RESEARCH PAPER

Open Data for Clinical AI: A Comprehensive Review of Disease Prediction Datasets

Sindhu Rajendran1*and Chandrashekar B S2

[1]*Department of ECE, RV College of engineering, Bengaluru, India*
[2] *Department of ECE, Jain Deemed-to-be-University, Bengaluru, India*
[1]*sindhur@rvce.edu.in and* [2]*cshastry2@gmail.com*
***Corresponding Author:** sindhur@rvce.edu.in*

Received: 28th Feb, 2026; Revised: 6th March 2026; Accepted: 7th April, 2026; Available Online: 20th April, 2026

ABSTRACT
With the advancements in the medical sector world-wide, the use of Machine learning has been in use. In order to use these machine learning models for prediction and diagnosis of certain diseases one of the main components is datasets. The need for high-quality datasets in healthcare prediction models is critical due to the data-driven nature of machine learning. These models rely on comprehensive, accurate, and representative datasets to make reliable predictions that can impact real-world patient outcomes. This paper provides an insight about the different components in the datasets present for diseases such as Osteoporosis, Heart disease, Diabetes, Respiratory, Syncytial Virus, Interactive Thyroid, Parkinson's and Sepsis. Also a comparative study on the parameters of the datasets in the Indian perspective and globally are also discussed.

Keywords: Health Datasets; Disease Prediction; Machine Learning in Healthcare; Dataset Quality Assessment; Clinical Data Analytics
How to cite this article: Rajendran S, Chandrashekar BS. Open Data for Clinical AI: A Comprehensive Review of Disease Prediction Datasets. Int J Drug Deliv Technol. 2026;16(56s): 1323-1332. DOI: 10.25258/ijddt.16.56s.147
Source of support: Nil.
Conflict of interest: None

1. INTRODUCTION

An interesting fact in today's healthcare practice is that the use of machine-learning-based predictive models has become more common. These models help many clinicians make decisions, alongside improving patient outcomes, and also increase operational efficiency. The reliable performance of such models is quite contingent upon the availability of high-quality as well as well-formatted datasets that constitute the primary input, that's necessary for identifying complex relationships between clinical variables and health outcomes. Thus, healthcare datasets are crucial for many applications, including diagnosing diseases, predicting patient outcomes, and creating personalized treatment plans. Therefore, the accurate generalizability is ensured through the use of large, representative datasets during model training and validation, thereby guaranteeing the applicability of predictive models across heterogeneous populations and clinical contexts. This highlights the intense value of longitudinal data in predicting negative outcomes in patients such as hospital readmissions, disease progression or recovery after surgery. This will mainly allow the clinicians to intervene right before a negative health outcome, that could have been serious or problematic. It has to be noted that beyond their clinical utility, these datasets also heavily facilitate the optimization of healthcare resource allocation, provide hospitals the empowerment to smoothen and refine the patient flow along with staffing configurations, and the systematic and effective utilization of medical equipment. As our various healthcare systems approach full digitization, the need and demand for rich, heterogeneous datasets will likely compound. This would eventually render them as a fundamental tool or element of modern healthcare innovation. This collective inquiry aims to furnish a comparative overview of more than twenty such publicly available disease datasets that comprise clinical, imaging, and genomic data modalities. Thus, this current research offers:

1. A detailed side by side comparison of the diversity as well as the data coverage;
2. An analysis into the geopolitical biases, specifically the disparity between the U.S specific datasets and the India based repositories that dominate various global collections;
3. The classification of some underrepresented categories of diseases such as chronic lifestyle-related disorders, pediatric disease, and their associated cohorts and diseased populations where existing, publicly available repositories currently are minimal;
4. A detailed discussion of how structural features, such as representation of dimensions,sampling strategies, and historical pathways that shape the generalizability and operational utility of machine-learning models in model healthcare practice.

**Author for Correspondence: sindhur@rvce.edu.in*

2. RELATED WORK

A detailed study of the current methodologies used for the publically available datasets for various diseases is discussed in this section. key bone health and metabolic indicators are highlighted by Osteoporosis datasets. For example, the Osseous Osteoporosis Screening dataset (Brazil) records multiple factors like BMI, age, bone mineral density (BMD), and related factors like vitamin D levels and calcium. These align well with clinical needs, since the direct risk factors for osteoporosis and fractures are low BMD and deficiencies in calcium/vitamin D. The structure of osteoporosis datasets varies; one is a structured table of clinical readings and patient attributes, the other comprises medical images plus an annotation file. Additionally, while one dataset includes DX-measured BMD, there is no incorporation of frailty indicators or fall risk factors that clinicians consider when evaluating fracture risk in osteoporotic patients. Publicly available datasets on heart disease generally encompass the core cardiovascular risk and diagnostic variables of interest. Examples include the Cleveland Heart Disease dataset, which reports age, sex, systolic and diastolic blood pressure, cholesterol (in terms of its total amount and as high-density lipoproteins) and any results of the tests of the diagnostics like the electrocardiogram (ECG) and an exercise stress test. One more similar data on heart-failures that are collected in Faisalabad rather than in Pakistan is showing measurable past of hypertension (hypertension), diabetes state, anaemia, and such clinical indexes as ejection fraction and creatinine in serum. These features are considered closely related to those clinical cardiovascular event risk factors which are known, hypertension, diabetes, and dyslipidaemia. Both the Cleveland and the Faisalabad heart-failure datasets supply both angina symptoms and ECG outcomes, which highlights the diagnostic characteristics of coronary artery disease; moreover, the Faisalabad-heart-failure dataset also has an outcome endpoint (death within a pre-structured follow-up period) but lacks in details of the longitudinal parameters more than just a single measure of survival time and binary event label. Both of the datasets do not mention physical activity, dietary habits, or stress variables, which are however analytically proven important variables in cardiovascular risk assessment which is quite hard to measure. In conclusion, even though these datasets capture conventional indicators such as lipids, blood pressure, and diagnostic outcomes, they do not provide the complete profile that cardiologists consider to be sufficient to patient care, including lifestyle factors, familial tendencies, and other contextual influences. Although,some basic clinical characteristics are covered by the diabetes datasets which are currently available, there are still gaps in their ability to fully depict the clinical picture of diabetes. Age, blood pressure, BMI, plasma glucose concentration from an oral glucose tolerance test, number of pregnancies (for female patients), a diabetes pedigree function, and a one-time 2-hour serum insulin test are among the physiological measurements that are the focus of the Pima Indians Diabetes dataset. These characteristics are in line with the clinical knowledge and correlate to risk factors or diagnostic criteria for type 2 diabetes, such as high blood glucose and insulin resistance, obesity (BMI), and genetic predisposition (pedigree). This dataset, however, is notable for only including female patients from a particular ethnic group (Pima Native Americans), meaning it does not account for broader genetic backgrounds or sex-specific differences (no males). The more extensive ”Diabetes130-US dataset provides a more extensive clinical picture of diabetic patients admitted to hospitals: it contains demographics (age, gender, race), hospital metrics (admission type, length of stay), laboratory tests (like glucose levels, HbA1c results), and details on diagnoses and medications. Numerous clinical facets pertinent to the management of diabetes inpatients are covered here. However, because the focus of this hospital dataset is on in-hospital outcomes, out-of-hospital factors might not be fully represented. In conclusion, lifestyle factors (diet, exercise), chronic complications, and wider population diversity are underrepresented in current datasets, despite the presence of the core biomedical features of diabetes (glucose, BMI, etc.). China’s national virus surveillance dataset is the RSV-related dataset that was highlighted in the survey. Although the data is limited from a clinical standpoint, it is very helpful for public health monitoring because it provides information on the time, location, and context of RSV cases (e.g., climate factors). When diagnosing RSV in a patient, doctors frequently take into account the patient’s symptoms, vital signs such as body temperature and respiration rate, and possibly oxygen saturation. As a result, although the dataset satisfies epidemiological requirements, it falls short in meeting clinical diagnostic requirements for RSV. Although not stated directly in the data matrix, important characteristics like fever or respiratory rate are suggested by the nature of RSV. In conclusion, many features are lacking this data set that are necessary for patient-level predictions, but it is relevant medically for population-level predictions. The UCI repository’s Thyroid Disease dataset, which includes a range of patient information and thyroid-related lab results, is the one that is referenced. Important characteristics include age, sex, and levels of several thyroid hormones, including TSH, T3, and T4 (total and possibly free T4), as well as binding proteins like TBG (thyroxine-binding globulin) and uptake values. In addition, it also contains thyroid function indicators, including other diagnostic flags and a potential goiter indicator. These characteristics closely match the criteria used by medical professionals to identify thyroid conditions. For example, primary hypothyroidism is suggested by a high TSH and low T4, which this dataset may be able to detect. The dataset, which focuses on laboratory diagnostics, is a little out of date. This thyroid dataset is organized as one record per patient evaluation, containing that patient’s lab results and some categorical flags.

Two Parkinson’s disease datasets were identified: one that focused on voice measurements and the other that was

more comprehensive (PPMI). Specialized voice features—measures of dysphonia like fundamental frequency variation, jitter, shimmer, and harmonic to-noise ratio—can be found in the Parkinson's tele-monitoring voice dataset. These characteristics accurately depict the soft, monotone, or hoarse voice impairment that is frequently seen in Parkinson's patients. Although voice change is a clinically recognized symptom of Parkinson's disease and can be used to track the disease's progression, it is only one facet of the illness. However, the Parkinson's Progression Markers Initiative (PPMI) dataset is multimodal and clinically rich, containing cerebrospinal fluid biomarkers, brain imaging results (MRI and dopamine transporter SPECT scans), age, sex, and clinical scores (UPDRS), which measure tremor, rigidity, bradykinesia, etc. It covers objective imaging to differentiate Parkinson's from other disorders, the gold-standard clinical rating (UPDRS), and even biochemical markers being studied for progression, all of which are highly aligned with the medical needs for Parkinson's disease. In essence, PPMI offers the type of comprehensive data that a neurologist would collect during several visits. One drawback of PPMI, a research cohort, is that it might not accurately reflect the entire Parkinson's population. Furthermore, genetic information isn't specifically mentioned in either PPMI or the voice dataset, though PPMI may contain some genomic information. PPMI is a controlled research dataset that is medically comprehensive, whereas the voice dataset is medically useful for a very specific use-case. Parkinson's data comes in two flavors – one is a simple table of voice features per recording (with an associated patient ID and clinical score), the other is a complex multi-modal database. The sepsis dataset, which focuses on vital signs, lab results, and interventions, is a synthetic derivative of ICU data (MIMIC-III). It comprises lab tests, records of fluid or vasopressor administration, and longitudinal vital signs such as temperature, heart rate, blood pressure, and oxygen saturation (SpO). These are precisely the features that intensivists look for in patients with sepsis; in fact, they match the criteria used to define sepsis. These time-series data make the dataset ideal for training models that use vital signs trends to predict septic shock or detect early sepsis. Early antibiotic initiation is crucial in real care; a model that could account for whether or not antibiotics were administered could improve outcome prediction. Lastly, since the population is MIMIC-III derived, it is slanted toward adult patients in US tertiary hospitals; sepsis or pediatric sepsis in settings with limited resources may not be represented. In conclusion, the dataset includes basic physiological information about sepsis that is medically significant for both detection and treatment; however, it leaves out background health information and some treatment details that would also be helpful in a clinical setting.

3. COMPARATIVE RESEARCH AND CONTEXTUAL BENCHMARKING

In addition to surveying the publicly available datasets, it is important to point this work within the broader research landscape. It has to be noted that several landmark studies have explored the application of machine learning in healthcare and the associated dataset challenges.

Ching et al. [23] outlined the advantages and disadvantages of deep learning in biology and medicine. The study particularly highlighted limitations in interpretability, privacy, and data heterogeneity problems that are consistent with findings in this manuscript, especially for the datasets related to sepsis and Parkinson's disease. By describing the gaps in individual-level clinical features and the lack of standardization across publicly available datasets, the current work supplements this.

The MIMIC-III database, which was first presented by John-son et al. [24], has grown to be a vital tool in ML-based intensive care unit research. The analysis of sepsis-related data in this study has made extensive use of citations to this particular dataset. Their research establishes a standard for how the various predictive models, particularly in critical care, can be greatly aided by structured, high-resolution time series data. Our synthesis reaffirms the importance of MIMIC-III database and compares it to other datasets that lack various and specific longitudinal or physiological time-series data.

Also, Wu et al. [25] gave a thorough analysis of machine learning for medical diagnostics, detailing its development, present applications, and future prospects. This source offers a theoretical framework for analyzing the direct effects of dataset structure and selection on clinical reliability, overfitting risk, and model generalizability. The disease-specific analysis and dataset taxonomy described in this manuscript offer specific instances of these theoretical ideas in action. These studies serve as foundational references and validate the approach. By extending their work with a structured taxonomy and comparative dataset analysis, this manuscript contributes an applied and data-centric perspective to the evolving domain of machine learning in healthcare.

4. METHODOLOGY

We surveyed the literature from 2020–2025 on ML in disease prediction, focusing on IEEE, Elsevier, Springer, and Nature publications. For each disease domain, we identified representative datasets. We then compiled a comparison table (Table 1) that integrates information from multiple sources: for instance, Tu et al.'s osteoporosis dataset, Wang et al.'s sepsis data, etc.

The table 1 specifies the different types of diseases for which the diagnostic support or the prediction model was made using the following parameters such as biographic data and other specific data required for analysis.

5. RESULTS

The table 2 specifies the comparison of all the eminent diseases and the sources of all the datasets used for the

prediction model, it even specifies the availability of these datasets. The figure 1 specifies the distribution of dataset origin across the world wherein the review has highlighted the disproportion in the geographic provenance of publicly accessible healthcare data that have been used in machine learning (ML)-driven disease prediction. The Indian dataset only represented less than 5 percent, as only one out of the 21 datasets analysed is the APTOS 2019 Diabetic Retinopathy dataset. Most of the data used have been collected all across the world, with the largest presence held by the United States, the United Kingdom, China, Brazil, and consortiums of countries. Such an unbalanced situation highlights a huge gap in regional data availability, especially in the case of India, whose epidemiological patterns and health issues are unique. Whereas it is undeniable that the use of global datasets will prove to be informative, it can limit the generalizability and contextual applicability of ML models upstream in the Indian healthcare system.

The imbalance can be supported with a visual pie chart, the necessity of which lies in the fact that the data on Indian-origin datasets should be better represented in publicly available health data repositories

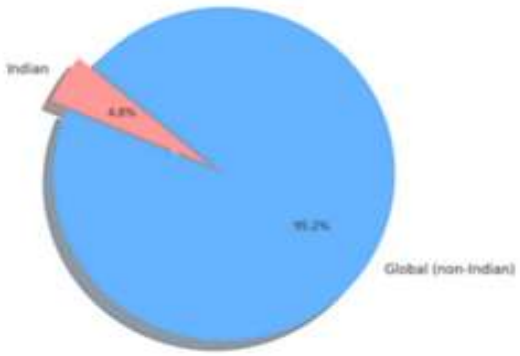

Fig 1: Distribution of Dataset Origins: Indian vs Global

6. DATA SET GAP ANALYSIS

Public datasets in healthcare are becoming more widely available, but there are still a number of significant restrictions that may prevent the creation of reliable and broadly applicable machine learning (ML) models. Four major gaps that are frequently found in the examined datasets are described in this section.

6.1 Lack of environmental and social determinants

The majority of publicly accessible datasets ignore environmental and social determinants of health (SDOH) like socioeconomic status, education, occupation, housing conditions, and air quality in favor of clinical, demographic, or biometric features. For instance, behavioral factors that have a substantial impact on cardiovascular outcomes—such as stress, physical activity, and smoking exposure—are left out of heart disease datasets like Cleveland or Faisalabad. The ecological validity of ML predictions is diminished when such contextual factors are excluded, especially when it comes to population-level risk stratification.

6.2 Absence of longitudinal patient follow-ups

The PIMA Diabetes and Cleveland Heart Disease datasets are two examples of cross-sectional datasets that only offer single time-point snapshots. This limits ML models' capacity to learn dynamics of disease progression or temporal patterns. Applications such as relapse prediction, early detection, and individualized treatment planning require longitudinal data. Although time-series vital signs are included in the MIMIC-III derivative for sepsis, most datasets do not have this level of granularity.

6.3 Underrepresentation of pediatric populations

There is a significant gap in the datasets used to represent pediatric patients. Most ML models use adult-focused datasets, frequently with an age bias toward middle-aged or older populations. The PIMA and MIMIC-III datasets, for instance, are not appropriate for pediatric predictive modeling because they primarily represent adult populations. Research in important areas where early intervention can have a significant impact, like childhood obesity, developmental disorders, or congenital diseases, is limited by the lack of pediatric-specific datasets

6.4 Data privacy and algorithmic bias issues

Data richness may be impacted by privacy-preserving techniques like de-identification or data masking that are frequently included in public datasets. Datasets may also be impacted by demographic bias, which occurs when there is a lack of diversity in the patient cohort with regard to geography, gender, or race. For example, only female Native American subjects are included in the PIMA dataset, and many other datasets are US-centric. These restrictions lessen generalizability across global health contexts and lead to biased model outputs. Furthermore, ML pipelines may exhibit hidden bias propagation as a result of opaque dataset preprocessing.

7. EVALUATION METRICS

Robust evaluation metrics are essential for assessing machine learning (ML) model performance in clinical settings. However, the quality and completeness of the underlying dataset have a significant impact on the reliability and reproducibility of these metrics.

7.1 Model generalizability

The ability of a model to function well on unknown data from various populations or environments is known as generalizability. Over fitting and limited external validity can result from datasets that are too small, lack diversity, or miss important contextual variables. When applied to larger populations, the PIMA dataset, which is solely focused on Native American women, presents generalizability issues. In test sets, models trained on such datasets might show high accuracy, but in real-world deployments across diverse patient groups, they might not perform well.

7.2 Precision, recall, and F1-score

Both class imbalance and label noise, which are prevalent in public health datasets, greatly affect these conventional classification metrics. The percentage of true positives among all predicted positives is measured by precision. The number of true positives that were accurately identified is measured by recall. The two are balanced by F1-score.These scores may be misleadingly inflated or deflated by poorly annotated or unbalanced datasets (such as datasets with a large number of negative cases compared to few sepsis-positive cases). Unbalanced data distributions in disease prediction tasks

Table 1. Mapping of Clinical/Biological Features to Disease Categories

Disease	BMI	Age	Gender	Heart Rate	BP	Cholesterol	Glucose Level	Hypertension	Vitamin D	BMD	Calcium	Stroke
Osteoporosis	✓	✓							✓	✓	✓	
Heart Disease		✓		✓	✓	✓		✓				✓
Diabetes	✓	✓					✓					
Respiratory Syncytial Virus	✓	✓		✓	✓							
Interactive Thyroid		✓	✓									
Parkinson's												
Sepsis	✓		✓									

Table 1(continued). Mapping of Clinical/Biological Features to Disease Categories

Disease	Pregnancy	Diabetes	Insulin	Family History	Body Temp.	Respiratory Rate	Thyroid Hormone Features	Tremors	Rigidity	Brady kinesia	Albumin
Osteoporosis											
Heart Disease		✓									
Diabetes	✓		✓	✓							
Respiratory Syncytial Virus				✓	✓	✓					
Interactive Thyroid							✓				
Parkinson's								✓	✓	✓	
Sepsis											✓

frequently resulting in high recall but low precision, which can cause needless alerts or interventions in clinical settings, according to Ching et al. [23].

7.3 Model drift in production

Non-representative training data, a lack of real-time updates,

or the absence of longitudinal features all contribute to model drift, which is the gradual deterioration of model performance brought on by changes in the data distribution. While many cross-sectional datasets do not allow limited drift analysis through temporal ICU records, datasets such as MIMIC-III do.

Drift detection and correction mechanisms in medical machine learning pipelines are underdeveloped, according to a recent review by Wu et al. [25]. This problem stems from dataset constraints rather than model architecture.

7.4 Benchmark references

Several studies use the same publicly available datasets to provide performance benchmarks in order to contextualize this manuscript:

MIMIC-III: Depending on the model architecture and data window, early warning systems for sepsis attain F1-scores ranging from 0.65 to 0.82.

PIMA Diabetes: Deep learning techniques perform marginally better but frequently overfit because of data sparsity, whereas logistic regression models typically produce 70–78% accuracy.

These benchmarks demonstrate how temporal consistency, feature richness, and dataset structure have a direct impact on clinical applicability in addition to performance metrics.

Table 2. Overview of disease-specific datasets used in health-related ML studies

Disease	Dataset Name	Size (records/patients)	Data Type	Geographic Coverage	Access	Key Clinical features	References
Osteoporosis	Osseous Osteoporosis Screening (DXA + phalanx sensor)	505 patients	Clinical	Brazil	Public (Zen-odo)	Age, BMI, Osseus attenuation, DXA-derived BMD, etc.	[1]
Osteoporosis	Vertebral Fracture CT (VerSe 2019)	141 patients (160 CT scans)	Imaging (CT)	Multi-center (international)	Public (OSF)	Age, lumbar spine BMD, vertebral fracture annotations (Genant grade), foreign bodies	[2]
Diabetes	Pima Indians Diabetes (UCI ML Repository)	768 subjects	Clinical	USA (Arizona Pima population)	Public (UCI)	Age, pregnancy count, plasma glucose, BP, insulin, BMI, pedigree, etc.	[3]
Diabetes	Diabetes 130-US Hospitals (UCI ML Repository)	101,766 hospital Stays	Clinical	USA (130 hospitals)	Public (UCI)	Age, gender, race, admission type, time in hospital, lab tests (glucose, A1c), diagnoses, medications, etc.	[4],[5]
Thyroid Disorders	Thyroid Disease Dataset (UCI ML Repository)	9,172 records	Clinical	Unspecified (mixed)	Public (UCI)	Age, gender, thyroid hormone levels (TSH, T4, T3), TBG, TBG uptake, possible goiter indicator, etc.	[6]
Heart Disease	Cleveland Heart Disease (UCI ML	303 subjects (270 used)	Clinical	USA (Cleve-land)	USA (Cleve-land)	Age, gender, blood pressure, cholesterol, ECG results, etc.	[7]

	Repository)						
Heart Disease	Heart Failure Clinical Records (Faisalabad)	299 patients	Clinical	Pakistan	Public (Kaggle/UCI)	Age, gender, anemia status, high BP, diabetes, ejection fraction, platelets, creatinine, sodium, etc.	[8]
Respiratory Syncytial Virus (RSV)	Chinese Respiratory Virus Surveillance (Ren et al., npj Clim)	19,161 laboratory confirmed entries (2016–2021)	Clinical (lab test results + environment)	China (31 provinces)	Restricted (CDC data)	Patient age, gender; virus type (RSV, influenza, adenovirus, etc.); time, region; meteorological factors	[9]
Influenza	Same Chinese Respiratory Virus Surveillance Dataset	19,161 (subset)	Clinical	China	Restricted	(See RSV dataset now)	[10]
COVID-19	RICORD Chest X-ray (RSNA COVID-19 Open Radiology)	998 CXR exams (361 patients)	Imaging (X-ray)	Multinational (4 centers)	Public (TCIA)	Annotations: age, sex(demographic); COVID-19 diagnostic labels; opacification patterns and airspace disease grading	[11]
Sepsis	Health Gym Synthetic Sepsis (MIMIC-III-derived)	2,164 ICU patient records	Time-series (vitals, labs)	USA (MIMIC-III)	Public (PhysioNet)	Longitudinal vitals (HR, BP, SpO, temp), lab tests, fluid/vasopressor administration (synthetic)	[12]

Table 2(continued). Overview of disease-specific datasets used in health-related ML studies

Disease	**Dataset Name**	**Size (records/patients)**	**Data Type**	**Geographic Coverage**	**Access**	**Key Clinical features**	**References**
Parkinson's Disease	Parkinson's Telemonitoring (Oxford Voice Dataset, UCI)	195 voice recordings (31 subjects)	Clinical/Signal (audio features)	UK (Oxford)	Public (UCI)	Dysphonia measures (vocal fundamental frequency, jitter, shimmer, HNR, etc.)	[13]

Parkinson's Disease	PPMI Parkinson's Pro-gression Markers Initiative)	683 subjects (423 PD, 196 HC, 64 SWEDD)	Multimodal (imaging, clinical, fluid)	International (ADNI-like consortium)	Restricted (registration)	Age, sex, clinical scores (UPDRS), MRI and DATSPECT imaging, CSF biomarkers	[14]
Alzheimer's Disease	ADNI (Alzheimer's Disease Neuroimaging Initiative)	821 (ADNI-1 phase; ∼2000 total)	Multimodal (imaging, clinical, fluid, genomic)	International (US-led)	Restricted (application)	Age, sex, cognitive scores, MRI, PET, amyloid/tau, CSF biomarkers, genetics	[15]
Depression (MDD)	REST-meta-MDD (DIRECT consortium, R-fMRI)	1,300 MDD + 1,128 HC (total 2,428 R-fMRI scans)	Imaging (resting-state fMRI)	China (25 sites)	Public (after release)	Age, sex; fMRI-derived network connectivity metrics (DMN, etc.)	[16]
Breast Cancer	Cancer Wisconsin Diagnostic Breast Cancer (WDBC, UCI)	569 patients	Clinical (FNA features)	USA (Wisconsin)	Public (UCI)	Cell nuclei features (radius, texture, perimeter, area, smoothness, etc.) extracted from FNA biopsy	[17]
Lung Cancer	LIDC-IDRI (Lung CT scans with nodule annotation)	1,018 CT studies (1,010 subjects)	Imaging (CT)	International (US institutions)	Public (TCIA)	CT images with expert-marked lung nodule boundaries and diagnostic ratings	[18],[19]
Rheumatoid Arthritis	ImmuneACCESS TCR-seq (Adaptive Biotech)	86 samples	Genomic (TCR sequences)	Public immune bank	Public (immuneACCESS)	T-cell receptor beta-chain CDR3 sequences	[19],[20]
Diabetic Retinopathy	APTOS 2019 Blindness Detection (Kaggle)	3,661 retinal images	Imaging (Fundus)	Asia (APTOS site, India)	Public (Kaggle)	DR severity grade (0–4), patient ID	[21]
Glaucoma	REFUGE Challenge (optic disc/cup segmentation)	1,200 fundus images (800 annotated)	Imaging (Fundus)	China (mult-device)	Public (TCIA)	Optic disc/cup segmentation masks, glaucoma yes/no	[22]

8. CONCLUSION

In order to present a thorough overview of publicly accessible disease datasets and their structure, the paper has restructured and extended the body of existing literature. The diseases osteoporosis, heart disease, diabetes, respiratory, syncytial virus, inter-active thyroid, Parkinson's, and sepsis have all been thoroughly

examined. Also covered are more details on publicly accessible datasets of illnesses like cancer, autoimmune disorders, and eye conditions. This paper provides information on publicly accessible datasets for the majority of diseases that can be utilized for prediction models based on machine learning or artificial intelligence. Enhancements to the datasets structure can also be made to increase the prediction model's precision and accuracy. In the present study, we undertake a rigorous comparative assessment of a diverse corpus of health-related datasets, with an eye to their clinical relevance. Our review shows that there are a few noticeable findings:

1. A pronounced set of geographic disparities characterizes the data ecosystem, especially within Indian healthcare contexts, where collection practices are often uneven and regionalized.
2. The ethnic, demographic, and clinical variance is less in the dataset landscape. In particular, cases of pediatrics and non-Western populations are underrepresented.
3. A dataset is suitable in clinical application based on its structure, its richness and profile in time. Datasets that are poorly maintained, lack longitudinal information, or provide insufficient contextualization are less likely to support models with high reliability, low bias, and minimal drift.

In sum, our evaluation articulates precisely how the availability, diversity, and stewardship of health-related datasets influence the efficacy and ethical deployment of clinical machine learning models.

REFERENCES

[1] "Osteoporosis screening using machine learning and electro-magnetic waves," PubMed Central, 2023. [Online]. Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10409756/

[2] "Vertebral Fracture CT (VerSe 2019) dataset," Semantic Scholar.[Online].Available:https://pdfs.semanticscholar.org/4dfa/65914ec9120528ea6acf1ed593fbcaf1a20b.pdfI. S. Jacobs and C. P. Bean, "Fine particles,thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] T. Wang et al., "Generalizability of machine learning models for diabetes detection: a study with Nordic Islet Transplant and PIMAdatasets," Sci.Rep., 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-87471-0

[4] M. Brown et al., "Diabetes 130-US hospitals for years 1999–2008," UCI Machine Learning Repository. [Online]. Available:https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

[5] W. Smith et al., "Pima Indians Diabetes Dataset," UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes

[6] A. Gupta and B. Keskar, "Detecting Thyroid Disease Using Optimized Machine Learning Model Based on Differential Evolution," Int. J.Comput. Intel. Syst., 2023. [Online]. Available: https://link.springer.com/article/10.1007/s44196-023-00388-2

[7] D. Detrano et al., "Cleveland Heart Disease Dataset," UCI Machine Learning Repository. [Online]. Available:https://archive.ics.uci.edu/ml/datasets/heart+disease

[8] R. Ahmed et al., "Heart Failure Survival Prediction Using Novel Transfer Learning-Based Probabilistic Features," PubMed Central,2023.[Online].Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11042000/

[9] J. Ren et al., "Development of a respiratory virus risk model with environmental data based on interpretable machine learning methods,"npj Clim. Atmos. Sci., 2025. [Online]. Available:https://www.nature.com/articles/s41612-025-00894-4

[10] J. Ren et al., "Chinese Respiratory Virus Surveillance Dataset," npj Clim.Atmos.Sci.,2025.[Online].Available: https://www.nature.com/articles/s41612-025-00894-4

[11] K. Armato et al., "MIDRC-RICORD-1C Chest X-ray Dataset," Cancer ImagingArchive.[Online].Available: https://www.cancerimagingarchive.net/collection/midrc-ricord-1c/

[12] F. Silva et al., "Synthetic Acute Hypotension and Sepsis Datasets Based on MIMIC-III and Published as Part of the Health Gym Projectv1.0.0," PhysioNet,2022.[Online].Available: https://physionet.org/content/synthetic-mimic-iii-health-gym/

[13] N. Tsanas et al., "Parkinson's Telemonitoring Voice Dataset," UCI MachineLearningRepository.[Online].Available:https://archive.ics.uci.edu/ml/datasets/parkinsons

[14] J. A. Marek et al., "The Parkinson's Progression Markers Initiative (PPMI)- establishing a PD biomarker cohort," PubMed Central, 2018.[Online].Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6292383/

[15] M. W. Weiner et al., "The Worldwide Alzheimer's Disease Neuroimaging Initiative: ADNI-3 updates and global perspectives," PubMedCentral, 2022. [Online]. Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8719344/

[16] Y. Yan et al., "The DIRECT Consortium and the REST-meta- MDD Project: Towards Neuroimaging Biomarkers of Major Depressive Disorder,"PubMedCentral,2019.[Online]. Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10917197/

[17] Y. Yan et al., "REST-meta-MDD (R-fMRI) dataset," PubMed Central, 2019.[Online].Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10917197/

[18] A. Shafique and S. Tehsin, "Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques," PubMed Central, 2022.[Online].Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9398810/

[19] A. Armato et al., "LIDC-IDRI – The Cancer Imaging Archive (TCIA)," 2015.[Online].Available: https://www.cancerimagingarchive.net/collection/lidc-idri/

[20] J. Sherwood et al., "Information on Autoimmune Disease Datasets," Sci. Rep., 2025.[Online].Available: https://www.nature.com/articles/s41598-025-88477-4/tables/3

[21] "APTOS 2019 Blindness Detection," Kaggle. [Online]. Available:https://www.kaggle.com/c/aptos2019-blindness-detection

[22] A. Chaks´ .u et al., "Chaks´ .u: A Glaucoma Specific Fundus Image Database,"Sci.Data,2023.[Online].Available: https://www.nature.com/articles/s41597-023-01943-4

[23] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones et al., "Opportunities and obstacles for deep learning in biology and medicine," J. R.Soc.Interface, vol. 15, no. 141, pp. 20170387, Mar. 2018.

[24] A. E. Johnson, T. J. Pollard, L. Shen et al., "MIMIC-III, a freely accessible critical care database," Sci. Data, vol. 3, no. 1, pp. 1–9, 2016.

[25] Y. Wu, X. Xu, and T. Zhang, "Machine learning for medical diagnosis: history, state of the art and perspective," Artif. Intell. Med., vol. 110,p.101962, Jun. 2020.