Publicly Available Datasets for Diabetes Classification Modeling

1. Introduction: The Significance of Datasets in Diabetes Classification Modeling

Diabetes Mellitus stands as a significant global health challenge, imposing substantial financial strains on economies and profoundly affecting the well-being of millions worldwide.[1] In the United States alone, the prevalence of this chronic disease impacts a vast segment of the population annually, underscoring the critical need for effective management and early intervention.[1] The increasing application of machine learning methodologies in medical health, particularly for the prediction of diabetes, reflects a growing reliance on data-driven solutions to address this pervasive condition.[2] Early and accurate diagnosis is paramount in mitigating the long-term complications associated with diabetes, leading to improved patient outcomes and reduced healthcare costs.[1]

Machine learning approaches have emerged as powerful tools for achieving rapid and reliable detection of diabetes, offering the potential for enhanced diagnostic accuracy, improved cost-effectiveness, and more efficacious treatment strategies.[3] The application of machine learning-based prediction models holds promise in facilitating the early identification of individuals at risk, even before the condition significantly deteriorates.[4] Various sophisticated algorithms are now being employed to develop artificial intelligence models capable of forecasting diabetes, signifying a transformative shift towards proactive healthcare management.[4] The timely detection afforded by these methods enables healthcare professionals to implement appropriate medication regimens, dietary adjustments, and exercise plans, ultimately contributing to a healthier life for individuals at risk.[3]

However, the effectiveness and reliability of machine learning models are intrinsically linked to the quality and characteristics of the datasets used for their training and evaluation. Existing machine learning approaches have sometimes been constrained by the utilization of limited datasets, resulting in a lack of generalizability and suboptimal accuracy.[3] The automation of diabetes diagnosis and the accurate assessment of its severity through machine learning and deep learning techniques necessitate the availability of large and comprehensive datasets to ensure the robustness and broad applicability of the resulting models.[3] The Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights (AI-READI) project was conceived to address the recognized gap in well-designed, high-quality, large, and inclusive multimodal datasets, highlighting the critical importance of robust data resources in advancing the field.[5] The limitations of current models due to data scarcity and lack of diversity underscore the vital need for researchers and practitioners to have access to a wide array of datasets to further refine and enhance diabetes classification methodologies.

2. Structured Datasets for Diabetes Prediction

2.1. Overview of Popular Platforms

Kaggle stands as a prominent platform within the data science community, widely recognized for hosting data science competitions and providing a rich repository of publicly available datasets. Numerous resources [1] confirm Kaggle as a central hub for a diverse collection of datasets, including a significant number specifically related to diabetes. This platform's accessibility and the vibrant community support it fosters make it an invaluable resource for individuals seeking data for machine learning projects focused on diabetes prediction. The sheer volume of diabetes-related datasets available on Kaggle underscores its importance as a primary source for both researchers and practitioners in this domain.

The **UCI Machine Learning Repository** represents another well-established and respected source for machine learning datasets, frequently utilized in academic research. Several sources [6] mention the availability of diabetes-related datasets within the UCI repository. This repository's emphasis on academic datasets implies a level of curation and reliability that can be particularly beneficial for rigorous model development and scholarly investigations. Its long-standing presence in the field also signifies its historical importance in providing foundational data resources for machine learning research, including those focused on diabetes.

2.2. Detailed Examination of Key Structured Datasets

2.2.1. Pima Indians Diabetes Database

The Pima Indians Diabetes Database is accessible through both Kaggle [3] and the UCI Machine Learning Repository.[6] Originating from the National Institute of Diabetes and Digestive and Kidney Diseases, this dataset was designed to predict whether a patient has diabetes based on a set of diagnostic measurements.[10] The data pertains to a specific demographic group: female patients of Pima Indian heritage, who were at least 21 years old and resided near Phoenix, Arizona.[24]

The dataset comprises several numeric-valued attributes, including the number of times a patient has been pregnant (Pregnancies), their plasma glucose concentration at two hours in an oral glucose tolerance test (Glucose), diastolic blood pressure (BloodPressure), triceps skin fold thickness (SkinThickness), two-hour serum insulin level (Insulin), body mass index (BMI), diabetes pedigree function (DiabetesPedigreeFunction), and age (Age).[24] The target variable, denoted as 'Outcome', is binary, with values of 0 indicating no diabetes and 1 indicating the presence of diabetes.[24]

The Pima Indians Diabetes Database is extensively used in diabetes prediction research and serves as a common benchmark for evaluating the performance of new machine learning algorithms.[4] Its widespread adoption has even led to research specifically addressing inherent characteristics of the dataset, such as the class imbalance between diabetic and non-diabetic individuals.[25] The enduring popularity of this database has established it as a fundamental resource in the field, making it an excellent starting point for researchers and practitioners. However, its focus on a specific demographic might limit the generalizability of models trained on it to broader populations, and the known issue of class imbalance requires careful consideration during model development to avoid biased outcomes.

2.2.2. Diabetes Health Indicators Dataset (BRFSS 2015)

The Diabetes Health Indicators Dataset, specifically the 2015 version, is available on both Kaggle [1] and the UCI Machine Learning Repository.[20] This dataset is derived from the Centers for Disease Control and Prevention's (CDC) Behavioral Risk Factor Surveillance System (BRFSS) survey.[1] The BRFSS is an annual, nationwide health-related telephone survey that collects data on health-related risk behaviors, chronic health conditions, and the utilization of preventive health services from a large sample of over 400,000 American adults each year.[1] The 2015 version of this dataset, as hosted on Kaggle, contains responses from 253,680 individuals and includes 21 feature variables.[1]

The data fields in this dataset encompass a range of health indicators and lifestyle factors, including Diabetes_012 (with categories for no diabetes, prediabetes, and diabetes), HighBP (high blood pressure), HighChol (high cholesterol), CholCheck (cholesterol check within the past five years), BMI (Body Mass Index), Smoker (history of smoking at least 100 cigarettes), Stroke (ever told they had a stroke), HeartDiseaseorAttack (coronary heart disease or myocardial infarction), PhysActivity (physical activity in the past 30 days), Fruits (consume fruit one or more times per day), Veggies (consume vegetables one or more times per day), HvyAlcoholConsump (heavy alcohol consumption), AnyHealthcare (any kind of healthcare coverage), NoDocbcCost (could not see a doctor due to cost), GenHlth (general health status), MentHlth (number of days with poor mental health), PhysHlth (number of days with poor physical health), DiffWalk (difficulty walking or climbing stairs), Sex, Age, Education, and Income.[1] The target variable can be approached in two ways depending on the specific analysis: as a three-class variable (Diabetes_012) or as a binary variable (Diabetes_binary), where 1 indicates either prediabetes or diabetes.[1]

This dataset is commonly used for predicting diabetes risk based on a comprehensive set of health indicators and lifestyle behaviors captured through the BRFSS survey.[1] The BRFSS dataset offers a significantly larger and more diverse sample compared to the Pima Indians Diabetes Database, potentially leading to the development of models with greater generalizability. The inclusion of behavioral and socioeconomic factors alongside traditional health metrics provides a richer context for prediction. The flexibility of having both a three-class and a binary target variable allows for different modeling strategies and research questions to be addressed.

2.2.3. Diabetes Prediction Dataset

The Diabetes Prediction Dataset is available on Kaggle [6] and is presented as a comprehensive resource for predicting diabetes using a combination of medical and demographic data.[6] This dataset includes information on key factors such as age, gender, body mass index (BMI), hypertension, presence of heart disease, smoking history, HbA1c level, and blood glucose level.[6]

The specific data fields within this dataset are Age, Gender, BMI, Hypertension, Heart_Disease, Smoking_History, HbA1c_level, Blood_Glucose_level, and Diabetes, where the 'Diabetes' field serves as the target variable.[6] The target variable is binary, with a value of 1 indicating the presence of diabetes and 0 indicating its absence.[6]

This dataset is specifically designed for the purpose of building machine learning models capable of predicting the likelihood of diabetes based on readily available medical history and demographic information.[6] The inclusion of the HbA1c level, a crucial indicator of long-term blood sugar control, makes this dataset particularly valuable for predictive modeling efforts. The demographic diversity, encompassing both age and gender, suggests that models trained on this data may exhibit better generalizability compared to datasets focused on a single demographic group.

2.2.4. Other Relevant Structured Datasets

Beyond the three key datasets discussed above, Kaggle hosts a wide array of other diabetes-related datasets, each with its own unique set of features and potential applications. For example, datasets from Aravindpcoders, Mathchi, and Ishandutta are mentioned in research [3], indicating their relevance within the field. The "Diabetes Dataset" by ehababoelnaga [7] provides information on individuals diagnosed with diabetes, including demographic attributes, medical history, and clinical measurements. The "Healthcare-Diabetes" dataset by nanditapore [8] includes a variety of health-related attributes for diabetes risk assessment and prediction. The "Diabetes Health Dataset Analysis" by rabieelkharoua [9] contains comprehensive health data for a cohort of patients, including demographic details, lifestyle factors, medical history, and clinical measurements. The "Diabetes Dataset" by aravindpcoder [11] focuses on medical information and laboratory analysis of diabetes patients from an Iraqi population. The "100,000 Diabetes Clinical Dataset" by priyamchoksi [12] offers a large-scale dataset with health and demographic data of 100,000 individuals. "Diabetes.csv" by pentakrishnakishore [13] provides a snapshot of women's health characteristics relevant to diabetes. The "Synthetic Diabetes 2 Type Prediction Dataset" by nigoraxonnasimova [14] contains synthetic data for predicting type 2 diabetes based on factors like age, sex, physical activity, and BMI. The "Diabetes prediction datasets" by kevintan701 [15] includes synthetically generated health information for diabetes risk assessment. Finally, the "Diabetes Dataset" by ankitbatra1210 [16] provides an overview of various types of diabetes, including genetic and lifestyle attributes.

The sheer variety of these datasets on Kaggle allows users to select resources that best align with their specific research questions or modeling objectives. Some datasets may focus on particular populations, include unique features not found elsewhere, or even be synthetically generated for specific analytical purposes. The "100,000 Diabetes Clinical Dataset" [12], with its large number of instances, is particularly noteworthy for training complex machine learning models. The "Synthetic Diabetes 2 Type Prediction Dataset" [14] offers a large and balanced dataset specifically designed for predicting type 2 diabetes. The dataset by ankitbatra1210 [16] is unique in its inclusion of information on various less common types of diabetes.

The UCI Machine Learning Repository also hosts another "Diabetes" dataset [19], which is characterized by its multivariate and time-series nature. This dataset contains records that include the date, time, a code indicating the type of measurement taken (such as regular insulin dose, NPH insulin dose, or blood glucose measurements at different times of the day), and the corresponding value. This time-series dataset offers a different perspective compared to the static, tabular datasets discussed previously, as it focuses on the temporal aspects of diabetes management. It could be particularly valuable for developing models aimed at predicting future blood glucose levels or insulin requirements rather than simply classifying the presence of diabetes at a single point in time.

2.3. Key Features of Prominent Structured Datasets

The following table summarizes the key features of three prominent structured datasets frequently used for diabetes prediction: the Pima Indians Diabetes Database, the Diabetes Health Indicators Dataset (BRFSS 2015), and the Diabetes Prediction Dataset.

Feature	Pima Indians Diabetes Database	Diabetes Health Indicators Dataset (BRFSS 2015)	Diabetes Prediction Dataset
Source (Platform)	Kaggle, UCI	Kaggle, UCI	Kaggle
Number of Instances	768	253,680	100,000
Number of Features	8	21	8
Key Features	Pregnancies, Glucose, BP, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age	Diabetes_012, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, Alcohol, Healthcare, Cost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income	Age, Gender, BMI, Hypertension, Heart_Disease, Smoking_History, HbA1c_level, Blood_Glucose_level
Target Variable	Outcome (Binary)	Diabetes_012 (3-class) / Diabetes_binary (Binary)	Diabetes (Binary)
Data Types	Numeric	Binary, Integer	Integer, Float, Object

3. Unstructured and Multi-Modal Datasets

3.1. Exploring Datasets with Clinical Notes and Medical Images

3.1.1. AI-READI Dataset

The Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights (AI-READI) project, hosted on Fairhub [5], represents a significant endeavor to create a large-scale, inclusive, and ethically sourced multimodal dataset specifically for research focused on type 2 diabetes mellitus (T2DM).[5] This project aims to collect cross-sectional data from 4,000 individuals and longitudinal data from 10% of this cohort across the United States, ensuring a balanced representation across self-reported race/ethnicity, gender, and stages of diabetes disease.[5]

The AI-READI dataset encompasses a rich variety of data modalities, including cardiac electrocardiogram (ECG) data in WaveForm DataBase (WFDB) format, clinical data potentially containing clinical notes in Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) format within CSV files, environmental sensor data in Earth Science Data Systems (ESDS) format, retinal data from Fluorescence Lifetime Imaging Ophthalmoscopy (FLIO), Optical Coherence Tomography (OCT), Optical Coherence Tomography Angiography (OCTA), and retinal photography, all in Digital Imaging and Communications in Medicine (DICOM) format (providing valuable medical images), as well as data from wearable activity monitors and continuous blood glucose monitoring devices in Open mHealth format.[5] To protect participant privacy, all protected health information (PHI), as well as information related to sex, race/ethnicity, and medication usage, has been removed from the dataset.[5] Access to the AI-READI dataset requires a multi-step process involving logging in through a verified identification system, agreeing to use the data exclusively for type 2 diabetes-related research, and accepting the terms of a custom license that includes specific restrictions on data usage, security measures, and the conditions for secondary sharing.[5]

The AI-READI dataset stands out as a crucial resource for the user due to its explicit focus on multimodal data, which includes both medical images (retinal scans) and clinical data that may contain clinical notes within the standardized OMOP CDM structure. This variety of data types opens up significant opportunities for developing more sophisticated and potentially more accurate machine learning models for diabetes research. The project's commitment to ethical sourcing and ensuring balanced representation across diverse demographic groups addresses critical considerations in the field of medical AI. However, the specific access requirements and the custom license terms necessitate careful review by the user to ensure they can comply with the conditions for utilizing this valuable resource.

3.1.2. Veradigm Cardiometabolic Clinical Registry Datasets

Veradigm, a provider of healthcare data and technology solutions, has introduced disease-specific cardiometabolic Clinical Data Registry datasets, which include de-identified information about patients' health and care related to specific conditions, including both Type 1 and Type 2 Diabetes.[28] These datasets leverage the power of Natural Language Processing (NLP) to extract valuable insights from both structured and unstructured clinical data, effectively harmonizing data originating from over 80 different Electronic Health Records (EHRs) into a research-ready data model.[28]

These registries contain a combination of traditional structured data fields, likely encompassing standard information found in EHR systems, along with unstructured data such as clinical notes, radiology reports, and other textual medical records. Veradigm employs advanced NLP techniques to analyze this unstructured data, extracting clinically relevant information that can then be integrated with the structured data to provide a more comprehensive view of disease patterns, treatment outcomes, and overall patient care.[28] While the specific data fields included in these registries are not detailed in the provided snippets, the core value proposition lies in the combination of structured and NLP-enhanced unstructured data. Access to Veradigm's Clinical Data Registries is facilitated through Veradigm itself, as they are a commercial provider of these specialized healthcare data solutions.

Veradigm's datasets are particularly relevant to the user's interest in unstructured data, as they explicitly incorporate and analyze clinical notes and other forms of unstructured information using NLP methodologies. The harmonization of data from a multitude of different EHR systems addresses a significant challenge in real-world healthcare data analysis, which is the heterogeneity of data formats and standards. The availability of disease-specific registries, including those for Type 1 and Type 2 Diabetes, allows for highly targeted research efforts. However, it is important to note that access to these datasets is proprietary and would require direct engagement with Veradigm, potentially involving specific agreements or financial considerations.

3.1.3. AnswerY Unstructured Database

AnswerY™ is a proprietary qualitative database and platform developed by Amplity, containing a vast collection of over 80 million HIPAA-compliant, unstructured patient-provider interactions and insights derived from medical transcripts.[29] This rich database leverages advanced Natural Language Processing (NLP) techniques to extract valuable information and gain a deeper understanding of the "why" behind treatment decisions made in clinical practice.[29] A compelling study that utilized

AnswerY demonstrated a significant underreporting of hypoglycemia (low blood sugar) in traditional structured data sources, such as claims data and Electronic Health Records (EHRs), revealing that the prevalence of hypoglycemia was estimated to be 2 to 9 times higher when analyzed using the unstructured data in AnswerY.[29] This finding underscores the limitations of relying solely on structured data and highlights the critical value of incorporating unstructured data sources for a more accurate and comprehensive view of patient health, particularly for individuals with diabetes.

The primary data source within AnswerY is the unstructured text derived from medical transcriptions. This type of data can include detailed descriptions of patient symptoms, discussions about treatment options and decisions, and the underlying rationale behind clinical management strategies. While the specific structured fields are not the focus of AnswerY, its strength lies in the depth and context that the unstructured text provides. Access to the AnswerY database is proprietary to Amplity and is not publicly available.

AnswerY serves as a powerful illustration of the rich insights that can be obtained from analyzing unstructured medical data, specifically in the form of clinical transcripts. The study's finding of significantly underreported hypoglycemia in structured data emphasizes the potential of unstructured data to provide a more complete and accurate picture of patient health in diabetes management. While the user likely cannot directly access the AnswerY database without engaging with Amplity, it provides a compelling example of the type of valuable information that can be derived from analyzing similar unstructured data sources if they are available.

3.2. Discussion of the Potential and Challenges of Using Unstructured Data

Unstructured data, including clinical notes, medical images, and other textual or visual information, offers a significant potential to enrich the development of diabetes classification models. Clinical notes can capture a wealth of information about patients' subjective experiences, detailed descriptions of symptoms, the evolution of their condition over time, and the nuanced reasoning behind treatment decisions.[29] This level of detail often goes beyond the standardized fields available in structured datasets. Medical images, such as retinal scans obtained through Optical Coherence Tomography (OCT) or retinal photography, can provide direct visual evidence of microvascular complications associated with diabetes, like diabetic retinopathy.[5] Integrating this type of data into machine learning models could lead to more accurate and comprehensive diagnostic and predictive capabilities.

However, the utilization of unstructured data also presents several challenges that researchers and practitioners must address. Extracting meaningful and actionable information from unstructured text typically requires the application of sophisticated Natural Language Processing (NLP) techniques.[28] These techniques can be complex to implement and may require specialized expertise. Similarly, analyzing medical images necessitates knowledge of image processing and computer vision methodologies. The integration of unstructured data with existing structured data sources can also be a complex process, often requiring advanced data fusion techniques to effectively combine information from disparate formats. Furthermore, the quality and consistency of unstructured data can vary significantly depending on the source, the clinician's documentation style, and the specific practices of different healthcare settings. Ensuring data quality and handling missing or inconsistent information are critical steps in the process. Finally, ethical considerations surrounding patient privacy and data security are paramount when working with sensitive medical information, particularly unstructured data which may contain more free-form details. Robust de-identification and data governance protocols must be in place to protect patient confidentiality.

4. Accessing Patient Health Record Databases (PHRs)

4.1. Overview of Publicly Accessible PHR Resources

4.1.1. PCORnet

The National Patient-Centered Clinical Research Network (PCORnet) is a substantial research infrastructure program that has established a network of over 60 health systems across the United States.[30] These participating health systems have undertaken the significant task of mapping their diverse clinical data to a standardized data model, facilitating large-scale, comparative effectiveness research.[30] PCORnet annually aggregates longitudinal health information from over 30 million patients, offering a broad and geographically diverse representation of healthcare data, encompassing information from both inpatient and outpatient settings.[30] This network plays a vital role in supporting public health surveillance efforts and enabling researchers to conduct comparative studies across a wide range of medical conditions, including the capacity to assess disease severity and control over time using objective measures such as glycosylated hemoglobin levels in patients with diabetes.[30]

PCORnet operates on a distributed query infrastructure, which allows authorized users to submit specific research queries to the network. The system then coordinates responses from the participating health systems, aggregating the relevant data to provide a comprehensive answer to the user's query.[30] While PCORnet has established protocols that allow for the transfer of patient-level data to researchers under specific circumstances, the availability of a reusable process for obtaining aggregate data from the network's partners often allows for quicker completion of many research assessments.[30]

PCORnet presents a valuable opportunity for the user to access a vast amount of real-world clinical data relevant to diabetes research. The adoption of a standardized data model across the network facilitates the integration and analysis of data from multiple sources. The distributed query system provides a mechanism for researchers to explore the data and obtain relevant information without necessarily needing to download entire patient-level datasets, which can be particularly useful for initial exploratory analyses and population-level studies focused on diabetes prevalence, treatment patterns, and outcomes.

4.1.2. T1D Exchange

The T1D Exchange is a dynamic and comprehensive patient data platform specifically designed to accelerate all aspects of research related to type 1 diabetes (T1D).[31] It operates as a learning health system, bringing together nearly 50 pediatric and adult endocrinology centers located across the United States.[32] The T1D Exchange Quality Improvement Collaborative (T1DX-QI) leverages an innovative web-based platform known as the QI Portal to systematically gather and securely store electronic medical record (EMR) data from participating centers.[32] This collaborative effort aims to promote benchmarking, facilitate population health improvement, and ultimately advance the understanding and treatment of type 1 diabetes.[32] The T1DX-QI network has demonstrated its effectiveness in improving clinical outcomes for patients with type 1 diabetes and has utilized the collected EMR data to generate valuable real-world evidence.[32]

The data specifications for the T1D Exchange include a detailed set of over 120 unique variables, organized across seven core data files: Patients, Providers, Encounters, Observations, Conditions, Medications, and a dedicated Diabetes file.[32] The Diabetes file contains specific information on adverse outcomes experienced by patients, their glucose monitoring practices, insulin treatment plans, and detailed patient glucose readings.[32] To enable longitudinal tracking of patients over time, the T1D Exchange shares limited protected health information, specifically dates of service and five-digit zip codes, while utilizing unique patient identifiers to maintain patient privacy.[32] Researchers interested in accessing the T1D Exchange data can submit proposals for clinical trial protocols or request access to analyze existing cases within the network.[31] The data is securely stored within the United States, and access to the QI Portal is restricted to authorized users who have a legitimate need for the data, with privacy and security further enhanced through the implementation of two-factor user authentication.[32]

The T1D Exchange represents a highly specialized and comprehensive resource for researchers specifically focused on type 1 diabetes. The detailed data specifications, the longitudinal nature of the data collection, and the platform's demonstrated success in generating real-world evidence make it an invaluable asset for advancing knowledge in this area. The requirement for a formal research proposal for data access reflects the depth and sensitivity of the information contained within the platform, ensuring responsible and ethical data utilization.

4.1.3. Other Registries and Databases

Beyond PCORnet and the T1D Exchange, several other publicly accessible resources may offer valuable data or information relevant to diabetes research. **HCUPnet** is a free, online query system that provides access to a wealth of health statistics and information related to hospital inpatient and emergency department utilization, based on data from the Healthcare Cost and Utilization Project (HCUP).[33] While not solely focused on diabetes, this resource can provide insights into hospitalization rates and patterns associated with the condition.

Disease registries are clinical information systems that systematically track a defined set of key data for individuals with specific chronic conditions, including Alzheimer's Disease, cancer, diabetes, heart disease, and asthma.[33] One example mentioned is the **Surveillance, Prevention, and Management of Diabetes Mellitus DataLink (SUPREME DM)** [33], which has been specifically used in research to validate methods for identifying diabetes mellitus using electronic health record data.[34] Access to data from these registries may vary, but they represent a potential source of aggregated or even patient-level information for diabetes research.

Finally, resources like **ClinicalTrials.gov**, the **Cochrane Library**, and the **WHO International Clinical Trials Registry Platform (ICTRP)** serve as comprehensive registries of clinical trials conducted around the world, both publicly and privately supported.[33] These platforms can provide valuable information about ongoing and completed research related to diabetes prevention, treatment, and management, although they may not directly offer datasets for developing classification models.

4.2. Considerations for Access and Data Use

Accessing Patient Health Record databases, such as PCORnet and the T1D Exchange, often involves specific requirements and procedures designed to ensure responsible data use and the protection of patient privacy.[31] Researchers typically need to submit formal agreements, detailed research proposals outlining their study objectives and methodology, and may need to have institutional affiliations to gain access to these sensitive datasets. These processes are essential for maintaining the ethical standards of research and complying with relevant data privacy regulations.

Once access is granted, researchers are typically bound by data use agreements that specify the permissible ways in which the data can be utilized.[5] These agreements often restrict the use of data to statistical reporting and analysis purposes only, explicitly prohibiting any attempts to re-identify individual patients or to link the data with other datasets that could potentially lead to identification. Adherence to these terms and to broader privacy regulations, such as HIPAA in the United States, is paramount for researchers working with patient health information.

A common practice in the management of these databases is the de-identification of patient data.[28] This process involves removing or obscuring any information that could directly identify an individual, such as names, addresses, and specific dates of birth. While de-identification is crucial for protecting patient privacy, researchers should be aware of the level of de-identification applied to a particular dataset and any potential limitations it might impose on their specific research questions or analytical approaches.

5. Datasets Highlighted in Diabetes Research Literature

5.1. Analysis of Research Papers

The **Pima Indians Diabetes Database** stands out as a consistently utilized resource in numerous research papers focusing on the prediction of diabetes and the evaluation of various machine learning algorithms.[4] Its frequent appearance in academic literature has firmly established it as a foundational dataset within the field, making it a standard benchmark against which new methodologies are often compared.

Research endeavors also frequently leverage other publicly accessible datasets, including **Aravindpcoders' diabetes dataset, Mathchi's diabetes dataset, and Ishandutta's early-stage diabetes risk prediction dataset**, all of which are readily available on the Kaggle platform.[3] Notably, some research studies have even adopted the strategy of combining multiple of these datasets to create larger and potentially more robust training datasets, aiming to enhance the generalizability and performance of the resulting machine learning models.[3]

The **Behavioral Risk Factor Surveillance System (BRFSS) data** also emerges as a popular choice for research investigations focused on understanding diabetes prevalence, identifying key risk factors associated with the disease, and developing predictive models at a population level.[1] The large-scale and population-based nature of the BRFSS data make it particularly well-suited for addressing broader epidemiological and public health research questions related to diabetes.

Several research papers highlight the availability of **Chinese diabetes datasets, specifically the ShanghaiT1DM and ShanghaiT2DM datasets**, as well as the **T1DiabetesGranada dataset**, for research purposes.[37] These datasets offer valuable opportunities for the development of data-driven algorithms and technologies aimed at improving diabetes monitoring and management, potentially contributing to advancements in personalized diabetes care.

For research specifically focused on type 1 diabetes, the **OhioT1DM Dataset** and the **D1NAMO dataset** are highlighted as significant resources for studies investigating blood glucose level prediction and exploring non-invasive management techniques.[37] These specialized datasets cater to the unique characteristics and management challenges associated with type 1 diabetes, providing researchers with more granular and relevant data for this particular form of the disease.

More recent research efforts continue to evaluate the performance of machine learning techniques across a range of datasets, including the **PIMA dataset, the Diabetic Dataset 2019, and the BIT_2019 dataset**.[26] This ongoing evaluation and the emergence of new datasets underscore the dynamic nature of the field and the continuous efforts to refine and improve diabetes prediction methodologies through data-driven approaches.

5.2. Identifying Datasets Frequently Used in Machine Learning for Diabetes Prediction

Based on the analysis of research literature, the **Pima Indians Diabetes Database** and various datasets derived from the **CDC Behavioral Risk Factor Surveillance System (BRFSS)** consistently rank among the most frequently utilized publicly available datasets for machine learning-based diabetes prediction. The Kaggle platform serves as a primary source for accessing many of these commonly used datasets, as well as a growing number of other relevant datasets that are increasingly being adopted by the research community.

6. Key Considerations for Dataset Selection

6.1. Available Data Fields and their Relevance to Diabetes Classification

When embarking on the task of selecting a dataset for diabetes classification modeling, the user should meticulously examine the available data fields within each potential dataset and critically assess their relevance to predicting the presence or absence of diabetes.[1] Certain features, such as blood glucose level, HbA1c level (a measure of long-term blood sugar control), Body Mass Index (BMI), blood pressure measurements, and information regarding family history of diabetes, are well-established risk factors for the disease and are therefore highly pertinent for predictive modeling. Datasets that offer a broader spectrum of potentially influential features may enable the development of more accurate and comprehensive models capable of capturing the complex interplay of factors contributing to diabetes.

The user's model's ability to accurately predict diabetes will be fundamentally influenced by the quality and the direct relevance of the features included in the chosen dataset. A thorough understanding of the available data fields and their known associations with diabetes is therefore a crucial prerequisite for effective model building. Datasets that provide a more extensive array of pertinent predictors are generally more likely to support the development of robust and reliable classification models.

For users who are also interested in exploring the potential of unstructured data, it is important to consider the types of information that might be contained within clinical notes (such as detailed descriptions of patient symptoms, the specific course of treatment, and the progression of the disease) or within medical images (for example, retinal scans that can reveal signs of diabetic retinopathy).[5] This kind of data can often provide valuable contextual information and insights that are not typically captured in structured datasets, potentially leading to the development of more nuanced and accurate predictive models capable of leveraging a richer understanding of the patient's condition. However, the user should also be prepared to address the inherent complexities associated with processing and analyzing unstructured data.

6.2. Licensing and Terms of Use for Each Identified Dataset

It is absolutely essential for the user to carefully review and understand the licensing agreements and terms of use associated with each dataset they are considering for their model development.[5] Datasets hosted on platforms like Kaggle often come with open-source licenses, such as the Apache 2.0 license [8]), or may have specific terms and conditions defined by the individuals or organizations that contributed the data. Similarly, datasets available through the UCI Machine Learning Repository frequently utilize Creative Commons licenses, for instance, the CC BY 4.0 license that applies to the primary "Diabetes" dataset.[21] The AI-READI dataset, hosted on Fairhub, has its own custom license that includes specific requirements regarding the permitted uses of the data for research, the security measures that must be implemented, and the conditions under which the data can be shared.[5] For data originating from the CDC BRFSS, it is generally considered to be within the public domain and can be reproduced without explicit permission, although proper citation of the source is typically requested.[27] However, users should be aware that some state-specific BRFSS data may have additional agreements or restrictions [38], and the National Center for Health Statistics (NCHS) Data User Agreement outlines specific terms and conditions for the use of NCHS data, including prohibitions on attempting to re-identify individuals.[35]

A thorough understanding of these licensing terms is crucial for ensuring ethical and legal compliance in the user's research. The user must verify that the license associated with their chosen dataset permits the intended use, which in this case is the development and potential deployment of a diabetes classification model. While the public domain status of CDC data offers considerable flexibility, the user must still adhere to the requested citation guidelines. For datasets with more restrictive licenses, such as the custom license for the AI-READI dataset, the user must carefully review and ensure they can meet all the stipulated requirements for accessing and utilizing the data.

6.3. Common Usage and Benchmarking in the Field

Certain datasets, notably the Pima Indians Diabetes Database and various versions of the BRFSS datasets, have become widely adopted within the diabetes research and machine learning communities as standard resources for benchmarking the performance of new algorithms and comparing research findings.[4] Selecting one of these commonly used datasets can offer significant advantages to the user. It allows them to directly compare the performance metrics of their developed model against those reported in existing research literature, providing a valuable context for evaluating the effectiveness of their approach and understanding the current state-of-the-art in diabetes prediction. Furthermore, utilizing a dataset that is well-established in the field increases the likelihood of finding relevant research papers, code repositories, and other resources that can provide guidance and support throughout the model development process.

7. Conclusion: Recommendations for Selecting Appropriate Datasets for Diabetes Classification

Based on the user's stated need for datasets to develop a machine learning model for classifying whether a patient has diabetes, a diverse range of publicly available options exist, each with its own strengths and characteristics. The most suitable choice will depend on the user's specific research objectives, the types of features they wish to incorporate into their model, the desired scale and diversity of the data, the licensing terms associated with the dataset, and their interest in leveraging unstructured data sources.

For users seeking a well-established and widely recognized structured dataset to begin their work, the **Pima Indians Diabetes Database**, available on both Kaggle and the UCI Machine Learning Repository, represents a solid starting point due to its long history of use in the field and its clearly defined set of features. If the user requires a larger and more demographically representative structured dataset that also includes a broader range of health indicators and lifestyle factors, the **Diabetes Health Indicators Dataset (BRFSS 2015)**, accessible through both Kaggle and UCI, is highly recommended. The **Diabetes Prediction Dataset** on Kaggle offers a focused collection of key medical and demographic features, including the important HbA1c level, making it a strong contender for building a targeted predictive model.

For users interested in exploring the potential of unstructured data, the **AI-READI Dataset** on Fairhub presents a valuable opportunity to work with a rich multimodal dataset that includes medical images and potentially clinical notes. However, the user must carefully review and comply with the specific access requirements and license terms associated with this resource. Investigating Veradigm's cardiometabolic clinical registries might also be a viable option if access can be obtained, as these registries incorporate unstructured data through the use of Natural Language Processing techniques. Researchers specifically focused on type 1 diabetes might find the **OhioT1DM Dataset** and the **D1NAMO dataset** particularly relevant for their work. For those interested in the temporal aspects of diabetes management, the time-series **Diabetes dataset from the UCI Machine Learning Repository**, containing information on insulin doses and blood glucose levels over time, could be a valuable resource.

Ultimately, the user should carefully weigh the characteristics of each potential dataset against their specific needs and research goals. Regardless of the dataset chosen, it is advisable to benchmark the performance of the developed model against results reported in the literature for commonly used datasets like the Pima Indians Diabetes Database and the BRFSS datasets. This will provide a crucial frame of reference for evaluating the model's effectiveness and identifying potential avenues for further improvement.

Works cited

1. Diabetes Health Indicators Dataset - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
2. Predicting Diabetes Mellitus With Machine Learning Techniques - Frontiers, accessed April 1, 2025, https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2018.00515/full
3. Enhanced detection of diabetes mellitus using novel ensemble feature engineering approach and machine learning model, accessed April 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11458802/
4. Machine Learning Algorithm-Based Prediction of Diabetes Among Female Population Using PIMA Dataset - MDPI, accessed April 1, 2025, https://www.mdpi.com/2227-9032/13/1/37
5. Flagship Dataset of Type 2 Diabetes from the AI-READI Project, accessed April 1, 2025, https://fairhub.io/datasets/2
6. Diabetes prediction dataset - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset
7. Diabetes - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/ehababoelnaga/diabetes-dataset
8. Healthcare Diabetes Dataset - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes
9. Diabetes Health Dataset Analysis - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/rabieelkharoua/diabetes-health-dataset-analysis
10. Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes - PubMed Central, accessed April 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10378239/
11. Diabetes Dataset - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/aravindpcoder/diabetes-dataset
12. Comprehensive Diabetes Clinical Dataset(100k rows) - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/priyamchoksi/100000-diabetes-clinical-dataset
13. Diabetes Prediction Dataset - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/pentakrishnakishore/diabetes-csv
14. Diabetes prediction dataset - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/nigoraxonnasimova/synthetic-diabetes-2-type-prediction-dataset
15. Diabetes Prediction datasets - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/kevintan701/diabetes-prediction-datasets
16. Diabetes Dataset - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/ankitbatra1210/diabetes-dataset

17. Diabetes Health Indicators Dataset - Kaggle, accessed April 1, 2025, https://www.kaggle.com/code/ratchaphonp/diabetes-health-indicators-dataset/data
18. Diabetes Health Indicators - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/siamaktahmasbi/diabetes-health-indicators/data
19. Datasets - UCI Machine Learning Repository, accessed April 1, 2025, https://archive.ics.uci.edu/datasets?skip=10&take=10&sort=desc&orderBy=NumHits&search=
20. Datasets - UCI Machine Learning Repository, accessed April 1, 2025, https://archive.ics.uci.edu/datasets?search=&Keywords=Survey
21. Diabetes - UCI Machine Learning Repository, accessed April 1, 2025, https://archive.ics.uci.edu/dataset/34/diabetes
22. UCI Machine Learning Repository: Diabetes Data Set, accessed April 1, 2025, http://archive.ics.uci.edu/ml/machine-learning-databases/00000/UCI%20Machine%20Learning%20Repository%20Diabetes%20Data%20Set.htm
23. CDC Diabetes Health Indicators - UCI Machine Learning Repository, accessed April 1, 2025, https://www.archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators
24. Diabetes Dataset - Kaggle, accessed April 1, 2025, https://www.kaggle.com/datasets/mathchi/diabetes-data-set
25. An Effective Methodology for Diabetes Prediction in the Case of Class Imbalance - MDPI, accessed April 1, 2025, https://www.mdpi.com/2306-5354/12/1/35
26. Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets - Frontiers, accessed April 1, 2025, https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1499530/full
27. BRFSS Frequently Asked Questions (FAQs) - CDC, accessed April 1, 2025, https://www.cdc.gov/brfss/about/brfss_faq.htm
28. Veradigm Introduces Cardiometabolic Clinical Registry Datasets to ..., accessed April 1, 2025, https://investor.veradigm.com/news-releases/news-release-details/veradigm-introduces-cardiometabolic-clinical-registry-datasets
29. Unstructured Data & NLP Reveal Hidden Hypoglycemia Cases ..., accessed April 1, 2025, https://amplity.com/news/unstructured-data-nlp-reveal-hidden-hypoglycemia-cases
30. Public Health Surveillance in Electronic Health Records: Lessons From PCORnet - CDC, accessed April 1, 2025, https://www.cdc.gov/pcd/issues/2024/23_0417.htm
31. Resources for Researchers | American Diabetes Association, accessed April 1, 2025, https://professional.diabetes.org/research-grants/resources-researchers
32. Making Diabetes Electronic Medical Record Data Actionable: Promoting Benchmarking and Population Health Improvement Using the T1D Exchange Quality Improvement Portal, accessed April 1, 2025, https://diabetesjournals.org/clinical/article/41/1/45/147670/Making-Diabetes-Electr

onic-Medical-Record-Data

33. Data Resources in the Health Sciences: Clinical Data - Library Guides, accessed April 1, 2025, https://guides.lib.uw.edu/hsl/data/findclin
34. Assessing the Effect of Electronic Health Record Data Quality on Identifying Patients With Type 2 Diabetes: Cross-Sectional Study - JMIR Medical Informatics, accessed April 1, 2025, https://medinform.jmir.org/2024/1/e56734
35. Data User Agreement | National Center for Health Statistics - CDC, accessed April 1, 2025, https://www.cdc.gov/nchs/policy/data-user-agreement.html
36. BRFSS Data Use Responsibilities and Guidelines - Washington State Department of Health, accessed April 1, 2025, https://doh.wa.gov/data-and-statistical-reports/data-systems/behavioral-risk-factor-surveillance-system-brfss/brfss-data-use-responsibilities-and-guidelines
37. [PDF] Chinese diabetes datasets for data-driven machine learning - Semantic Scholar, accessed April 1, 2025, https://www.semanticscholar.org/paper/4fbb688b8734a573b78914a3b0609e7a66c9ad19
38. BRFSS - CDPH - CA.gov, accessed April 1, 2025, https://www.cdph.ca.gov/Programs/CCDPHP/DCDIC/CDSRB/Pages/BRFSS.aspx