Enhancing Healthcare Efficiency: Automated Abdominal Pain ICD-10 Code Classification via Automatic/Deep Classifiers

Vishnu N. Nair, MSE Data Science
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Abstract
ICD-10 codes are alphanumeric designations used by healthcare professionals to categorize and standardize diagnoses for billing purposes. Healthcare professionals are faced with the burdensome task of manually assigning ICD-10 codes to clinical notes from patient visits. This is particularly evident in the case of abdominal pain, which constitutes 9.6 million annual Emergency Department visits in the US[1]. This manual coding process is time-consuming, error-prone, and diverts providers from direct patient care. In response, we propose the implementation of automatic and deep learning classifiers for the efficient and accurate classification of abdominal pain ICD-10 codes. By automating this process, this study aims to significantly reduce the burden for healthcare providers, allowing for greater focused approaches for patient care. Through this, we hope to enhance accuracy and ultimately optimize the quality of care provided to the many patients that are hospitalized due to abdominal pain.

Introduction

As stated previously, the International Classification of Diseases, also referred to as ICD-10, is a widely used and globally accepted system to classify diseases and health conditions for patients. This system was put into place by the World Health Organization (WHO) and uses a mix of letters and numbers to identify various medical diagnoses and ailments. These codes are primarily used for insurance purposes and allow for healthcare administrators to communicate efficiently with one another to foster the creation of accurate billing processes. For this reason, ICD-10 codes are quite integral to the operation of the healthcare system across the world.

During a patient visit, certain healthcare providers (such as nurses, scribes, or physicians themselves), make notes of the patients symptoms, conditions, history of present illness, etc. They form a comprehensive outline of the patients ailments and procedures conducted, including any imaging reports, medical records, bloodwork, family history, and clinical notes in general. After receiving clinical assessment by a physician, medical coders spend time looking over the notes and related information, thereby designating the appropriate code. This is a time-consuming and laborious process. In a study done in 2014, coding times were quantified. They found that the average coding time using ICD-9 standards was 25.52 min for a singular patient and for ICD-10, 43.23 minutes[2]. This suggests that medical coders now need to spend an increased amount of time (69% to be exact) to place codes on a singular patient visit[2]. Additionally, it was found in another study that ICD-10 coding accuracy done by humans as of 2021 is only 78.7%[3]. As we can see, there is room for improvement in efficiency as well as accuracy for this coding system, especially with the newly designed ICD-10 system that is taking coders strenuous amounts of time.

In this study, we focus on the topic of abdominal pain due to the sheer number of cases that the USA has annually. In 2012, abdominal pain was the cause for 15.9 million visits[4]. Gastroesophageal reflux, also a cause of abdominal pain, was the cause for 8.9 million visits that same year[4]. Some of the most common reasons for abdominal pain are: nonspecific abdominal pain, acute appendicitis, biliary disease, bowel obstruction, acute diverticulitis, acute pancreatitis, constipation etc. As medical coders navigate through the extensive volume of abdominal pain cases that can be caused by a plethora of reasons, this study seeks to shed light on the challenges posed by this prevalent symptom, aiming to explore potential avenues for improvement in the accuracy and efficiency of ICD-10 code assignment.

Solution

In this study, we aim to reduce the burden on healthcare professionals to manually review clinical notes in aims of assigning ICD-10 codes. Specifically, we aim to reduce the time burden while bolstering the accuracy of said coding for abdominal pain ailments. We propose the implementation of state-of-the-art automatic classifiers and neural networks, leveraging the power of artificial intelligence in the realm of medical coding. The use of automatic classifiers, powered by advanced machine learning algorithms, holds the promise of significantly reducing the time-intensive nature of code assignment. These classifiers can be trained on vast datasets of clinical notes, learning intricate patterns and associations that might elude traditional coding methods. By automating the initial phase of code assignment, healthcare professionals can redirect their focus to more nuanced aspects of patient care, thereby optimizing their time and expertise. In our study, we employed a diverse set of machine learning and deep learning techniques to predict ICD-10 codes for abdominal pain cases. The selected methods encompass Linear Support

Vector Machines (SVMs), Logistic Regression, Naive Bayes, Convolutional Neural Networks (CNNs), and more generalized Neural Networks.

Linear SVMs operate by finding optimal hyperplanes that are best able to separate classes in a dataset. In our case, it would determine the most effective linear boundary between the embedded clinical notes to predict which ICD-10 code it corresponds to. The goal is to maximize the margin, the distance between the hyperplane and the nearest instances of each class, while minimizing classification errors. Logistic regressors are able to transform a linear combination of features into a probability score for predicting the likelihood of a particular ICD-10 code. A Naive Bayes model was employed as a baseline model essentially since it relies on the Bayes' theorem to calculate the probability of a given class. Both Naive Bayes and Logistic Regression models are simple and computationally efficient ones that we used to obtain initial baseline results. Finally, CNNs were applied to the text data. They operate by treating the text as a sequence of tokens and are able to extract patterns from the sequential data via identification of key phrases within clinical notes. Its ability to discern these complex relationships make it a robust approach for feature extraction and code prediction.

Implementing automatic classifiers and neural networks for predicting ICD-10 codes in abdominal pain cases presents a myriad of potential benefits for healthcare professionals and the broader healthcare system. As stated before, the adoption of these advanced technologies offers a substantial reduction in the time burden associated with manual code assignment. By automating this process using state-of-the-art classifiers, healthcare professionals and administrators can redirect their time and expertise to more nuanced aspects of healthcare. These models have the potential to rapidly assign medical codes while also doing it more accurately than a human worker may be able to do. Improved accuracy not only contributes to more precise representation of patients' conditions but also helps mitigate potential coding inaccuracies, fostering a higher standard of healthcare quality. Overall, the integration of machine learning classifiers have the potential to optimize resource allocation, streamline healthcare workflows, and enhance overall quality of patient care in regards to abdominal pain diagnoses.

There are several limitations to this solution, the first and foremost being the dependency on the quality and representativeness of the training data. The effectiveness of the automatic classifiers and neural networks heavily relies on the diversity and accuracy of the clinical notes used for training. Incomplete or biased datasets may result in models that fail to generalize well to the wide spectrum of abdominal pain cases encountered in real-world clinical settings. Additionally, the interpretability of the models poses a challenge, particularly with complex neural networks like CNNs. Furthermore, there will be challenges that arise in keeping the models up-to-date due to the dynamic, rapidly changing landscape of healthcare medical coding. Regular updates and continuous monitoring are essential to ensure that the automated system aligns with the latest medical standards. Addressing these limitations is crucial for the successful use of machine learning classifiers for ICD-10 abdominal pain code prediction.

Method

MIMIC-IV (Medical Information Mart for Intensive Care IV) is a thorough medical, publicly available dataset that is widely used in healthcare research studies. It was created by clinicians and researchers at the Massachusetts Institute of Technology (MIT). The dataset is hosted by the Laboratory for Computational Physiology at MIT and is open access for researchers across the world. The dataset contains de-identified EHR data from Beth Israel Deaconess Medical Center in Boston. Specifically, it contains a wide variety of information including medication records, diagnostic codes, clinical notes, patient demographics, lab results, bloodwork, and imaging information. From this, many researchers are able to develop machine learning models and conduct other studies using this data.

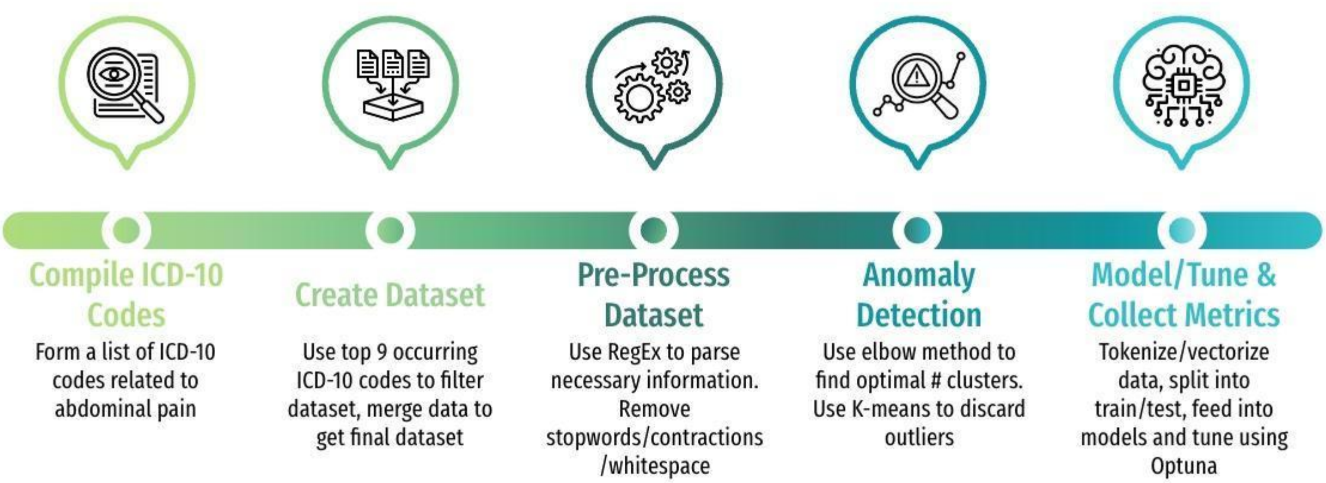
In order to integrate the predictive models for multi-label classification, a thorough data preparation process was conducted. Firstly, the "DISCHARGE", "DIAGNOSES_ICD", and "ADMISSIONS" were merged to create a comprehensive dataset. A thorough list of ICD-10 codes were then compiled that relate to abdominal pain. The dataset we created was then filtered on this list that was made, thereby tailoring said data to the scope of our study. Finally, we chose the top 9 classes (ICD-10) codes with the highest value counts for modeling. Recognizing the potential impact of class imbalance on model performance, a resampling technique was then employed. This was done by undersampling the majority classes in order to fix the imbalance, ensuring that our models are trained on a more equal representation of the abdominal pain classes. The codes chosen correspond to constipation, gastrointestinal hemorrhage, diverticulitis of large intestine, other intestinal disorders, diaphragmatic hernia, unspecified Crohn's disease, ischemic bowel disorders, stomach/duodenum disease, celiac disease, and anal fissure.

In the preprocessing phase of our study, regular expressions played a crucial role in extracting the desired information from the clinical notes at hand. We specifically chose to focus on sections including “History of Present Illness”, “Chief Complaint”, and other acute symptoms/conditions experienced by the patient detailed in their clinical notes. Using regular expressions, we were able to identify and extract crucial details relating to the patient’s abdominal pain condition. The corpus across many notes were analyzed closely; the unstructured and variant nature of the notes made it difficult to develop generalized expressions, but it worked in the end. Then, further regular expressions were utilized to remove special characters, whitespaces, stopwords, etc. Doing so, the corpus we worked with for each observation was significantly cleaned, making it model ready at last.

The use of regular expressions in this context ensured precision in extracting targeted information from the clinical notes, contributing to a more focused and accurate dataset for subsequent analysis. By customizing regular expressions to align with the specific patterns and structures within the medical narratives, this approach facilitated the creation of a more refined dataset, essential for the successful implementation of machine learning models in predicting abdominal pain ICD-10 codes. Prior to building machine learning models for this problem, anomaly detection was conducted using the K-means algorithm. The elbow method was employed in order to first find the optimal number of clusters for the dataset. This allowed for the selection of a meaningful number of clusters for anomaly detection. The K-means algorithm was then executed, using a 95 percentile threshold to discard any outliers. In other words, if a datapoint exceeds this threshold in terms of distance from the centroid of its cluster, it is discarded. This allowed for us to retain data points that fell closely within the distance of the main cluster densities. By systematically discarding outliers, the study aimed to enhance the models' ability to discern patterns and relationships within the abdominal pain dataset, ultimately leading to more robust and accurate predictions of ICD-10 codes.

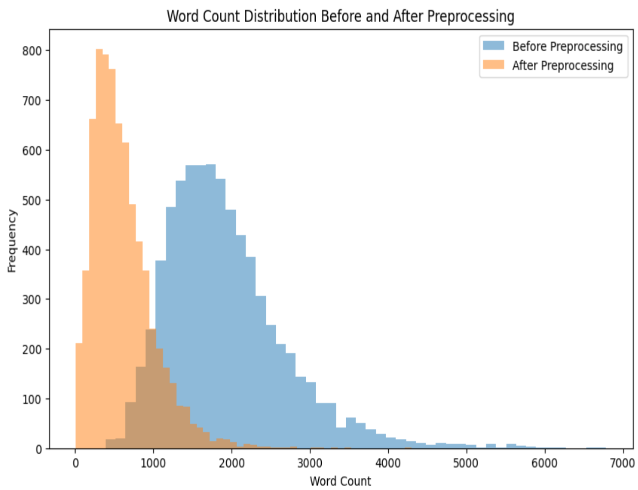
Finally, our dataset was ready for modeling. In order to develop the predictive algorithms for abdominal pain ICD-10 code assignments, we employed a diverse set of machine learning and deep learning models. Linear Support Vector Machines (SVM) and Logistic Regression in the scikit-learn library were chosen initially for their simplicity and efficiency in multi-label classification tasks. Next, stronger learners such as XGBoost (using scikit-learn) and CNN’s (using PyTorch) were built in order to handle potential complex relationships and deliver robust results. The CNN was built from scratch using PyTorch, employing dropout and MaxPooling layers in order to prevent overfitting due to the limited amount of data being worked with. To optimize the performance of these algorithms, hyperparameter tuning was conducted using Optuna, a hyperparameter optimization framework. Optuna uses Sequential Model-Based Optimization (SMBO) to model probabilistic distribution of the hyperparameters. By sampling different hyperparameter configurations, it is able to effectively find the optimal ones via updating a probabilistic model iteratively.

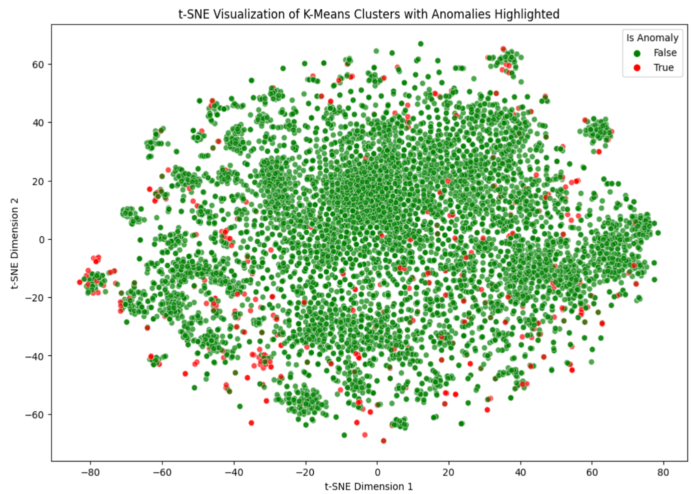
Workflow Diagram



Results/Discussion

The first step after completing data pre-processing was to visualize our cleaned data and conduct anomaly detection. In Figure 1, we see the effect that removing contracts, stopwords, whitespaces, and filtering for specific portions of each patient note corpus (history of present illness, chief complaint, acute conditions, etc.) have on our text data. It is safe to assume that our data pre-processing portion was successful and effective. Our anomaly detection algorithm removed 381 outliers from our dataset using 9 clusters for K-means, ensuring that noise is minimized during the modeling portion of our study. At this point, our model was





We collected accuracy, precision, and recall metrics for all models that we ran. Here is a table summarizing our results below:

Metric	Logistic Regression	Linear SVM	XGBoost	CNN
Precision	0.60	0.58	0.63	0.51
Recall	0.58	0.56	0.62	0.49
Accuracy	0.58	0.56	0.62	0.49

Table 1. Metric collection from default baseline models

Metric	Tuned Logistic Regression	Tuned Linear SVM	Tuned XGBoost	Tuned CNN
Precision	0.59	0.55	0.62	0.53
Recall	0.58	0.54	0.61	0.52
Accuracy	0.58	0.54	0.61	0.52

Table 2. Metric collection from tuned models

To dive deeper into the functionality regarding our two best models: XGBoost and Logistic Regression, the two heatmaps in Figure 1 illustrate the predictive power of these models across all 9 classes of abdominal pain.

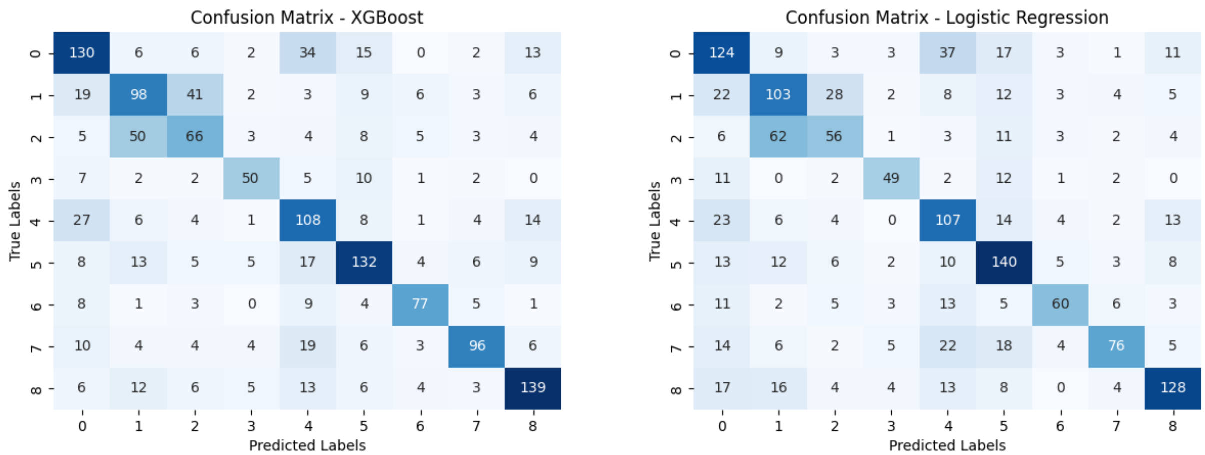


Figure 1. Heatmaps of model performance for best performing models: XGBoost and Logistic Regression

Overall, we see that all models are performing to an objectively mediocre standard, but decently well considering the fact that our dataset is quite limited in size. With only 9,000 observations, these models are predicting with around 60% accuracy across the board. While this level of accuracy might be considered acceptable, it's crucial to delve deeper into the precision and recall metrics to gain a comprehensive understanding of the models' behavior. Overall, the results obtained with the baseline model were somewhat expected aside from the CNN doing worse than all of the more simple models. In regards to the tuned models, we did not attain the expected results. It was expected that Optuna would find hyperparameters that would improve model performance, but the models did approximately the same or sometimes a bit worse after tuning. This may be attributed to the fact that we only ran 10 trials of Optuna training due to RAM/memory/time constraints for this project. Due to this, the Optuna framework did not have enough trials to probabilistically search through the hyperparameter space, explore diverse values, and discover optimal ones. The limited dataset size of 9,000 observations might have played a role in hindering the effectiveness of hyperparameter optimization as well. With a smaller dataset, the models might not be able to generalize well to unseen data, making it challenging for the optimization process to identify robust parameter configurations. Increasing the dataset size, as previously suggested, remains a potential avenue for improvement. Additionally, it is important to keep in mind that hyperparameters are often interconnected, and changes in one parameter can affect the optimal values of others as well. This intricate relationship might pose challenges for automated optimization processes like Optuna. Manually exploring interactions between hyperparameters would be an ideal future work.

Conclusion

Overall, the models, including the baseline and tuned versions, exhibit potential for abdominal pain ICD-10 code classification. These models present a viable solution for healthcare workers by offering a baseline level of accuracy. The use of automated classification models can significantly save time for healthcare professionals, enabling them to focus on more complex cases or critical tasks. Moreover, as the dataset size increases and hyperparameter tuning becomes more extensive, the models are likely to further enhance their accuracy and become even more valuable tools for efficient and reliable abdominal pain classification in the medical field, saving much time and manual burden for healthcare workers.

References

1. Wang L, Haberland C, Thurm C, Bhattacharya J, Park KT. Health outcomes in US children with abdominal pain at major emergency departments associated with race and socioeconomic Status. PLoS One. 2015 Aug 12;10(8)
2. Weems S, Heller P, Fenton SH. Results from the Veterans Health Administration ICD-10-CM/PCS Coding Pilot Study. Perspect Health Inf Manag. 2015 Summer;12(Summer):1b. PMCID: PMC4558479. PMID: 26396553

3. Olagundoye O, van Boven K, Daramola O, Njoku K, Omosun A. Improving the accuracy of ICD-10 coding of morbidity/mortality data through the introduction of an electronic diagnostic terminology tool at the general hospitals in Lagos, Nigeria. BMJ Open Qual. 2021 Mar;10(1):e000938
4. Fagerström A, Paajanen P, Saarelainen H, Ahonen-Siirtola M, Ukkonen M, Miettinen P, et al. Non-specific abdominal pain remains as the most common reason for acute abdomen: 26-year retrospective audit in one emergency unit. Scand J Gastroenterol. 2017;52(10):1072-1077.
5. Peery AF, Dellon ES, Lund J, Sandler RS, Kappelman MD, Shaheen NJ. Burden of gastrointestinal disease in the united states: 2012 Update. Gastroenterology. 2012;143(5):1179-1187.e3.