Predicting Associated Comorbidities of Obesity from MIMIC IV Clinical Notes

Stanford CS224N Custom Project

Peyton Chen
Department of Computer Science
Stanford University
peytonc@stanford.edu

Ana Selvaraj
Department of Computer Science
Stanford University
aselvara@stanford.edu

Om Jahagirdar
Department of Bioengineering, Department of Computer Science
Stanford University
ojahagir@stanford.edu

Abstract

Our goal is to create a successful NLP deep-learning model to predict diseases associated with obesity from clinical notes, namely diabetes and hypertension. This is important for the field of automated machine learning from a biomedical perspective and can improve health outcomes while cutting healthcare costs (Waring et al., 2020); since it reduces the amount of human labor if we could predict common health conditions just from clinical notes. For our clinical notes dataset, we use MIMIC-IV as it is a large and free database comprising recent de-identified health-related data.

We compare and analyze multiple models' performance and optimizations for predicting diabetes and hypertension. These model variants include Logistic Regression with BoW, BERT, BioDischargeSummaryBERT (trained on ICU discharge summaries) and BioDischargeSummaryBERT with class weights to combat class imbalance. Our baseline, data pre-processing and graph generation code were written from scratch, while the other models were modified and tuned versions of pretrained models.

1 Key Information to include

- Mentor: Ansh Khurana
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

Using machine learning in the biomedical field to reduce human labor has been seen as a potential solution to reduce rising, unaffordable healthcare costs in America (Waring et al., 2020). Using clinical notes to predict diseases can help nurses and doctors to promptly verify and diagnoses those conditions quickly. This is a time-saving process since clinical notes are long and require human experts to interpret and suggest possible diagnoses.

By predicting diagnoses of common conditions associated with morbid obesity, an epidemic in America, that is associated with other health conditions, we can automate the mapping of clinical notes to diagnoses and reduce the need for tedious human labor from experts. Current research has not examined the performance of transformers on this task. Previous papers to predict diseases with

the top common ICD codes have used AWD-LSTM and achieved a model accuracy of 80% such as Nuthakki et al. (2019) using the MIMIC III dataset. Similarly, researchers have used BERT to predict common ICD code diagnoses with accuracy scores of around 90%.

Our task is slightly different because we want to predict health conditions (which are more sparse amongst the data) not specific common ICD codes. For example, diabetes, a condition associated with obesity, has multiple different ICD codes. However, in our task, we predict the diagnosis of any type of diabetes as well as the diagnosis of any type of hypertension. This was done because of the sparseness of diagnoses of each type of diabetes in the MIMIC IV dataset (which has de-identified patient data from a specific medical center from 2011 to 2019). Furthermore, our data is long text since MIMIC IV contains discharge summaries that merge all clinical notes for a patient into one row while MIMIC III has individual clinical notes.

Our key contribution is comparing multiple models' prediction of the presence of comorbidities associated with obesity in light of the class imbalance often seen in medical data. Because of the fact that not all comorbidities had meaningful positive samples, we chose two common comorbidities: diabetes and hypertension. However, our model structures can generalize to other comorbidities. These models are simple logistic regression with Bag of Words features, BERT, BioDischargeSummaryBERT, and BioDischargeSummaryBERT with computed class weights. We found that our tuned version of the BioDischargeSummaryBERT with class weights to account for class imbalance performed best.

3 Related Work

Our task was inspired by the 2008 i2b2: Informatics for Integrating Biology and the Bedside challenge about 'Recognizing Obesity and Co-morbidities in Sparse Data' Uzuner (2009). It involved classifying 1237 discharge summaries on predicting the presence of obesity and its associated comorbidities for those patients. However, their dataset was too small for us to use our models on so we decided to use MIMIC IV instead. Across the comorbidities, they recommended using F_1 macro and micro scores.

An important paper (Solt et al., 2009) to attempt this challenge found promising results with an SVM and macro F_1 score 0.67 and micro F_1 score 0.95. This inspired us to implement a baseline model that was a simple logistic regression classifier with Bag of Words as features.

Another influential piece of work for our project was by Yao et al. (2019) which sought to improve clinical text classification. It introduced deep learning to attempt the i2b2 obesity challenge. The authors used a CNN and reported a macro F_1 scores 0.67 and micro F_1 scores 0.96, similar to the SVM. However, they observed, when evidence for a condition was explicit in the text, their model performed better than the SVM and was more effective for learning hidden features.

Using NLP models on clinical notes for other tasks is a dense research area. However, most research has been on LSTM architectures to classify relations from clinical notes. One paper (Singh et al., 2020) used BERT to predict the most common ICD diagnoses from clinical notes from MIMIC III to get an F_1 accuracy of 92%. This showed us that using transformers for clinical text classification was still a promising step.

We also found it interesting to analyze the different publically available BERT embeddings that are used for clinical text classification (Alsentzer et al., 2019). For example, BioDischargeSummaryBERT is trained on MIMIC III ICU discharge summaries and is good on domain-specific tasks. This prompted us to consider what performance boost one would gain by using BioDischargeSummaryBERT for our task compared to BERT.

Based on the research we surveyed, we wanted to compare the performance of different variants of models to predict comorbidities associated with obesity by classifying discharge summaries from the whole MIMIC IV dataset.

4 Approach

The models we ran were:

- Simple logistic regression classifier with Bag of Words as features (Primary Baseline)

- BERT (Secondary Baseline)
- BioDischargeSummaryBERT
- BioDischargeSummaryBERT with class weights

We used BERT transformer architecture to see how it would perform on our given task and whether it would be more effective than simpler logistic regression models. Since BioDischargeSummaryBERT is trained on clinical notes from MIMIC, we hoped BioDischargeSummaryBERT would perform better than BERT since it was trained on clinical texts from MIMIC III. We also wanted to replicate the results of existing research to see whether multilabel BERT models would generalize well to our specific task. We used class weights because the diagnoses for diabetes and hypertension are sparse causing tremendous class imbalances in our preliminary results. This is particularly common in medical datasets and so it is important to combat it with appropriate methods Rahman and Davis (2013).

Class weights were calculated as follows: n_samples / (n_classes * np.bincount(y)). These weights are then passed into the loss functions for each of the models (except the baseline) to be incorporated when calculating the loss.

Our baseline model was simple, built with sklearn's LogisticRegression configured to a maximum of 10,000 iterations and sklearn's CountVectorizer to generate the bag of words features. This was our initial baseline model based on the method described here Solt et al. (2009) whose results are almost the same as a CNN on a prediction task (Uzuner, 2009) similar to ours.

Our BERT models used bert-based-uncased and padded each tokenizer input to the max length of 318. The models first applied a dropout layer of 0.5, a linear transformation layer to reduce the number of channels from 768 to 2 and finally, we applied the rectified linear unit function (ReLU). We used cross-entropy loss with our without class weights. We found this configuration after hyperparameter tuning on smaller versions of the subset (that contained 1000 rows). We used Adam for the optimizer since it performs better on sparse data.

Our code is here: `https://github.com/ana13S/cs224n_project`. We used the BERT tutorial here as a starting point and heavily edited them to use our layers, loss functions and evaluation metrics.

5 Experiments

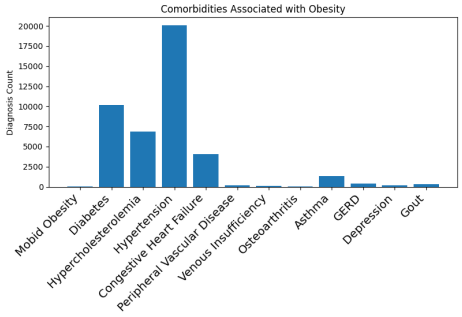
5.1 Data

Our dataset is MIMIC-IV v2.2 (Johnson et al., 2021) which details critical care data for over 40,000 patients from Beth Israel Deaconess Medical Center.
To gain access to this data, we completed IRB training on medical data ethics and compliance laws. We also submitted an application to the database owners detailing our project to receive approval and credentialing to use the dataset.

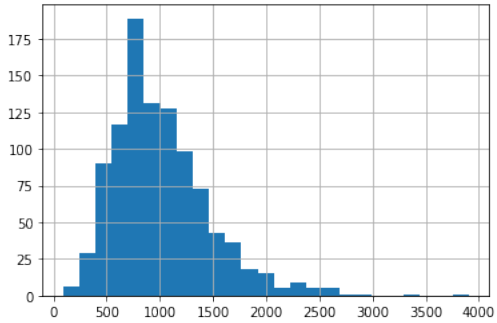
All patient identifiers have been removed from the data. When handling the data, we took precautions to make sure none of our generated csv files were public so we would be following our data usage agreement.

Our associated task is predicting whether a patient was diagnosed with any type of diabetes or hypertension. This was done by looking at the icd_title of each diagnosis of a specific patient and marking a patient based on whether their diagnosis description has the string 'diabetes' or 'hypertension'.

For our task prediction, we used diabetes and hypertension since the diagnoses are the least sparse among all comorbidities of obesity 5.1. An important thing to note is that previous research on the prediction of the top 10 common ICD diagnoses in MIMIC does not include comorbidities associated with obesity Singh et al. (2020). This means that the same model will perform differently in the prediction of the presence of common ICD diagnoses compared to the prediction of diabetes and hypertension (since these diagnoses are much more sparse).



(a) Diagnoses Counts of Comorbidities



(b) Sequence Length of Pre-processed Discharge Summaries

Figure 1: Dataset Features

For data processing of the discharge summaries, we removed special characters, words with digits, repeated data tokens in each discharge summary, punctuation and stop words using nltk. Generally, our text sequences can be considered long text 5.1 compared to other common datasets BERT is used with such as Tweet Sentiment analysis. Our complete dataset was 1 GB with 145,915 unique patient notes. This was filtered from 450K patient notes by randomly choosing one note per patient. Our training-testing-validation split is 80-10-10. We made sure each dataset was stratified properly and retained accurate class proportions with iterative_train_test_split using skmultilearn. Each row in our model input csv file would have a long text sequence representing the pre-processed discharge summary and binary numbers for each of the comorbidities.

5.2 Evaluation method

Our main evaluation method will be derived from Yao et al. (2019) This means in addition to the normal loss and accuracy metrics, we will be comparing F_1 macro and F_1 micro scores. F_1 macro scores will be based on the average performance for each class (Present/Absent) and this is so that a small class will be weighted the same as the larger class. F_1 micro scores, on the other hand, will also be used, but it gives all samples equal weight, which will mean that results will be skewed by the performance of the larger classes, an issue for our dataset due to extreme class imbalance.

5.3 Experimental details

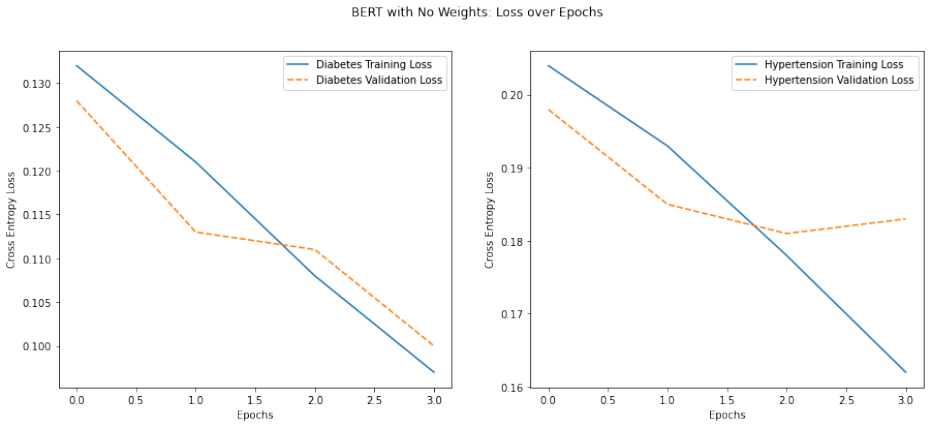
We ran all but one of our experiment on Google Colab Pro, mentioned in the BioDischargeSummary-BERT with weights section where we used AWS instead.

- **Simple logistic regression classifier with Bag of Words as features (Primary Baseline)**: For this model, the only dataset we were able to run on this model was a special dataset with only 5,000 entries as above this many entries the model consumed RAM at such an amount that the Python kernel would crash and shut down, demonstrating a limitation to this approach.
- **BERT (Secondary Baseline):** For all transformer models, we ran it for 4 epochs with a learning rate of 1e-6. For this model, we ran our default dataset which was a proportional 10% (meaning the proportion of positive to negative samples for the truncated dataset remained the same) of the the total 145,915 entries. We used a multi-label splitting algorithm such that all the models were trained on the same train-validation-test splits. We predicted both diabetes and hypertension without class weights in consideration in the loss function and were ran for 4 epochs.
- **BioDischargeSummaryBERT :** For this model, we also ran our default dataset which was a proportional 10% of the the total 145,915 entries. We predicted both diabetes and hypertension without class weights in consideration in the loss function and were ran for 4 epochs.

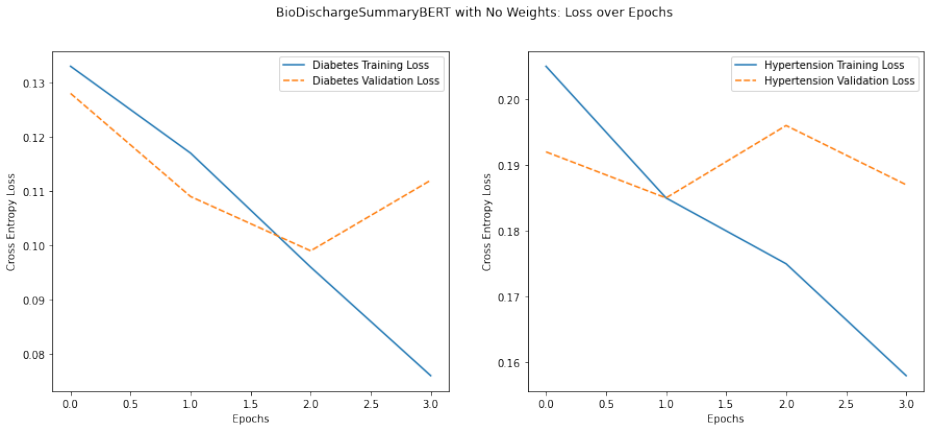
- **BioDischargeSummaryBERT with weights:** We ran the same model in the same configuration above with the exception of addition of class weights for both diabetes and hypertension. We expected this model, specifically for diabetes classification, to perform the best, so we utilized an AWS p3.2xlarge instance to train the BioDischargeSummaryBERT with weights on diabetes for the full 145,915 dataset.

5.4 Results

- **Logistic Regression with BoW feature classifier (Primary Baseline)**
  For the diabetes baseline, we observed overfitting, in addition to the incapability of processing our full dataset as mentioned above. It yielded a micro-F_1 and macro-F_1 score of 1.0 during training but a micro-F_1 score of 0.626 and a macro-F_1 score of 0.399.
  For the hypertension baseline, we see something similar, with even worse overfitting and the same limitations as the diabetes baseline. It yielded a micro-F_1 score and macro-F_1 score of 1.0 during training but a micro-F_1 score of 0.286 and macro-F_1 score of 0.238.

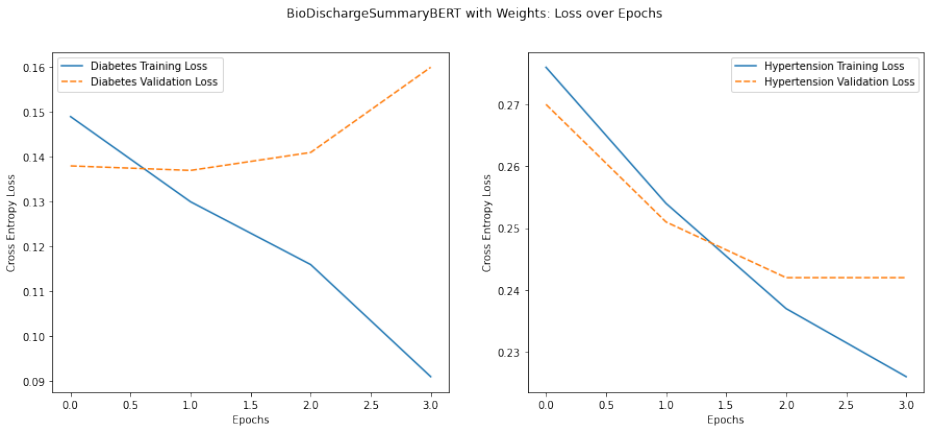


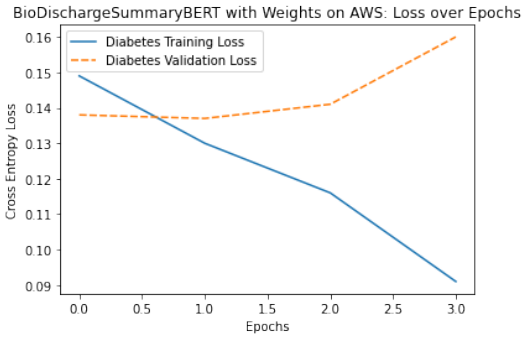
- **BERT (Secondary Baseline)**
  For the diabetes BERT classifier, we observed that both the training and validation loss similarly drop over time, suggesting that the model is fitting such that it generalizes past the training data. It yielded a test micro-F_1 score of 0.93 macro-F_1 score of 0.54. This discrepancy of scores, which we see in other models shown below, is indicative of the class imbalance we know to occur in the dataset that is common in medical data.
  For the hypertension BERT classifier, we observed that both the training loss similarly drops over time, but the validation loss drops and then slightly increases in the last epoch. This suggests that the model may overfit to the training data in the final epoch. It yielded a test micro-F_1 score of 0.86 macro-F_1 score of 0.47.



- **BioDischargeSummaryBERT**

For the diabetes BioDischargeSummaryBERT classifier, we observed that both the training loss similarly drops over time, but the validation loss drops and then slightly increases in the last epoch. This suggests that the model may overfit to the training data in the final epoch. It yielded a test micro-F_1 score of 0.92 macro-F_1 score of 0.57.
For the hypertension BioDischargeSummaryBERT classifier, we observed that both training loss drops as expected and while the validation loss varies, though ultimately ending lower than the initial validation loss. The heightened validation loss compared to the training loss perhaps suggests impaired generalizability of the model. It yielded a test micro-F_1 score of 0.86 macro-F_1 score of 0.46.





- **BioDischargeSummaryBERT with class weights**

For the diabetes BioDischargeSummaryBERT classifier with class weights, training and validation loss similarly drop over time, suggesting that the model generalizes. It yielded a test micro-F_1 score of 0.93 macro-F_1 score of 0.56.
For the hypertension BioDischargeSummaryBERT classifier with class weights, we observed the same phenomena as for the diabetes BioDischargeSummaryBERT classifier with class weights. It yielded a test micro-F_1 score of 0.83 macro-F_1 score of 0.53.
The diabetes BioDischargeSummaryBERT diabetes classifier with class weights had higher micro/macro-F_1 scores than the hypertension BioDischargeSummaryBERT diabetes classifier with class weights. Moreover, its validation/train loss suggests its training did not cause overfitting. Hence, we decided to train the hypertension BioDischargeSummaryBERT classifier with class weights on all 145,915 samples.

For the full dataset on the diabetes BioDischargeSummaryBERT classifier with class weights, we observed better results, in particular the macro-F_1 score than on the truncated dataset. It yielded a test micro-F_1 score of 0.909 and a macro-F_1 score of 0.683, which suggests better generalization and has a better macro-F_1 score than the score from our reference paper (Yao et al. (2019)). From these results, we believe that with more epochs, the model could perform even better and will produce similar results when applied to other comorbidities.

6 Analysis

For any given model, we see both a higher F_1 macro-score and F_1 micro-score for the diabetes classifier compared to the hypertension classifier. This is an initially surprising result, as the frequency of hypertension is close to double that of diabetes in the dataset, suggesting the hypertension classifiers would be less susceptible to shortcomings relating to class imbalance. This finding may be explained by the fact that hypertension is a common symptom, while diabetes is a common disease. Hypertension is associated with multiple diseases and disease terminologies and may be described alongside diverse conditions in each note, while diabetes has a more controlled terminology associated with it. This controlled diabetes terminology may be better learned by the BERT models than the decentralized hypertension-associated terminology. The BERT models may also need more time to learn the relatively complex features that indicate hypertension and may be captured with more training.

The linear regression classifier with BoW features severely overfitted to the training data and had poor micro/macro-F_1 scores. Moreover, it was very computationally inefficient, requiring the amount of training data to be significantly scaled down to allow for count-vectorization without exhausting RAM, demonstrating that it is computationally expensive and intractable for long complex medical vocabularies. Thus, the BERT models demonstrate a feasible and better-performing approach.

The BERT and BioDischargeSummaryBERT models performed very similar with regard to micro/macro-F_1 scores. However, the BioDischargeSummaryBERT model validation loss curves were more irregular and suggested less generalizability of the trained model. This may be due to the BioDischargeSummaryBERT being trained specifically trained on ICU discharge summaries from MIMIC-III, while our dataset contains only general hospital and ED discharge notes from MIMIC-IV. The ICU, made up of only in-patients who stay in the hospital, is a very different environment from the ED and general hospital, which have mostly out-patients who temporarily visit from outside the hospital. Therefore, the structure of notes on these in-patient patients may be different than those from our out-patient notes. Hence, our training may be more robust and generalizable from base BERT compared to BioDischargeSummaryBERT, which further trains base BERT on ICU discharge notes.

The BioDischargeSummaryBERT classifiers with class weights train much more robustly and are more generalizable compared to their BioDischargeSummaryBERT classifier counterparts without class weights. This may be due to the class weighting of the loss function helping account for issues relating to class imbalance in the dataset. Since the positive examples of a comorbodity make a larger contribution to the loss function, classifying them correctly is incentivized for the model. Hence, useful features relating to the desired classification are more likely to be learned, leading to better generalizability.

The BioDischargeSummaryBERT classifier with the full dataset with weights on AWS trained better than the BioDischargeSummaryBERT classifier counterparts with the truncated dataset. While the test micro-F_1 score on the truncated dataset was marginally worse, we saw big gains on the macro-F_1 score on the full dataset. We see better generalizability having been trained on the full dataset, with the model identifying far more positive samples than the model trained on the truncated dataset.

7 Conclusion

We created a successful NLP deep-learning model to predict diseases associated with obesity from clinical notes, namely diabetes. Our BioDischargeSummaryBERT classifier trained on the full dataset with class weights gave our micro-F_1 score of 0.909 which is comparable to our reference and macro-F_1 score of 0.689 which is better than our reference with room to improve. We evaluated the performance of models including Logistic Regression with BoW, BERT, BioDischargeSummary-BERT, and BioDischargeSummaryBERT with class weights to combat class imbalance. We found that the primary baseline Logistic Regression classifier was computationally intractable and performed

poorly on the lengthy, complex medical vocabulary. The BERT-based models are more sustainable and better-performing. The secondary baseline BERT model and BioDischargeSummaryBERT model had similar performance, but BioDischargeSummaryBERT generalized worse and performance on the validation dataset was more variable. This may have been due to the over-specialization of the BioDischargeSummaryBERT model to specifically in-patient ICU patients as opposed to our out-patient hospital and ED patients. However, this effect may be rescued by class-weighting the cross-entropy loss function, as the BioDischargeSummaryBERT model with class weights shows robust training with generalizability and validation loss that decreases alongside the training loss. This robustness was validated by the success of our BioDischargeSummaryBERT classifier with class weights trained on the full dataset, though there is room for improvement with increased epochs and hyperparameter tuning. In the future, we would like to run the BERT classifier with class weights over the full dataset and compare it against the BioDischargeSummaryBERT classifier with class weights. This may provide us more insight on potential over-specialization after further refinement. Furthermore, we would also like to compare these models against multi-label BERT classifiers. These may be able to glean further insights regarding these comorbidities that coincide with varying probabilities.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2021. Mimic-iv.

Siddhartha Nuthakki, Sunil Neela, Judy W Gichoya, and Saptarshi Purkayastha. 2019. Natural language processing of mimic-iii clinical notes for identifying diagnosis and procedures with neural networks. *arXiv preprint arXiv:1912.12397*.

Mostafizur Rahman and Darryl N. Davis. 2013. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3:224.

AK Singh, Mounika Guntu, Ananth Reddy Bhimireddy, Judy W Gichoya, and Saptarshi Purkayastha. 2020. Multi-label natural language processing to identify diagnosis and procedure codes from mimic-iii inpatient notes. *arXiv preprint arXiv:2003.07507*.

Illés Solt, Domonkos Tikk, Viktor Gál, and Zsolt T. Kardkovács. 2009. Semantic Classification of Diseases in Discharge Summaries Using a Context-aware Rule-based Classifier. *Journal of the American Medical Informatics Association*, 16(4):580–584.

Özlem Uzuner. 2009. Recognizing Obesity and Comorbidities in Sparse Data. *Journal of the American Medical Informatics Association*, 16(4):561–570.

Jonathan Waring, Charlotta Lindvall, and Renato Umeton. 2020. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial intelligence in medicine*, 104:101822.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*, 19(3):71.