A Natural Language Processing-Driven Machine Learning Approach for Diabetes Phenotyping

Abstract

Diagnosing diabetes using automated approaches would aid in increasing patient care, disease management and assist clinicians. In this project supervised algorithms were used to develop classifiers developed to detect type 1 diabetes and type 2 diabetes from MIMIC-IV clinical notes using three different approaches by leveraging natural language processing techniques, TD-IDF, word embeddings and the use of pre-trained large language model.

Introduction

Diabetes is considered a global health concern affecting the lives of approximately four hundred and twenty million people worldwide. Despite the treatment options available, it's still responsible for one million and a half deaths annually[1]. In this report, a data-driven approach is used to aid in diagnosing diabetes by analyzing clinical notes utilizing natural language processing methodology.

Problem Statement

Diabetes mellitus is a term for several diseases that affect the body's ability to utilize sugar (glucose). There are different types of diabetes depending on the main cause of them[2]. Type 1 diabetes is a chronic disease characterized by the pancreas' deficiency in producing insulin. Insulin is a hormone that regulates the sugar (glucose) for energy production. Genetics, viruses, and autoimmune reactions are the leading causes of type 1 diabetes. There are no ways to prevent type 1 diabetes from occurring, but it can be managed through the use of insulin injections and leading a healthy lifestyle to prevent further complications[3]. The other chronic type is type 2 diabetes, which affects the cells' response to insulin which makes the pancreas produce more insulin. The problem is that over time, insulin resistance results in elevated blood sugar levels. Like type 1 diabetes, it's incurable and it's mostly managed through the use of diabetes medications, diet, and physical activity[4]. Diabetes could lead to severe complications such as cardiovascular disease, which includes heart and blood vessels problems such as heart failure and atherosclerosis. Diabetic patients are more prone to it than non-diabetic people and it's considered the leading cause of death in diabetic patients[5]. Another complication is diabetic ketoacidosis (DKA), which is a life-threatening condition that happens when the insulin levels are insufficient which causes the body to break down fat to use for energy resulting in a chemical called ketone. Ketoacidosis could result in diabetic coma or death[6]. Another common complication is neuropathy, which is characterized by nerve damage that may result in infection and amputation[7]. Blindness and hearing loss are more common in diabetes patients than other people[8]. In a report published by the Centers for Disease Control and Prevention (CDC)[9], In 2020, around 16.8 million diabetes-related visits to the emergency department (ED) were reported. The risk factors in these visits included hypoglycemia, which is a condition caused by the drop of blood sugar level below the normal range[10], diabetic ketoacidosis, and hyperosmolar hyperglycemic syndrome, which is an acute complication of diabetes that happens when the blood sugar levels are elevated for a long period of time. 38.4% of the visits were admitted to the hospital and 7.86 million diabetes-related hospital discharges were reported. The leading causes of hospitalization were cardiovascular diseases such as ischemic heart disease and strokes. Moreover, in 2019, around 61,422 people had end-stage kidney disease caused by diabetes. In 2021, the eighth leading cause of death was diabetes. There are risk factors that contribute to diabetes development such as smoking, obesity, physical inactivity, high cholesterol, and high blood pressure. In 2022, The costs of diabetes in the United States were estimated to be around 413 billion dollars. There was an increase in costs from 277 billion dollars in 2012 to 307 billion dollars in 2022 with an average cost of 12,022 dollars per person associated with diabetes. The main way to diagnose diabetes is by a blood test. Despite that, there were 8.7 million people in the United States who were unaware that they were considered diabetic in terms of laboratory ranges and never reported that they had diabetes. Despite medical advances, it was estimated that there was a 3% increase in diabetes prevalence from 2001 to 2020.

Solution

Diagnosis of diabetes in its early stages could prevent fatal consequences. The aim of this project was to develop classification models that can detect the phenotypes of chronic diabetes, Type 1 diabetes, and Type 2 diabetes from clinical notes. There have been automated approaches for diabetes phenotyping by using electronic health records and clinical notes. It mainly relied on rule-based algorithms. The proposed solution for this project was to develop machine learning classifiers by leveraging natural language processing techniques that converted the clinical notes into an appropriate form to be used as inputs. There have been three approaches followed in order to compare between the different natural language processing techniques. The benefits of using the classifiers include as mentioned earlier, early detection of diabetes which would help the patient manage the disease in its early stages. Also, it would aid clinicians in managing the time and effort needed to diagnose and provide better care for the patients. Moreover, the use of such data-driven techniques could help identify hidden patterns in the diabetes patient population. While such approaches are beneficial, the requirements to develop them are challenging. In order to train the classifiers, they need datasets that were labeled. Due to the project scale, the dataset was labeled by following an algorithm rather than having clinicians annotate and label each clinical note in order to ensure the accuracy of the labels. Hence, the reliability of the classifiers developed might not be as good as if a clinician labeled the notes. Moreover, not all techniques used in the project are interpretable which is a common problem in the machine learning field. Also, to ensure the generalizability of the classifiers, a large amount of data must be used which means that preprocessing such datasets and training the classifiers would require a long time and computational resources. Lastly, while there are two main types of chronic diabetes, several studies suggested other phenotype definitions according to the root causes. There have been 5 phenotypes identified which are, autoimmune phenotype, insulin-related phenotypes which include insulin deficiency and resistance, obesity phenotype, ageing phenotype, and a sex-related phenotype[11]. Since there has been no algorithm developed to identify these types due to their complexity, the main two types of chronic diabetes were selected.

Methods

In **Figure 1**, the flowchart of the developing diabetes phenotyping is illustrated. There were two main parts of the process, the first was patient labeling since the data wasn't labeled. The second part was developing the classifiers.

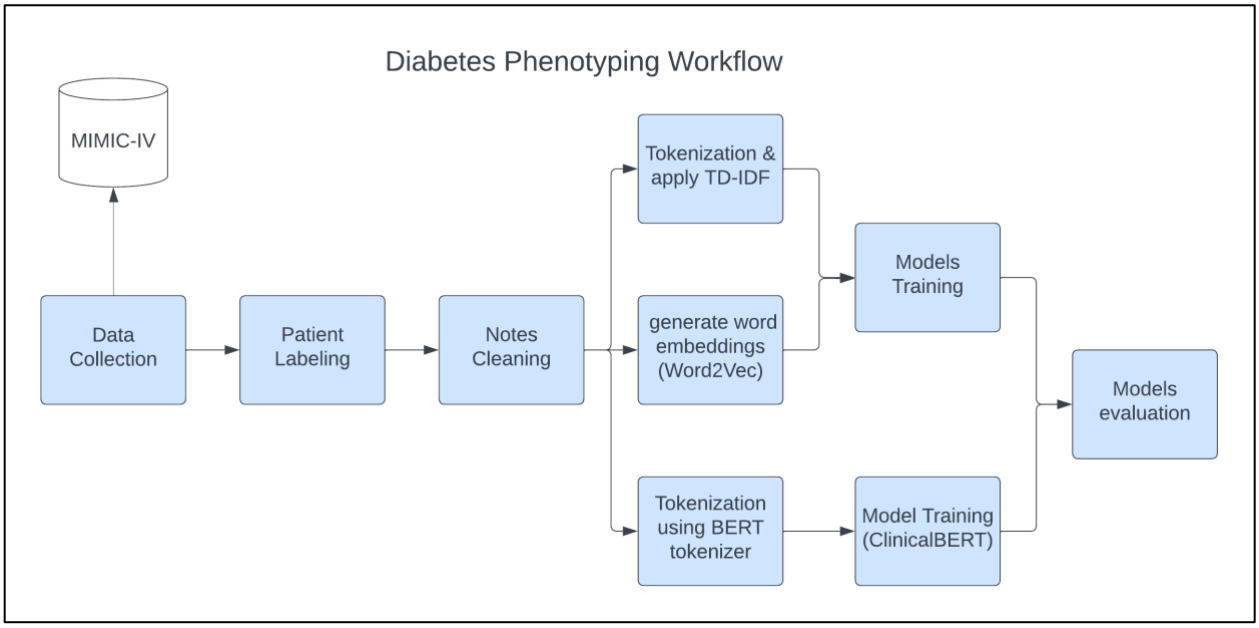

Figure 1. Workflow diagram for diabetes phenotyping using MIMIC-IV database.

Patients Labeling

The database that was used to develop the classifiers is the Medical Information Mart for Intensive Care version 4 (MIMIC-IV) database, which is a database that contains the de-identified electronic health records of 299,712 patients who were admitted to the ICU or emergency department of Beth Israel Deaconess Medical Center in Massachusetts

between 2008 – 2019. The database consists of different tables such as patients, admissions, medications, and discharge summaries[12,13]. The project consists of two main parts, the first is the labeling of the clinical notes for diabetes since classifiers are supervised machine learning models that require labeled datasets. In order to label the clinical notes, the patients with type 1 diabetes and type 2 diabetes were identified by leveraging a diabetes phenotype algorithm pseudo code developed by the Department of Medicine at Stanford University, which was published on the Phenotype KnowledgeBase website[14] to produce silver standard labels, a silver standard label is a label that was annotated by automated approaches, unlike gold standard labels which are labels annotated and validated by clinicians. The first step was to identify diabetes ICD codes available in the MIMIC-IV database by querying the d_icd_diagnoses table for diabetes ICD codes which are 250.x, E10.x, E11.x. After that, the diagnoses_icd table which contains the patients, and their diagnoses were filtered based on the extracted ICD codes. The second step was to check if these patients were prescribed any diabetes-related medications. the medications were extracted according to Upadhyaya et al list[15] which includes generic and brand names of diabetes medications in the US. After that the patients were filtered again based on whether they had prescriptions or not in the prescriptions table. if a prescription was not found, then the lab tests table was queried for the following tests: hemoglobin A1c and glucose as these lab tests are relevant to diabetes. After that, the patients were filtered based on the criteria stated in the Stanford algorithm using the labevents table, a glucose test is considered abnormal if it's equal to or more than 125 mg/dl. It was assumed that all glucose tests are for fasting glucose since the mentioned normal range for those tests is 70mg/dl to 100mg/dl. As for hemoglobin A1c, it's considered abnormal if the test result is equal to or more than 6.5%. by the end of this step, the diabetes patients were identified. The last step was to determine which phenotype they had by starting with how many diagnoses/ICD codes each patient had for type 1 diabetes and type 2 diabetes. type 1 diabetes ICD codes are 250.x1, 250.x3 or E10.x. As for type 2 diabetes ICD codes are 250.x0, 250.x2 or E11.x[16]. After counting the frequencies, if the patient ratio of type 1 diabetes to type 2 diabetes codes was more than 0.5 and there was a prescription for glucagon or there was no record of oral hypoglycemic medication other than metformin, the patient would be labeled as a type 1 diabetes patient. Other than that, the patient is labeled as a type 2 diabetes patient. After applying the algorithm, 2344 patients were labeled as type 1 diabetes patients and 34060 patients were labeled as type 2 diabetes patients. After that, the notes from the discharge notes table were filtered using the patients' IDs and labeled accordingly. In order to get the relevant clinical notes. The notes with the following words were selected: "diabetes, diabetic, insulin, sugar, A1C, glucose hyperglycemia, hypoglycemia, euglycemia, diabetic, diabet, fasting, hypoglycem, hypoglycemic, pancreas". After that, 8514 notes were identified as type 1 diabetes patients' clinical notes, and 94873 notes were identified as type 2 patients' clinical notes. 7000 notes from each phenotype were selected to have a balanced dataset of 14000 clinical notes. A subset of the dataset (7000 notes) was used instead of 14000 notes in the third approach due to the complexity of the approach as it required time and computational power to train.

Classification

There were 3 different approaches followed to develop the classifiers. All of them started by cleaning the notes by removing punctuations, numbers, special characters, new lines, stop words and lemmatization. After that the labeled notes were split into a training set containing 80% of the notes and a testing set containing 20% of the notes in order to be used in the modeling phase.

- TD-IDF approach: the first approach for classification was done by tokenizing the notes and applying the term frequency-inverse document frequency (TF-IDF) from the sklearn package, which measures how relevant and important a word in a document is. Three classification algorithms were used which are Naive Bayes, support vector machine (SVM) and logistic regression. The classifiers were evaluated using accuracy, F1 score, recall, and precision for training and testing sets. Moreover, the top 15 ranked positive coefficients for each phenotype were extracted from the logistic regression classifier.
- Word embeddings approach: the second approach was done by generating sentence-level vectorized word embeddings from the notes which is a way to represent words' semantic relationships by using Word2Vec algorithm from gensim package, an algorithm developed by Thomas et al. from Google[17]. After that, the training and testing sets were scaled using min-max scalar which scales the values between 0 and 1 since the inputs to algorithms such as Naïve Bayes must be positive values. The previous algorithms were used to construct the classifiers and were evaluated on training and testing sets.
- ClinicalBERT approach: the third approach which is considered the most complex approach was done by tokenizing the notes using a tokenizer that is tailored for use with transformers models from Hugging Face transformers package then padding the dataset since the use of pretrained large language models (LLMs) requires a fixed input length. After that, the tokenized dataset was trained using a pretrained model called "tiny-

clinicalbert", which is a distilled version of the BioClinicalBERT, an LLM model that was trained on all notes from MIMIC III[18]. Finally, the model was evaluated on the training and testing sets using the same metrics that were previously used.

Results

Evaluation

In **Table 1**, the first approach which used TD-IDF to preprocess the notes was evaluated for both phenotypes. Overall, the classifiers' results were better on the training set than testing set but generally, the testing set results are considered promising. The optimal model for this approach was logistic regression.

Table 1. TD-IDF approach evaluation results

Phenotype	**Training set**			**Testing set**		
Naïve Bayes	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
Type 1 diabetes	0.87	0.81	0.84	0.82	0.75	0.79
Type 2 diabetes	0.83	0.88	0.85	0.77	0.83	0.80
Support Vector Machine	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
Type 1 diabetes	0.90	0.81	0.85	0.87	0.76	0.81
Type 2 diabetes	0.83	0.91	0.87	0.78	0.89	0.83
Logistic Regression	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
Type 1 diabetes	0.94	0.92	0.93	0.89	0.85	0.87
Type 2 diabetes	0.92	0.94	0.93	0.86	0.89	0.88

In **Table 2**, the second approach which used word embeddings was evaluated. Naive Bayes's performance was approximately 10% lower than the previous approach. As for SVM and logistic regression, the performance was slightly lower than the previous one. Overall, the optimal model for this approach was logistic regression.

Table 2. Word embeddings approach evaluation results

Phenotype	**Training set**			**Testing set**		
Naïve Bayes	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
Type 1 diabetes	0.71	0.70	0.71	0.70	0.70	0.70
Type 2 diabetes	0.71	0.71	0.71	0.69	0.70	0.69
Support Vector Machine (SVM)	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
Type 1 diabetes	0.87	0.83	0.85	0.86	0.82	0.84
Type 2 diabetes	0.84	0.87	0.86	0.82	0.86	0.84
Logistic Regression	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
Type 1 diabetes	0.88	0.85	0.86	0.87	0.84	0.85
Type 2 diabetes	0.85	0.88	0.87	0.84	0.87	0.85

In Table 3, pretrained ClinicalBERT was trained and evaluated. This approach's results were better than the word embeddings approach and slightly lower than the TD-IDF approach.

Table 3. ClinicalBERT approach evaluation results

Phenotype	**Training set**			**Testing set**		
Transformers (tiny-clinicalbert)	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
Type 1 diabetes	0.90	0.87	0.88	0.84	0.83	0.84
Type 2 diabetes	0.88	0.90	0.89	0.84	0.85	0.84

Feature Importance

In **Figure 2**, the top features contributing to the classification of type 1 diabetes were reported. Words like transplant can be seen, this could be related to post-transplant diabetes mellitus or pancreas transplant which is a common surgery to treat diabetes. dka (diabetic ketoacidosis) and gastroparesis are diabetes complications. In **Figure 3**, the top features contributing to the classification of type 2 diabetes were reported. Most of the words are medications generic and brand names.

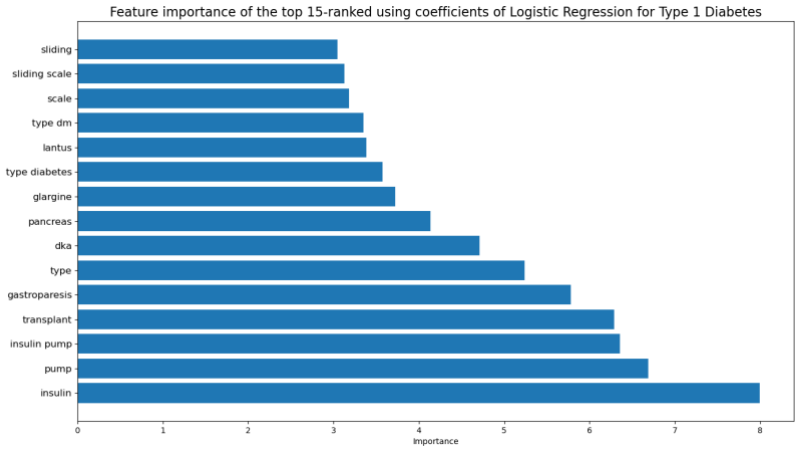

Figure 2. Feature importance of the top 15-ranked using coefficients of TD-IDF-Logistic Regression for Type 1 Diabetes

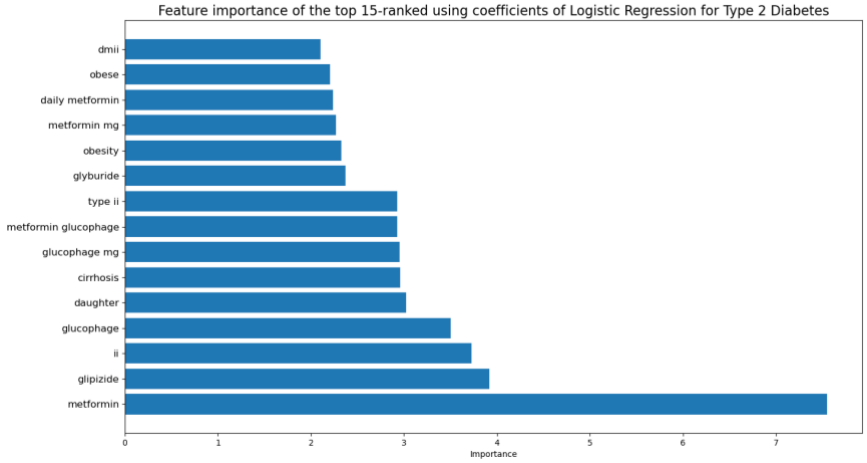

Figure 3. Feature importance of the top 15-ranked using coefficients of TD-IDF-Logistic Regression for Type 2 Diabetes

Conclusion

A textual data driven machine learning approach to detect chronic diabetes was developed leveraging natural language processing techniques. The dataset was initially labeled using a pseudo-code of a phenotyping algorithm and three different approaches were followed to develop the classifiers, term frequency-inverse document frequency, word embeddings and pretrained large language models. This project could be useful for doing further study into hidden patterns in the diabetes patient's population and developing interpretable models to understand the context behind automated diabetes detection.

References

1. WHO. Diabetes [Internet]. Who.int. World Health Organization: WHO; 2019. Available from: https://www.who.int/health-topics/diabetes
2. Mayo Clinic. Diabetes - symptoms and causes [Internet]. Mayo Clinic. Mayo Clinic; 2023. Available from: https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444
3. CDC. What Is Type 1 Diabetes? [Internet]. Centers for Disease Control and Prevention. CDC; 2022. Available from: https://www.cdc.gov/diabetes/basics/what-is-type-1-diabetes.html
4. Centers for Disease Control and Prevention. Type 2 Diabetes [Internet]. Centers for Disease Control and Prevention. 2023. Available from: https://www.cdc.gov/diabetes/basics/type2.html
5. Cardiovascular Disease | ADA [Internet]. diabetes.org. Available from: https://diabetes.org/about-diabetes/complications/cardiovascular-disease
6. Diabetes & DKA (Ketoacidosis) | ADA [Internet]. diabetes.org. Available from: https://diabetes.org/about-diabetes/complications/ketoacidosis-dka/dka-ketoacidosis-ketones
7. Understanding Neuropathy and Your Diabetes | ADA [Internet]. diabetes.org. Available from: https://diabetes.org/about-diabetes/complications/neuropathy
8. American Diabetes Association. Diabetes Complications | ADA [Internet]. diabetes.org. Available from: https://diabetes.org/about-diabetes/complications
9. CDC. National diabetes statistics report [Internet]. CDC. 2022. Available from: https://www.cdc.gov/diabetes/data/statistics-report/index.html
10. Mayo Clinic. Hypoglycemia - Symptoms and causes [Internet]. Mayo Clinic. 2020. Available from: https://www.mayoclinic.org/diseases-conditions/hypoglycemia/symptoms-causes/syc-20373685
11. 1.Gouda P, Zheng S, Peters T, Fudim M, Randhawa VK, Ezekowitz J, et al. Clinical Phenotypes in Patients With Type 2 Diabetes Mellitus: Characteristics, Cardiovascular Outcomes and Treatment Strategies. Current Heart Failure Reports. 2021 Aug 24;18(5):253–63.
12. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi L A, Mark R. MIMIC-IV (version 2.2). PhysioNet. 2023. Available from: https://doi.org/10.13026/6mm1-ek67.
13. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023 Jan 3;10(1).
14. Stanford University. Type 1 and type 2 Diabetes Mellitus | PheKB [Internet]. phekb.org. [cited 2023 Dec 14]. Available from: https://phekb.org/phenotype/1506
15. Upadhyaya SG, Murphree DH, Ngufor CG, Knight AM, Cronk DJ, Cima RR, et al. Automated Diabetes Case Identification Using Electronic Health Record Data at a Tertiary Care Facility. Mayo Clinic Proceedings: Innovations, Quality & Outcomes. 2017 Jul;1(1):100–10.
16. Zhang W, Cao T. Automated Type 2 Diabetes Case and Control Identification from the MIMIC-IV Database. AMIA Joint Summits on Translational Science proceedings AMIA Joint Summits on Translational Science [Internet]. 2023 [cited 2023 Dec 14];2023:602–11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10283086/
17. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space [Internet]. arXiv.org. 2013. Available from: https://arxiv.org/abs/1301.3781
18. Rohanian O, Nouriborji M, Jauncey H, Kouchaki S, Group ICC, Clifton L, et al. Lightweight Transformers for Clinical Natural Language Processing [Internet]. arXiv.org. 2023 [cited 2023 Dec 14]. Available from: https://arxiv.org/abs/2302.04725