year survival rate for CCI scores.

/// Classify at a document level

Topzorg voor uw levenskwaliteit

UNIVERSITY OF TWENTE.

complaint:
potential collum fracture r after fall

anamnesis:
heteroanamnesis due to **dementia**.
patient fell out of bed this morning, was no longer able to mobilize afterwards.

medical history:
hypertension, osteoporosis, **dvt**
2010 – **claudicatio intermittens**
2002 – knee fracture

lab: …

conclusion/therapy: …

Peripheral vascular disease

Dementia

/// Dataset

All documents:
Emergency department notes
Fractures due to trauma
age ≥ 70

Hip Fractures
n=3290

Hand annotated

TABLE 5.3: Occurrence rates of CCI categories in **DATA-HIP**

Category	Occurrence rate
Cerebrovascular disease	0.188
Dementia	0.170
Congestive heart failure	0.153
Diabetes, without chronic complications	0.147
Malignancy, except skin neoplasms	0.146
Chronic pulmonary disease	0.136
Peripheral vascular disease	0.121
Renal disease	0.089
Rheumatic disease	0.086
Myocardial infarction	0.078
Diabetes, with chronic complications	0.047
Hemiplegia / paraplegia	0.024
Metastatic solid tumor	0.020
Peptic ulcer disease	0.020
Mild liver disease	0.009
Moderate / severe liver disease	0.003
AIDS / HIV	0.000

/// Dataset: Class Imbalance

Topzorg voor uw levenskwaliteit zgt

UNIVERSITY OF TWENTE.

All documents:
Emergency department notes
Fractures due to trauma
age ≥ 70

Hip Fractures
n=3290

Hand annotated

TABLE 5.3: Occurrence rates of CCI categories in **DATA-HIP**

Category	Occurrence rate
Cerebrovascular disease	0.188
Dementia	0.170
Congestive heart failure	0.153
Diabetes, without chronic complications	0.147
Malignancy, except skin neoplasms	0.146
Chronic pulmonary disease	0.136
Peripheral vascular disease	0.121
Renal disease	0.089
Rheumatic disease	0.086
Myocardial infarction	0.078
Diabetes, with chronic complications	0.047
Hemiplegia / paraplegia	0.024
Metastatic solid tumor	0.020
Peptic ulcer disease	0.020
Mild liver disease	0.009
Moderate / severe liver disease	0.003
AIDS / HIV	0.000

/// Dataset: Phase 2

All documents:
Emergency department notes
Fractures due to trauma
age ≥ 70

Hip Fractures
n=3290

Hand annotated

Other Fractures
n=20897

Weakly annotated

TABLE 5.3: Occurrence rates of CCI categories in **DATA-HIP**

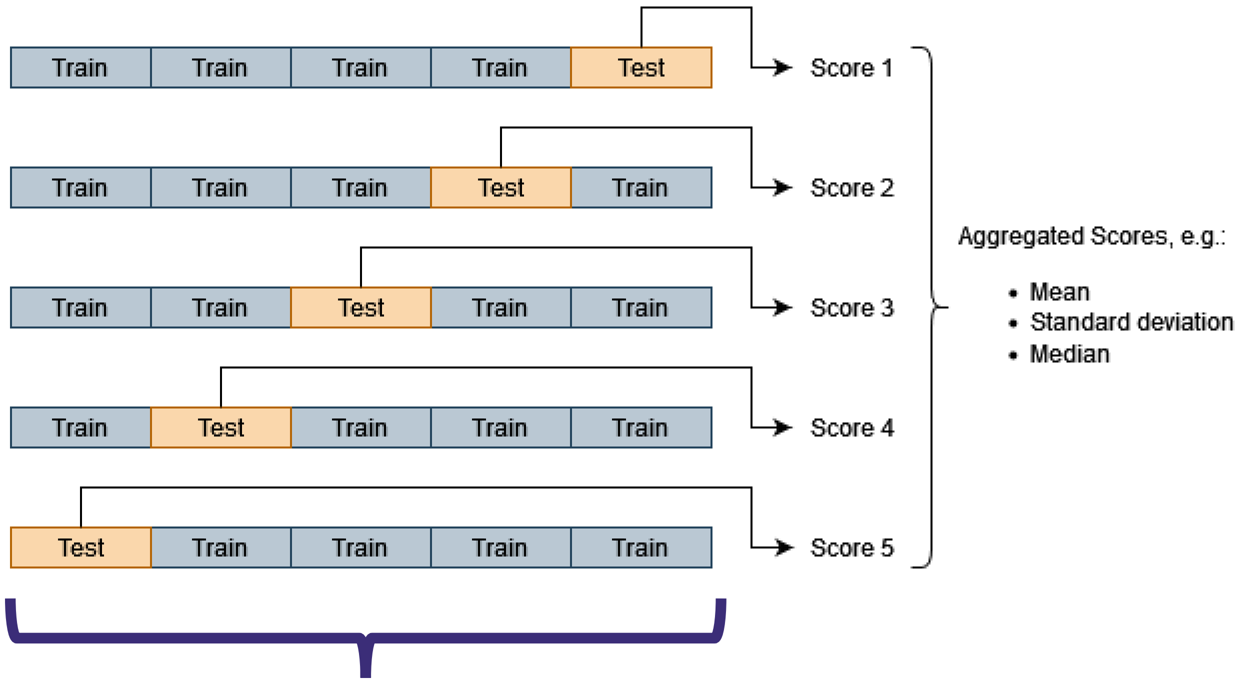
Category	Occurrence rate
Cerebrovascular disease	0.188
Dementia	0.170
Congestive heart failure	0.153
Diabetes, without chronic complications	0.147
Malignancy, except skin neoplasms	0.146
Chronic pulmonary disease	0.136
Peripheral vascular disease	0.121
Renal disease	0.089
Rheumatic disease	0.086
Myocardial infarction	0.078
Diabetes, with chronic complications	0.047
Hemiplegia / paraplegia	0.024
Metastatic solid tumor	0.020
Peptic ulcer disease	0.020
Mild liver disease	0.009
Moderate / severe liver disease	0.003
AIDS / HIV	0.000

/// K-fold Validation (K=10)

- Individual groups of diagnoses:

$$f_1 = 2\frac{Precision * Recall}{Precision + Recall}$$

- Entire document: accuracy (%documents with correct labels)



Hip Fractures n=3290

Hand annotated

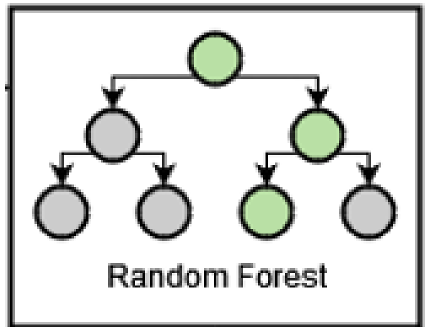
Phase 1: Full Supervision

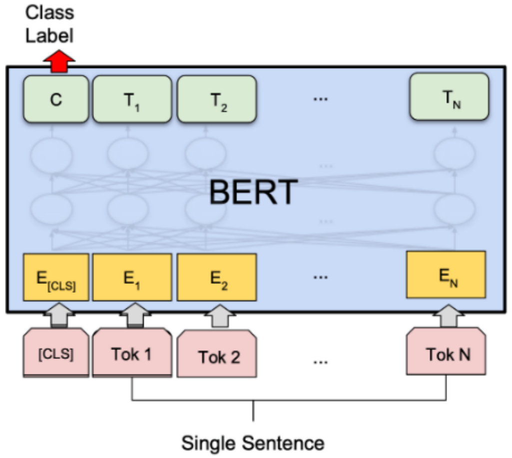
/// Full supervision

- Straightforward, baseline approach.

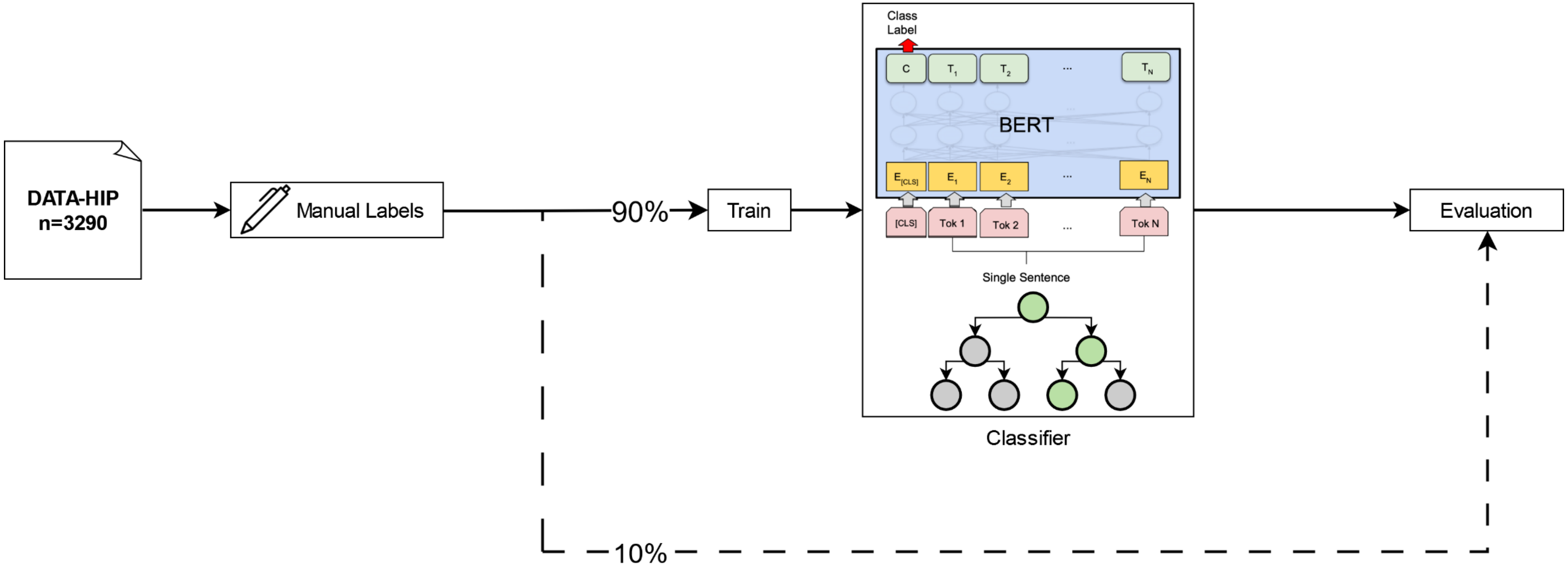
$$P(C|D) = \frac{P(D|C)P(C)}{P(D)} \propto P(D|C)P(C)$$

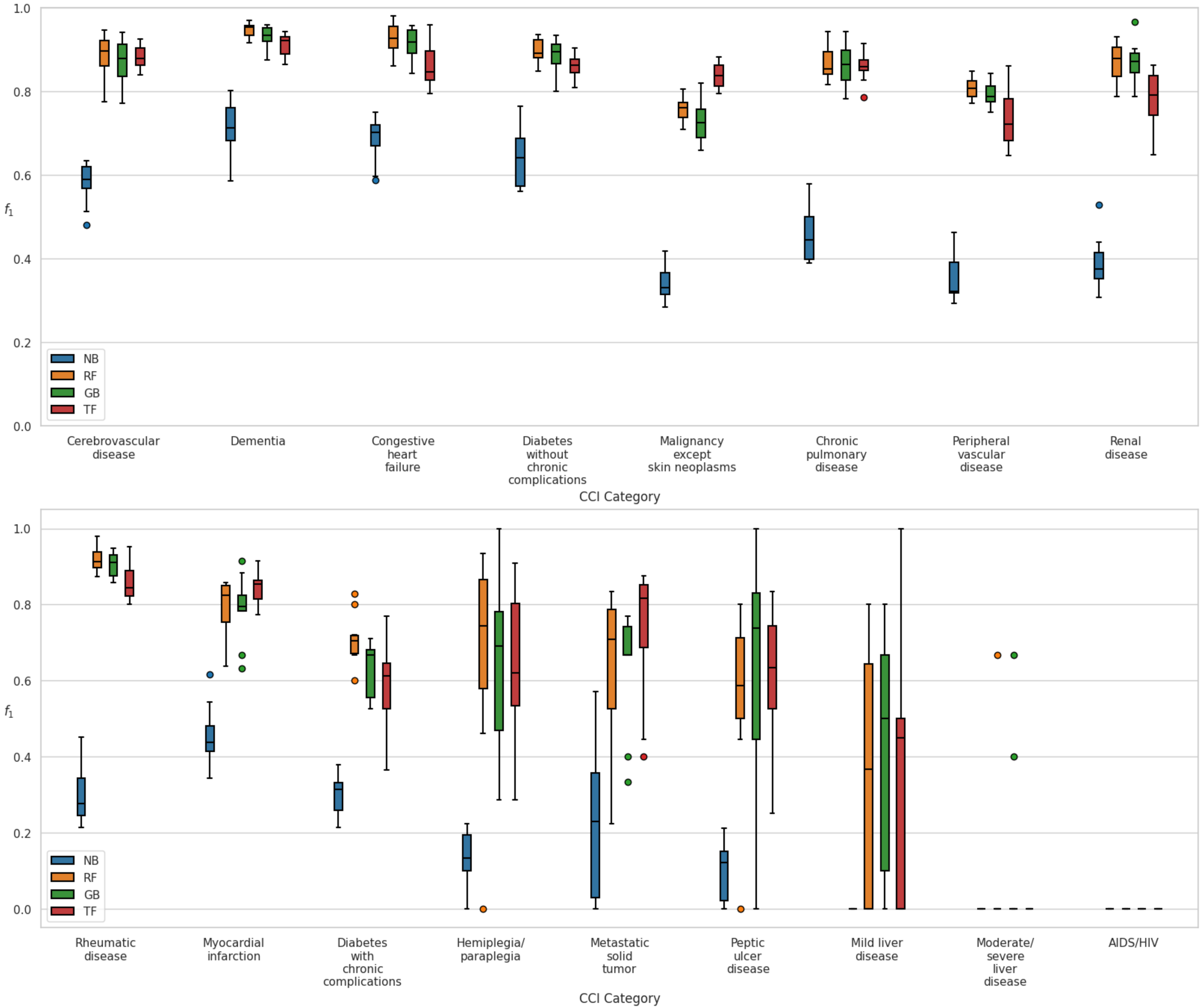
- 4 Considered models:
    - Naïve Bayes
    - Gradient Boosted Trees
    - Random Forest
    - Transformers (BERT / RoBERTa)





/// Full supervision (1 fold)





Topzorg voor uw levenskwaliteit

UNIVERSITY OF TWENTE.

- > 5% occurrence rate:
- < 5% occurrence rate:
- Best classification accuracy: Random Forest - 71%

Topzorg voor uw levenskwaliteit

UNIVERSITY OF TWENTE.

Phase 2: Weak Supervision

/// How can we generate enough examples of rare conditions?

- Literature links the Charlson Index to SNOMED CT[1]

- Can we look for the terms of relevant SNOMED concepts in our documents?

SNOMED CT
The global language of healthcare

[1] Stephen Fortin, Jenna Reps, and Patrick Ryan. “Adaptation and validation of a coding algorithm for the Charlson Comorbidity Index in administrative claims data using the SNOMED CT standardized vocabulary”

/// Research Question 2

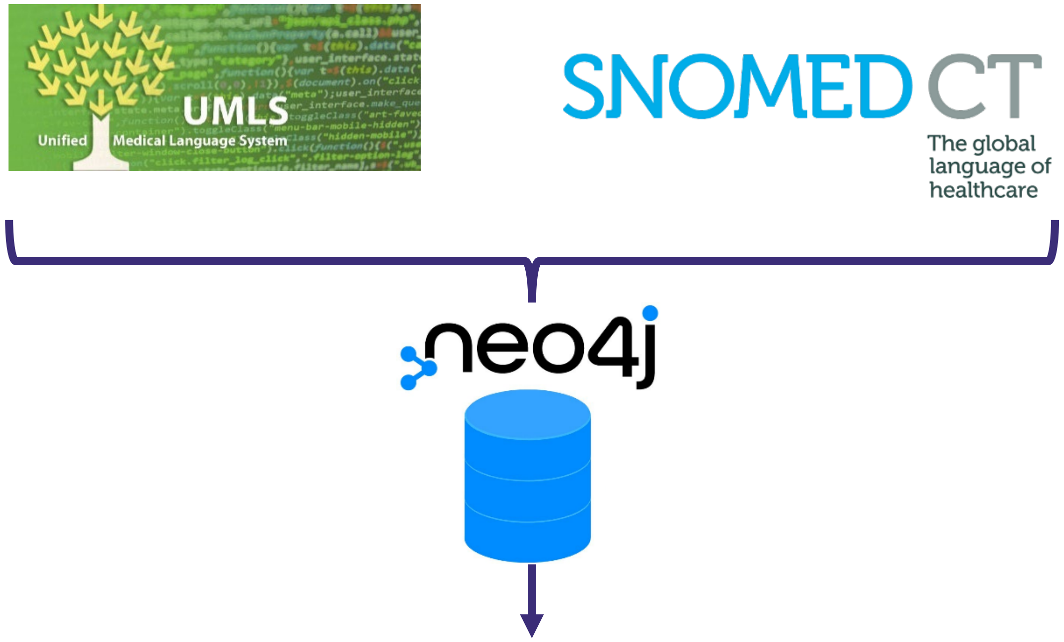
Q2: How can we leverage existing medical terminologies and ontologies in labeling sufficient training data?

/// Weak Supervision: General Approach

1. Aggregate terminologies onto SNOMED CT

2. Retrieve relevant terms for CCI categories from SNOMED

3. Check for occurrences of terms from retrieved list in unlabeled documents



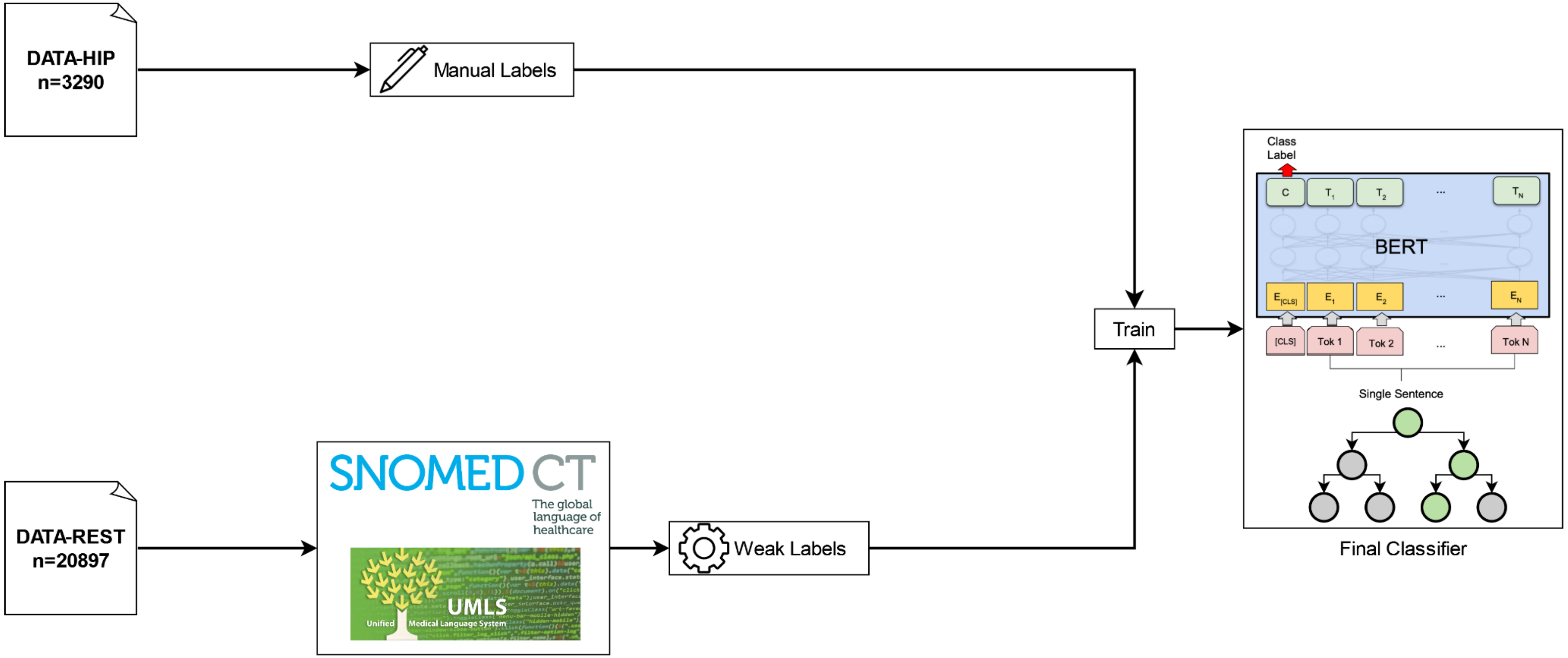
Per-class terminology list

- ❑ Term 1 in document?
- ❑ Term 2 in document?
- ❑ …
- ❑ …
- ❑ …

Other Fractures
n=20897

Weakly annotated

/// Weak Supervision Pipeline



/// Problem 1: Mismatch in language

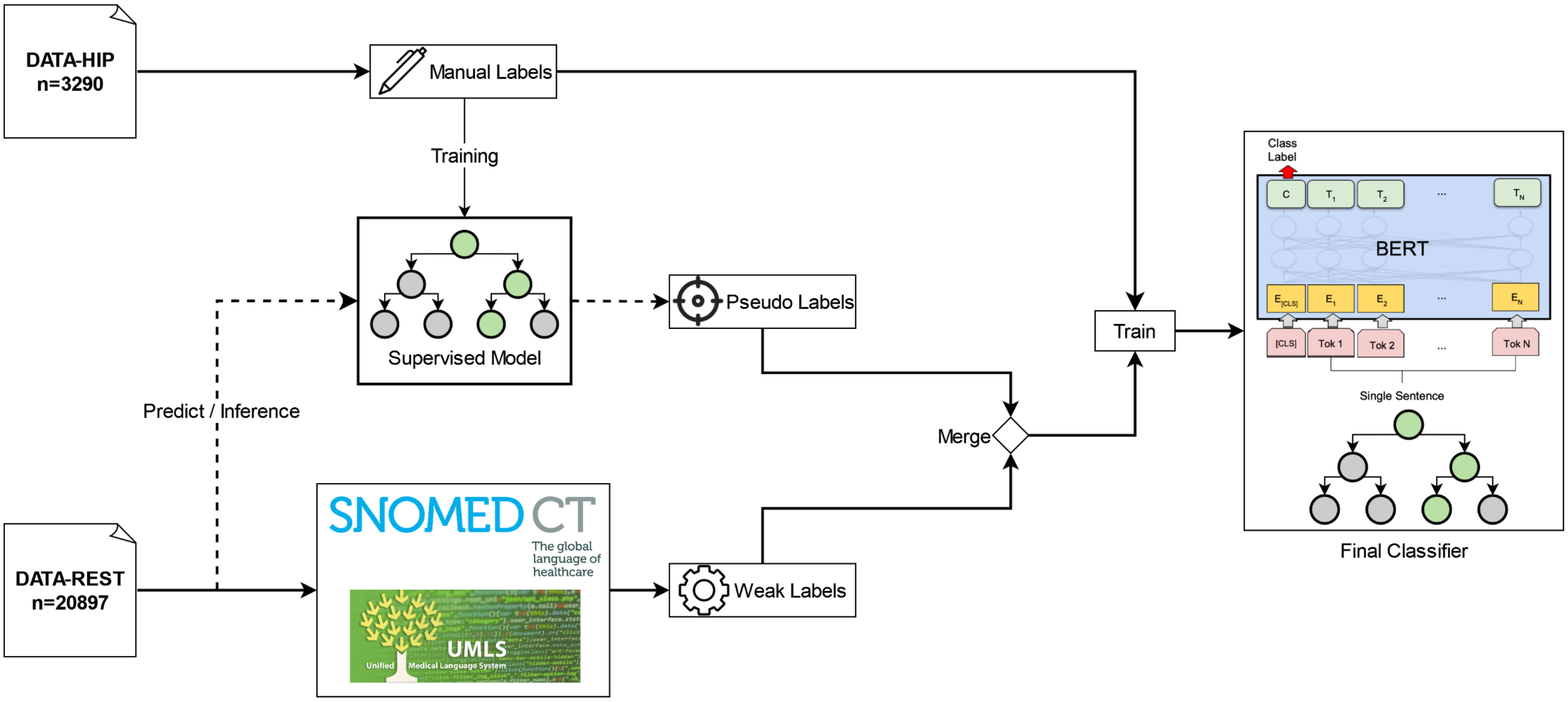
- Clinicians often use terms or phases that can not be found in medical terminologies like SNOMED CT.

“hemibeeld” instead of “hemiplegie” / “hemiparese”

“diabetes met voetafwijking” instead of “diabetische voet”

Our solution:
Pseudo-labeling:
1. *Train a supervised classifier based on hand-annotated data.*
2. *Have supervised classifier predict labels for unannotated data.*
3. *Augment keyword-based weak labels with predicted (pseudo-) labels.*

/// Weak Supervision + Pseudo-labeling



/// Problem 2: Abbreviations

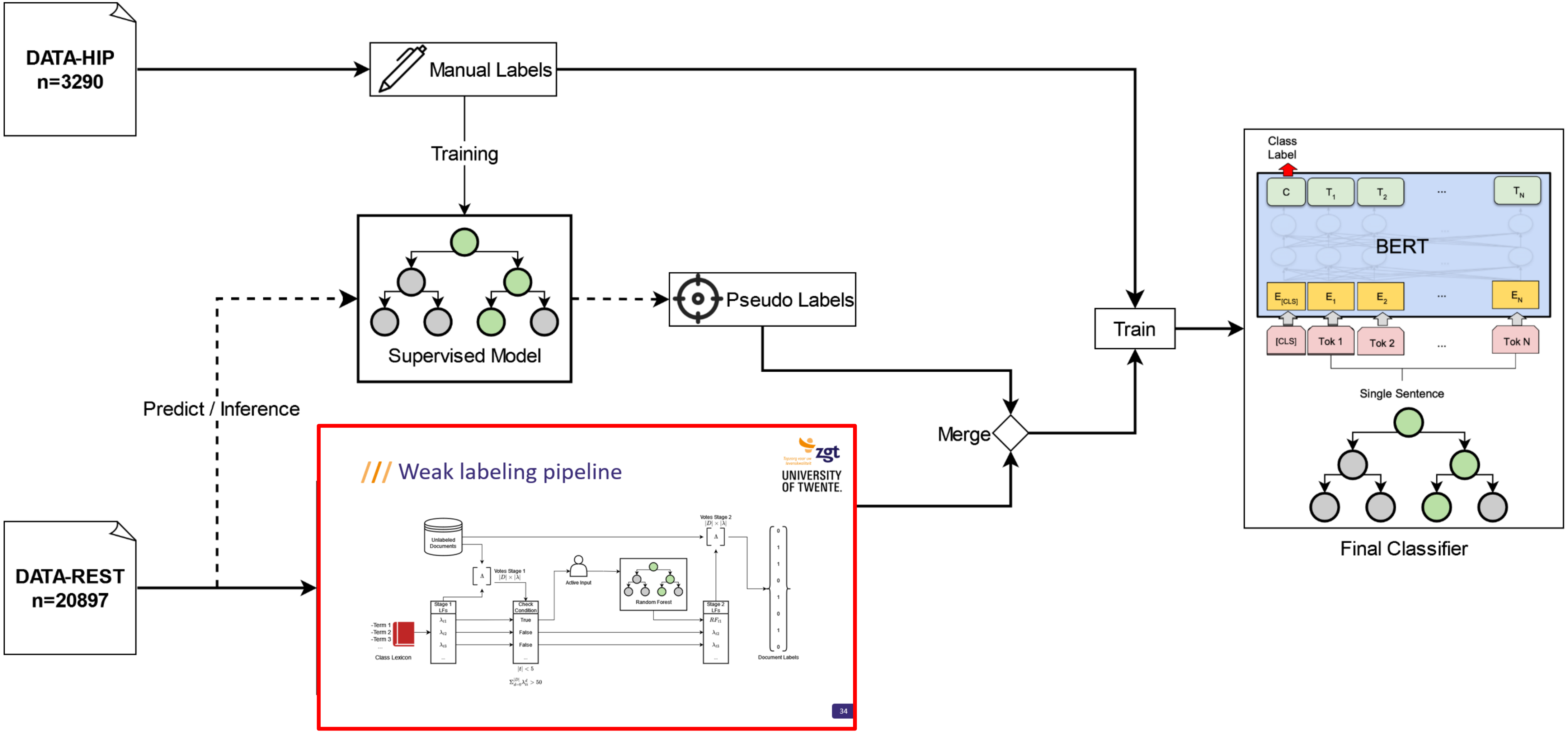
pta

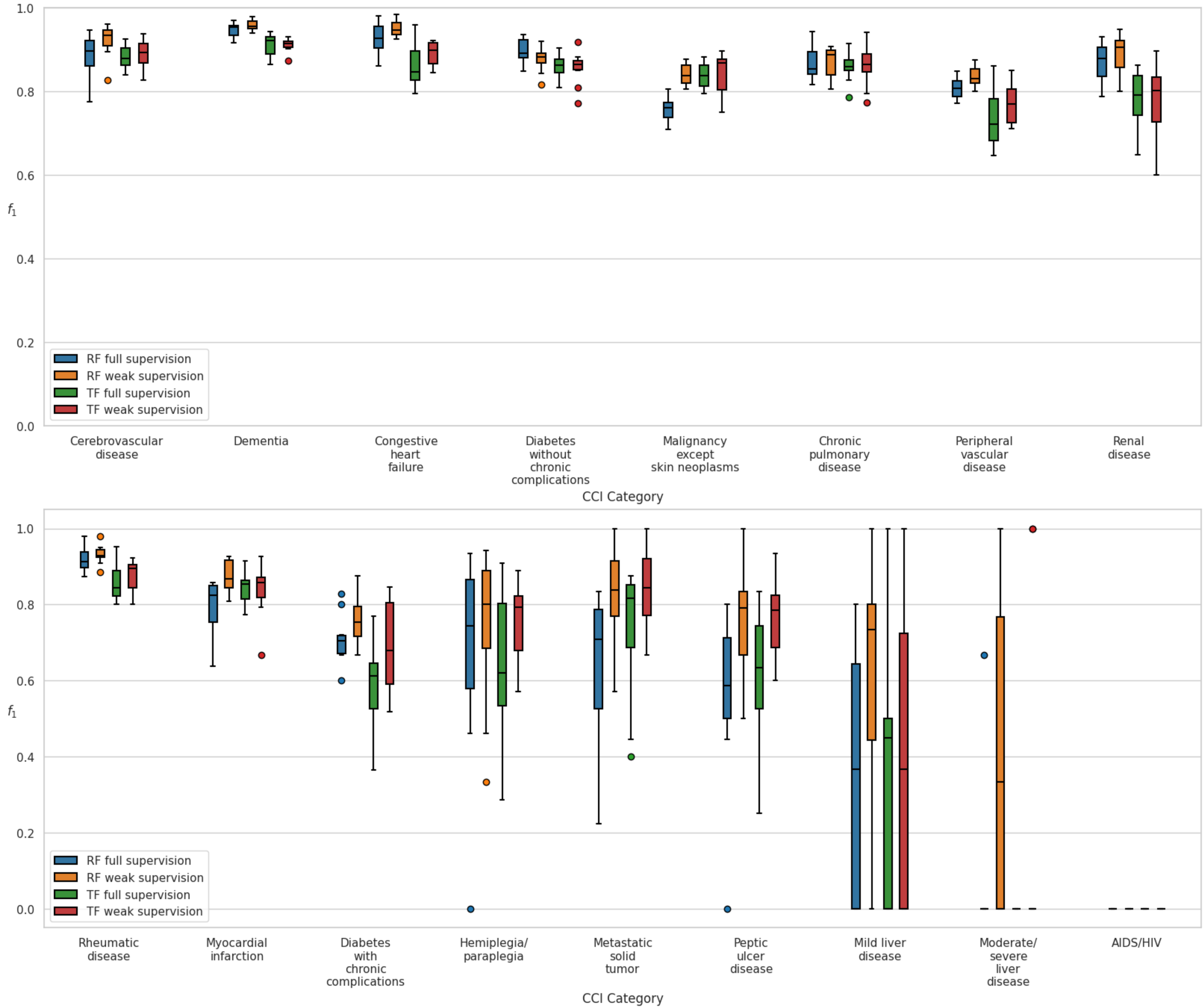
- angioplasty — Peripheral vascular disease
- bladder cancer — Peripheral vascular disease

mi

- myocardial infarction — Myocardial Infarction
- mitral valve insufficiency — Myocardial Infarction

/// Full Training Pipeline





ZGT Topzorg voor uw levenskwaliteit

UNIVERSITY OF TWENTE.

- Improvements in f1 score: 0.05-0.35 for <5% categories.
- Best classification accuracy: Random Forest - **75%**
- **92%** of documents were within 1 CCI point

/// Takeaways

- Random Forests + Weak supervision performed best.
  - Classification accuracy of **75%**. (71% w/o weak supervision)
  - Within 1 point of the correct CCI score in **92%** of test cases. (89% w/o weak supervision)

- Weak supervision with terminologies is effective at generating samples at low cost but care should be taken to bridge the language gap between terminologies and practice.
  - Small amount of hand-labeled data.
  - **Pseudo labeling.**
  - Maintain list of nonstandard vocabulary.
  - **Disambiguation of abbreviations.**

/// Applicability

Clinical Practice

- The achieved accuracy (75% + 92% within 1 point) is insufficient for completing structured “problem lists” in the EHR.
- May be used to present some aggregated metric of comorbidity (e.g. CCI score).

Research

- The achieved accuracy (75% + 92% within 1 point) may be sufficient for feature extraction and annotation in future research.
- This is especially the case for research regarding elderly patients.
- ZGT is currently continuing work regarding postoperative mortality prediction.

Topzorg voor uw levenskwaliteit

UNIVERSITY OF TWENTE.

Thank You!

sylvainbrouwer@gmail.com

https://github.com/SylvainBrouwer/

VOORUITSTREVEND

VERBINDEND

MET OPRECHTE AANDACHT

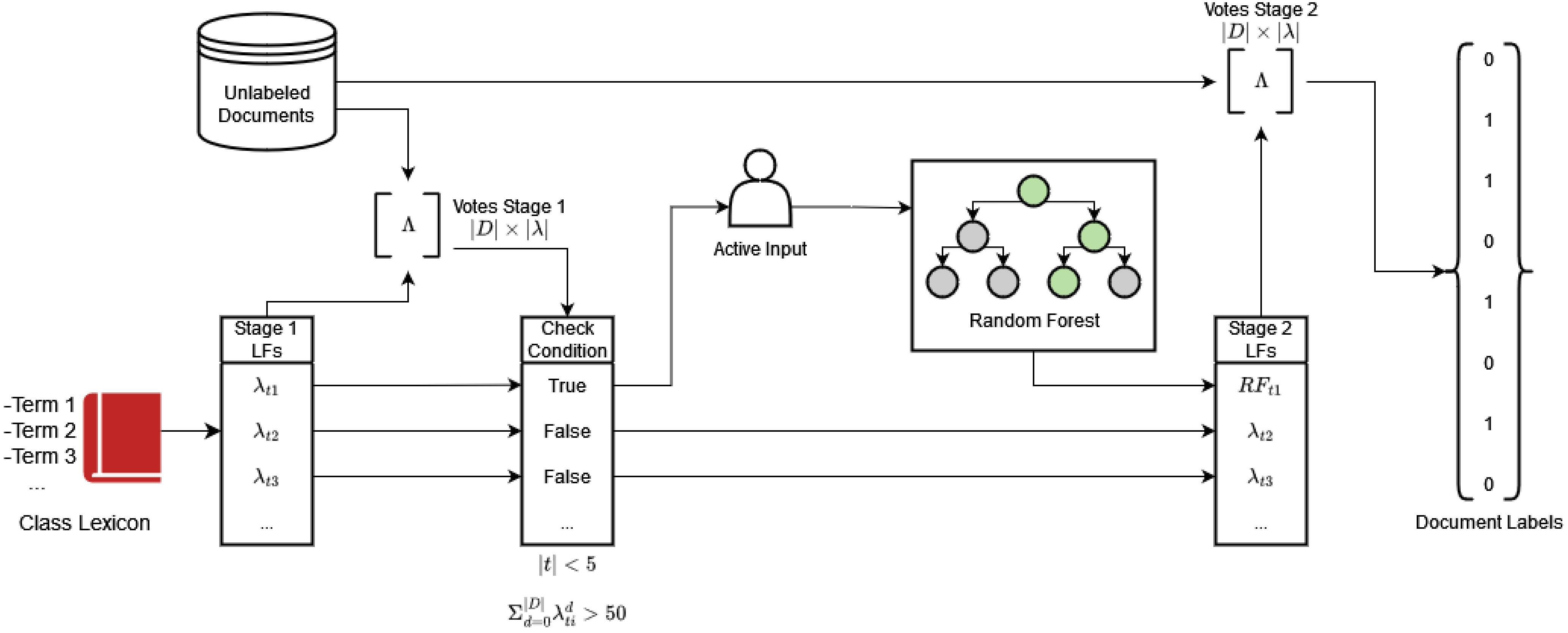
Attributions

Template:
- Hospital Group Twente (ZGT)

Images:
- https://www.istockphoto.com/nl
- https://www.chipsoft.com

Icons:
- Vitaly Gorbachev @ https://www.flaticon.com/authors/vitaly-gorbachev
- https://www.freepik.com/
- https://www.cleanpng.com/

/// Weak labeling pipeline



/// Weak labeling pipeline