

Article

The Lab Fingerprint of HIV Comorbidities [†]

Solomon Russom ^{1,2,3}, Dimitrios Kollias ^{1,2,4,*} , Saeid Pourroostaei Ardakani ^{1,2}  and Qianni Zhang ^{1,2} 

¹ Digital Environment Research Institute, London E1 1HH, UK; s.russom@qmul.ac.uk (S.R.); s.pourroostaei@qmul.ac.uk (S.P.A.); qianni.zhang@qmul.ac.uk (Q.Z.)

² Center for Multimodal AI, Queen Mary University London, London E1 4NS, UK

³ Guy's & St Thomas' NHS Foundation Trust, London SE1 7EH, UK

⁴ Centre for Predictive In Vitro Models, Queen Mary University London, London E1 4NS, UK

* Correspondence: d.kollias@qmul.ac.uk

[†] This article is a revised and expanded version of: Solomon, R.; Kollias, D.; Zhang, Q. Proactive hiv care: Ai-based comorbidity prediction from routine ehr data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 7274–7282.

Abstract

Despite the success of antiretroviral therapy, people living with HIV remain at heightened risk of multimorbidity spanning cardiovascular, renal, hepatic, oncologic and neuropsychiatric domains. We investigate whether routinely collected electronic health record data (30 laboratory variables plus seven demographic/social descriptors) can support early, multi-label classification of recorded comorbidities in a real-world cohort of 2200 HIV-positive patients receiving continuous care at a major London hospital. We benchmark classical machine and deep learning models under two settings: a demographic-aware configuration that includes sensitive attributes (age, gender, race and continent of birth) and a demographic-unaware configuration that excludes them. XGBoost yields the best macro- F_1 performance, and demographic-aware variants consistently outperform their unaware counterparts. Permutation feature importance revealed physiologically coherent drivers (e.g., creatinine/eGFR for renal and cardiometabolic labels, hemoglobin for hematologic labels, albumin for respiratory labels) and suggested that the relative contribution of demographic variables varied across comorbidity categories. These findings indicate that (i) routinely collected EHR data contain informative patterns associated with the multi-label comorbidity profiles of people living with HIV and (ii) carefully governed use of demographic context can improve accuracy while motivating transparent consideration of fairness and bias.

Keywords: HIV; comorbidity prediction; multimorbidity; clinical decision support; demographic-aware modeling; health disparities; population health; machine learning; deep learning; XGBoost; multi-label classification; electronic health records (EHR); feature importance; explainable AI



Academic Editor: Enzo Pasquale Scilingo

Received: 25 May 2026

Revised: 16 June 2026

Accepted: 23 June 2026

Published: 27 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Human Immunodeficiency Virus (HIV) is a chronic viral infection that progressively compromises the immune system, placing patients at increased risk of developing a range of comorbid conditions, including cardiovascular diseases, renal disorders, hepatic complications, and certain cancers [1]. While antiretroviral therapy (ART) has significantly improved life expectancy, HIV patients still experience a high burden of comorbidities due to persistent immune activation, drug toxicity, and socio-behavioral factors. Early detection

and proactive management of these conditions are essential to improving long-term patient outcomes and reducing healthcare system strain.

Traditionally, the identification and management of comorbidities in HIV care rely on clinician expertise, routine screening protocols, and retrospective evaluation of medical histories. However, these approaches may fail to capture subtle risk patterns and may delay timely interventions. As such, there is growing interest in leveraging data-driven methodologies to improve risk stratification and clinical decision-making.

Artificial Intelligence (AI), particularly through its subfields of machine learning (ML) and deep learning (DL), offers promising tools for analyzing complex healthcare datasets. ML and DL algorithms excel in processing high-dimensional, heterogeneous data to uncover nonlinear associations and predictive markers that may not be easily identifiable through conventional analysis. DL in particular has the advantage of learning nonlinear feature representations from structured clinical variables and may capture interactions among various features that are difficult to specify manually.

These technologies have demonstrated success in various clinical domains, from predicting disease onset to guiding treatment decisions [2–4]. In the context of HIV care, the integration of ML and DL with routinely collected data from electronic health records (EHRs) presents an opportunity to enhance early identification of comorbid diseases.

The potential impact of employing an ML/DL system to predict comorbid diseases in HIV patients using routinely collected blood test and demographic data is substantial. Such a system could transform current care pathways by enabling earlier risk detection, facilitating timely intervention, and supporting more personalised management strategies. Early identification of comorbid conditions is crucial in the context of HIV, as many of these diseases—such as liver and renal dysfunction, cardiovascular complications, and certain cancers—often progress silently and are only diagnosed at advanced stages, when treatment options become limited and outcomes poorer.

Currently, monitoring for comorbidities in HIV patients often relies on periodic evaluations and fixed clinical protocols, which may not adequately account for individual variation in risk or disease progression. As a result, high-risk individuals may remain undetected for extended periods. An ML/DL system trained on a hospital's EHR data could continuously analyze patterns in routine lab tests and social-demographic attributes to flag patients likely to develop specific conditions, long before clinical symptoms manifest. This would allow clinicians to prioritize patients for further screening, preventative therapies, or closer monitoring without waiting for overt clinical signs.

Moreover, ML/DL systems can process high volumes of patient data simultaneously and in real time, offering scalable tools for risk stratification across large patient populations. They can also uncover subtle, non-linear patterns in lab values that might be overlooked in traditional rule-based assessments. By integrating demographic and social factors, such systems can further personalize risk estimates, enhancing clinical decision-making.

Ultimately, the goal is to support a shift from reactive to proactive HIV care, where comorbidities are anticipated and managed before they escalate. This would not only improve patient safety and long-term health outcomes but also reduce unnecessary tests, hospital admissions, and healthcare expenditures. Efficient identification of high-risk patients allows better allocation of clinical resources, ensuring that specialist time and hospital capacity are directed where they are most needed.

The description above represents the longer-term clinical motivation for developing AI-supported systems in HIV care. The present study does not evaluate prospective forecasting of newly occurring comorbidities. Instead, the present study constitutes a retrospective feasibility analysis investigating whether various ML and DL models can classify existing multi-label comorbidity profiles recorded for individual patients using

structured EHR data, including routinely collected laboratory, demographic and social variables. Let us note that the DL models are used as structured-data classifiers rather than temporal or representation-learning models for raw EHR streams.

The source data for this study comes from two EHR systems within a major London hospital: the general hospital-wide Electronic Health Recording System and the specialty EHR system of the Sexual Health and HIV Services Department. Our dataset comprises laboratory blood test results with 30 microbiology and biochemistry features, and an additional 7 demographic and social features (including age, gender, ethnicity, and behavioral risk indicators) from around 2200 patients. The dataset also comprises disease diagnoses of these HIV patients extracted based on entries in the International Classification of Diseases (ICD) coding system, with hierarchical mapping from high-level categories to specific diagnostic entities.

We apply a range of classical ML and specialized DL models, including Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), LightGBM (LGBM), Artificial Neural Network (ANN) and Multilayer Perceptron (MLP) and TabNet. Each model is evaluated under two experimental settings: a *demographic-aware* approach, where demographic and social variables are included alongside laboratory data, and a *demographic-unaware* approach, which uses only laboratory and social variables. This design allows us to examine not only predictive performance but also the role and influence of demographic information in clinical risk modeling. The goal is to determine the effectiveness of these ML and DL models in predicting the presence of specific comorbidities in HIV patients.

Additionally, we adopt permutation feature importance to quantify how much each input contributes to predictive performance. Importantly, we interpret the magnitude of importance values, not just their rank. When a top feature's importance is substantially larger than the next (e.g., hemoglobin dominating hematologic labels), the model relies primarily on that single clinical marker; the remaining variables act as secondary context. Conversely, when several features have similar importance (e.g., infectious disease labels), the model leverages a multifactorial signal combining virologic control, hepatic function, hematologic status, and demographics. We also observe label-dependent roles for demographics: they carry decisive signal for mental health, neurologic, and musculoskeletal outcomes but are less influential for generic utilization codes (Z00–Z99), aligning with clinical intuition and epidemiology.

This work makes three primary contributions:

- Retrospective multi-label benchmarking in HIV care. We formulate recorded HIV comorbidity status as a multi-label classification problem and benchmark several classical machine learning and deep learning models on real-world EHR data from 2200 PLHIV (spanning 12 ICD-10 blocks) under a unified evaluation protocol.
- Assessment of demographic context through predictive performance and model interpretation. We compare models trained with all available demographic variables against models excluding demographic attributes and examine the resulting changes in predictive performance. For the best performing model (XGBoost), we use permutation feature importance to provide disease-specific interpretation profiles that recover clinically plausible markers (e.g., eGFR/creatinine for renal and cardiovascular labels; hemoglobin for hematologic labels; albumin for respiratory labels) while highlighting comorbidity categories in which demographic variables exhibit relatively higher importance (e.g., mental health, neurological, and musculoskeletal categories).
- Preliminary clinical interpretation and future research directions. We discuss the clinical plausibility of the identified label-specific drivers, the potential role of demographic information in HIV comorbidity modeling and the implications of these

findings for future clinical decision-support systems, fairness-aware evaluation and external validation studies.

The remainder of the paper is organized as follows. Section 2 reviews related work in ML/DL for clinical prediction and HIV comorbidity modeling. Section 3 details the materials and methods, including cohort construction, problem formulation, model architectures, metrics and implementation. Section 4 presents experimental results, comparing demographic-aware and demographic-unaware variants and analyzing permutation feature importance per ICD-10 block. Section 5 discusses clinical implications and fairness considerations. Section 6 concludes, discusses limitations and outlines directions for future research.

2. Related Work

The increasing availability of electronic health records (EHRs) has supported the development of machine learning (ML) and deep learning (DL) methods for clinical prediction, risk stratification and decision support. Large-scale studies have demonstrated that heterogeneous EHR data can be used to predict outcomes such as mortality, hospital readmission, prolonged length of stay and future diagnoses [5]. Reviews of the broader literature similarly highlight the potential of ML and DL for disease-onset prediction and longitudinal clinical forecasting using routinely collected health data [6,7]. The authors in [6], for example, found that laboratory measurements, demographic characteristics, diagnoses, medications, symptoms and clinical notes were commonly used as predictive variables. However, much of the reviewed research remained at the retrospective development or prototyping stage, with limited prospective evaluation and clinical implementation.

Within infectious disease research, ML has been applied extensively to early warning, prognosis and treatment outcome assessment. Representative applications include sepsis prediction from EHR data [8], COVID-19 mortality risk prediction [9,10] and tuberculosis treatment outcome prediction [11,12]. These studies demonstrate that laboratory measurements, demographic characteristics, clinical observations and treatment histories can reveal complex nonlinear patterns associated with adverse outcomes. Nevertheless, broad comorbidity and multimorbidity modeling in long-term HIV care has received comparatively less attention.

Recent systematic reviews provide a more focused overview of AI applications across the HIV care continuum. The authors in [13] reviewed 47 studies covering HIV testing, retention in care, virological response, clinical and immunological outcomes, and testing and treatment support. The reviewed methods included random forests, support vector machines, neural networks and other deep learning architectures. The authors also identified persistent challenges involving data quality, population representativeness, infrastructure, ethical governance and integration into routine healthcare systems. Similarly, the authors in [14] analyzed 98 studies using real-world clinical data and identified six principal application areas: comorbidity assessment, HIV drug resistance prediction, monitoring of HIV infection, treatment outcome prediction, treatment adherence assessment and treatment recommendation. Their findings demonstrated the expanding use of clinical and biological data in HIV-related ML research, while also showing that clinically validated decision-support systems, adherence interventions and treatment optimization tools remain underdeveloped.

Predictive modeling is particularly relevant for people living with HIV (PLHIV), as antiretroviral therapy has transformed HIV into a chronic condition while the risks of cardiovascular, renal, hepatic, metabolic, oncological, neurological and mental health comorbidities remain elevated.

The work in [15] is the closest prior study to our work because it used EHR data and ML to investigate comorbidity burden among 8253 people living with HIV in South Carolina. Specifically, the authors used 35 predictors from multiple domains and developed a logistic LASSO regression model to predict changes in non-AIDS comorbidity burden following HIV diagnosis. Comorbidity burden was quantified using the Charlson Comorbidity Index (CCI), and the outcome was formulated as a binary task indicating whether the CCI score increased or remained unchanged. In contrast, our study formulates recorded HIV comorbidity status as a multi-label classification problem across 12 ICD-10 diagnostic categories. This distinction is important because a comorbidity burden score provides a summary measure of overall disease burden, whereas multi-label classification preserves disease-specific information, enables simultaneous modeling of multiple comorbidities, and supports label-wise interpretation of model behavior.

Other studies have concentrated on individual clinical outcomes. The authors in [16] developed cohort-derived ML models for forecasting the onset of chronic kidney disease in PLHIV, while in [17], authors proposed an ML-based model for coronary heart disease risk assessment using electronic medical records. More recently, cardiovascular disease risk modeling has been investigated in older HIV/AIDS populations, with gradient-boosting approaches such as LightGBM demonstrating promising performance [18]. Table 1 presents a detailed comparison between the above-described prior HIV-related comorbidity prediction studies and our work.

Although this literature confirms the potential value of ML in HIV care, several methodological gaps remain. First, most studies focus on a single outcome, such as cardiovascular disease, renal dysfunction, treatment failure, viral suppression, mortality or overall comorbidity burden. Such formulations do not fully represent the multimorbid nature of HIV care, in which several conditions may occur simultaneously and may share biological, behavioral and demographic risk factors. Multi-label classification provides a suitable alternative because it allows multiple ICD-based disease categories to be modeled jointly rather than treating each condition as an independent endpoint.

Second, prior HIV-related ML studies commonly include demographic variables (such as age, sex, ethnicity) or route of infection as standard predictors, but they usually do not explicitly compare models trained with and without core demographic attributes. More broadly, fairness research in clinical prediction has shown that sensitive attributes may influence model performance and may also contribute to subgroup disparities, motivating subgroup-level evaluation and fairness-aware modeling [19,20]. However, these studies address fairness assessment or mitigation more directly, whereas our work focuses on quantifying the predictive contribution of demographic context in HIV comorbidity classification. We therefore distinguish our demographic-aware versus demographic-unaware comparison from a full algorithmic fairness audit.

Third, interpretability remains essential for clinical acceptance. Although gradient-boosting methods and neural networks can capture complex feature interactions, clinicians must be able to determine whether predictions are based on medically plausible laboratory and demographic patterns. Previous HIV-related studies have applied feature-importance and explainability techniques to selected outcomes, but disease-specific interpretation across a broad set of simultaneously modeled comorbidity categories remains limited.

To address these gaps, we formulate HIV comorbidity prediction as a multi-label classification problem covering 12 ICD-10 diagnostic blocks. We benchmark several classical ML and DL methods using routine laboratory measurements and demographic and social descriptors from approximately 2200 PLHIV. We also compare demographic-aware and demographic-unaware configurations to quantify the contribution of demographic context and examine its potential fairness implications. Finally, permutation feature importance

is used to identify the laboratory, demographic and social variables associated with each comorbidity category. The resulting framework provides a broader and more interpretable evaluation of ML-based HIV comorbidity modeling than approaches focused on isolated clinical outcomes.

Table 1. Comparison with prior HIV-related comorbidity prediction studies. CCI = Charlson Comorbidity Index; CKD = Chronic Kidney Disease; CHD = Coronary Heart Disease; CVD = Cardiovascular Disease; ICD = International Classification of Diseases.

Study	Cohort Size	Outcome/Labels	Demographics	Metrics
[15]	8253	CCI increase/no increase	Yes	Accuracy
[16]	12,761	CKD onset	Yes	ROC-AUC
[17]	Not reported	CHD risk	Yes	AUC
[18]	1916	CVD occurrence	Yes	F1
This	2200	12 ICD-10 labels	Yes & No	F1

3. Materials and Methods

3.1. Problem Statement

We define the dataset as a collection of N samples $(\mathbf{x}_i, \mathbf{y}_i)$, where $i = 1..N$, and each feature vector $\mathbf{x}_i \in \mathbb{R}^d$ represents patient-specific inputs, and the target $\mathbf{y}_i \in \{0, 1\}^m$ is a multi-hot vector indicating the presence or absence of m distinct comorbidities, derived from ICD-based diagnostic codes. In this study, $\mathbf{x}_i$ primarily consists of standard laboratory examinations, such as microbiological and biochemical markers optionally supplemented with demographic and social information.

Each instance is further associated with a set of sensitive descriptors, denoted $\mathbf{s}_i = \langle s_{\text{Age}}, s_{\text{Gender}}, s_{\text{Race}}, s_{\text{Continent}} \rangle$. These variables capture protected characteristics:

- $s_{\text{Age}} \in \{21\text{--}40, 41\text{--}60, 61\text{--}80, 81\text{--}100\}$
- $s_{\text{Gender}} \in \{\text{Male, Female, Other}\}$
- $s_{\text{Race}} \in \{\text{White, Black, Asian, Mixed, Other}\}$
- $s_{\text{Continent}} \in \{\text{Europe, Africa, Asia, South America, North America, Oceania}\}$

The categorization of race and continent follows the classification schemes adopted by the U.S. Census Bureau and aligns with conventions used by NHS England and the UK Office for National Statistics. Since these attributes correspond to legally and ethically sensitive factors, we treat them as protected variables that should not unjustly influence predictions.

To analyze their role, we construct two experimental settings. The first is a demographic-aware model, which estimates $p(\mathbf{y}_i | \mathbf{x}_i, \mathbf{s}_i)$ by explicitly incorporating both laboratory measurements and sensitive descriptors (i.e., demographic and social features). The second is a demographic-unaware model, which learns $p(\mathbf{y}_i | \mathbf{x}_i)$ using only non-sensitive laboratory and social features (i.e., it excludes the demographic attributes Age, Gender, Race and Continent of Birth). Comparing these two approaches enables us to investigate whether demographic attributes enhance predictive performance or risk embedding bias into the comorbidity predictions for HIV populations.

3.2. Machine Learning Models

We investigated a range of ML and DL methods to predict comorbidities in HIV patients using structured laboratory, demographic and social data. The task was formulated as a multi-label classification problem, where each patient could present with multiple comorbid conditions simultaneously. The evaluated models include Logistic Regression (LR) [21], Random Forest (RF) [22], XGBoost (XGB) [23], LightGBM (LGBM) [24], Artificial

Neural Networks (ANNs)/Multilayer Perceptrons (MLPs), and the TabNet architecture [25]. These models span both classical ML baselines and state-of-the-art DL methods, providing a comprehensive framework for comparison.

Logistic Regression (LR) : As a linear baseline, LR models the log-odds of a binary outcome as a weighted sum of input features, producing probabilistic predictions through the logistic function. In multi-label settings, we adopt the binary relevance strategy, training one independent classifier per comorbidity. LR is valued for its simplicity, interpretability and robustness in linearly separable cases, offering clinically meaningful feature coefficients that highlight the contribution of individual variables.

Random Forest (RF): RF is an ensemble method that aggregates predictions from multiple decision trees, each trained on bootstrapped samples and random feature subsets. For multi-label classification, we extend RF by fitting one model per outcome. Its strengths lie in capturing nonlinear dependencies, handling high-dimensional structured data and providing feature importance scores, which aid in understanding the clinical drivers of comorbidity risk.

Gradient Boosting Models: XGBoost (XGB) and LightGBM (LGBM): Both XGB and LGBM implement gradient boosting, sequentially refining decision trees to minimize prediction errors. While XGB emphasizes regularization to reduce overfitting, LGBM introduces histogram-based optimization and leaf-wise growth for faster training and improved scalability. Applied with a one-vs-rest strategy, these models excel at handling missing values, capturing complex feature interactions and achieving strong performance on large-scale EHR datasets.

Neural Networks: ANN/MLP: ANNs and MLPs extend predictive capacity beyond tree-based models by learning hierarchical feature representations. We employ feedforward architectures with multiple hidden layers, ReLU/LeakyReLU activations, and dropout regularization. Training is performed with backpropagation and binary cross-entropy loss applied to each label. These networks are particularly effective in modeling subtle and nonlinear patterns in laboratory values that may not be detected by classical approaches.

TabNet: TabNet is a purpose-built DL model for tabular data that incorporates sequential attention to select informative features during training. Unlike traditional MLPs, TabNet performs instance-wise feature selection, improving generalization and interpretability. In our framework, TabNet is extended to multi-label classification using a multi-output wrapper. Its combination of competitive accuracy and inherent explainability makes it especially suitable for heterogeneous clinical datasets.

Together, these models enable systematic comparison between interpretable linear methods, flexible ensemble learners and advanced DL approaches. By training both demographic-aware and demographic-unaware variants, we further assess how sensitive patient attributes (e.g., age, gender, race, continent of birth) influence predictive performance and fairness in comorbidity modeling.

3.3. Dataset

Source and Cohort: The study draws on longitudinal data from South East London Hospital, covering HIV-positive patients in continuous care between 2012 and 2023. Each clinical record corresponds to a specific patient encounter and contains the laboratory measurements available at that time. Consequently, a patient may contribute multiple records to the dataset. Each record was treated as an independent sample during model development. Records were integrated from two complementary sources: the hospital-wide electronic health record (EHR) system and a specialized EHR system maintained by the Sexual Health and HIV Services Department. The latter provides detailed documentation of outpatient HIV consultations, including clinical, laboratory, and social information, while

the hospital-wide EHR captures diagnoses and treatments from other specialties. Together, these systems offer a complete view of each patient's healthcare trajectory.

From these records, we identified a cohort of approximately 2200 individuals who regularly attended HIV outpatient services during the study period and for whom complete records were available. Each record comprises the patient's aggregated laboratory measurements, demographic and social descriptors and multi-label comorbidity status derived from ICD-10 diagnostic codes. The laboratory variables and diagnostic records were obtained from the same observation period. For each patient, the dataset contains 30 laboratory-based variables and 7 demographic/social descriptors, complemented by ICD-coded comorbidity diagnoses.

Input Features: We categorize predictive features into two main groups:

- *Laboratory Features (30 continuous variables):* Standard microbiology and biochemistry markers routinely collected during HIV care. These include liver and kidney function measures (e.g., AST, ALT, creatinine, eGFR), hematological indices (e.g., hemoglobin, RBC, WBC, platelet counts), immune-related parameters (e.g., lymphocytes, neutrophils, globulin), and viral load measurements. As detailed in Table 2, each feature is accompanied by a clinical description and its standard reference range. Table 3 depicts descriptive statistics across the cohort for each feature. These markers are clinically significant, as many comorbidities in HIV patients first manifest through abnormal biochemical profiles. These variables are included in both the demographic-aware and demographic-unaware approaches.
- *Demographic and Social Features (7 categorical variables):* Patient-level descriptors include the: (i) demographic features, i.e., Age Group, Gender, Race and Continent of Birth (these four variables are included only in the demographic-aware approaches); and (ii) Social Features, i.e., Sexual Orientation, Commercial Sex Worker status and Route of HIV Infection (these three variables are included in both the demographic-aware and demographic-unaware approaches). The distributions of these variables are summarized in Table 4. They capture both biological and socio-behavioral risk factors that may influence comorbidity profiles.

Table 2. Microbiology and Biochemistry Input Features Used for Comorbidity Prediction. These variables constitute the laboratory feature set and are included in both the demographic-aware and demographic-unaware approaches.

Feature	Description	Normal Range
%Hypo	Low blood sugar level	4–6 mmol/L
Adjusted Calcium	Calcium level corrected for albumin	2.2–2.6 mmol/L
Albumin	Protein made by liver marker of liver function	3.4–5.4 g/dL
ALP	Enzyme linked to liver or bone conditions	44–147 IU/L
AST	Enzyme elevated in liver/muscle damage	8–33 U/L
Bilirubin Total	Byproduct of red blood cell breakdown	3–20 µmol/L
Creatinine	Waste product indicating kidney function	45–120 µmol/L
Eosinophils	Immune cells linked to allergies/infections	0–0.4 × 10 ⁹ /L
eGFR	Kidney filtration estimate	>60 mL/min
GGT	Liver enzyme marking bile duct/liver damage	5–40 U/L
Globulin	Immune-related proteins	20–35 g/L
Hemoglobin (Hb)	Oxygen-carrying red cell protein	12–18 g/dL
HIV RNA Viral Load	Amount of HIV RNA in blood	50–Millions
Lymphocytes	Key immune cells	1000–4800/µL
MCH	Avg. hemoglobin per red cell	27–31 pg
MCHC	Avg. hemoglobin concentration	320–360 g/L
MCV	Avg. size of red blood cells	80–100 fL
Monocytes	Immune cells against infection	2–8%
MPV	Average size of platelets	7–9 fL

Table 2. *Cont.*

Feature	Description	Normal Range
Neutrophils	First response immune cells	2500–7000/ μ L
Platelets (PLT)	Blood clotting cells	150K–450K/ μ L
Potassium	Electrolyte for nerve/muscle function	3.6–5.2 mmol/L
Total Protein	Total proteins in blood	60–83 g/L
Basophils	Immune cells for allergic responses	0–1%
Phosphate	Mineral for bone and energy function	2.5–4.5 mg/dL
PCV	Proportion of red blood cells in blood	35.5–48.6%
RBC	Red blood cell count	3.92–5.65 M/ μ L
RDW	Variation in red blood cell size	12–15%
Sodium	Electrolyte for fluid/nervous balance	135–145 mEq/L
WBC	White blood cell count	4.5K–11K/ μ L

Table 3. Summary Statistics of Microbiology and Biochemistry Features.

Input Feature	Min	Average	Max
Hypo%	0	10.88	76.88
Adjusted Calcium	1.56	2.24	2.73
Albumin	21.5	40.32	53
Alkaline Phosphatase	26.67	106.01	2694.6
Aspartate Transaminase	11.14	37.32	760.56
Bilirubin Total	1.27	10.95	376.4
Creatinine	24.85	119.6	1572.1
Eosinophils	0	0.17	1.95
Estimated GFR	3	73.95	90
Gamma-glutamyl Transferase	2	79.81	3428.25
Globulin	13	34.38	93.58
Hemoglobin (Hb)	63.14	124.89	181.5
HIV RNA Viral Load	20	41,424.12	7,731,687
Lymphocytes	0.22	1.85	9.67
MCH	15.68	29.87	42.15
MCHC	256	317.53	366.5
MCV	55.68	93.97	127.85
Monocytes	0.07	0.38	1.53
MPV	5.5	8.67	15.97
Neutrophils	0.22	5.55	21.92
Platelets (PLT)	8.33	233.85	650
Potassium	3.09	4.33	7.2
Total Protein	39	74.71	123.92
Basophils	0	0.03	0.28
Phosphate	0.42	1.06	2.51
PCV	0.2	0.39	0.7
RBC	1.97	4.2	6.32
RDW	11.05	14.85	25.52
Sodium	126.95	138.47	151.02
WBC	1.16	6.11	16.95

Comorbidity Labels: Disease outcomes were derived from structured ICD-10 codes. Each patient may be associated with multiple diagnostic categories, yielding a multi-label classification problem. To balance granularity and interpretability, ICD codes were aggregated into 12 high-level categories (Table 5), spanning broad domains such as infectious diseases, circulatory and digestive disorders, endocrine and metabolic conditions and mental health. The most prevalent category—“Factors influencing health status and contact with health services” (Z00–Z99)—was observed in 87% of patients, while infectious and

circulatory system diseases affected 46% and 41% of the cohort, respectively. These labels capture the systemic and multi-morbid nature of HIV care.

Table 4. Distribution (in %) of Demographic and Social Features. The Social ones are included in both the demographic-aware and demographic-unaware approaches, whereas the Demographic ones are only included in the demographic-aware approach.

Demographic or Social Feature	Values	Distributions
Age	41–60	63.13
	61–80	21.01
	21–40	14.15
	81–100	1.72
Gender	Male	56.49
	Female	43.46
	Other	0.05
Race	Black	53.57
	White	33.77
	Other	6.86
	Mixed	4.73
	Asian	1.07
Continent of Birth	Africa	42.92
	Europe	38.87
	South America	5.34
	Asia	5.32
	Oceania	4.79
	North America	3.76
Sexual Orientation	Heterosexual	54.92
	Homosexual	23.42
	Other	19.38
	Bisexual	2.27
Commercial Sex Worker	No	90.63
	Current	8.81
	Previous	0.56
Route of Infection	MSW (Male-Female Sex)	61.87
	MSM (Male-Male Sex)	29.41
	PWID (Injected Drug Use)	3.11
	MTCT (Mother to Child)	2.64
	CWBP (Blood products)	1.56
	EVHC (Healthcare Exposure)	1.41

Evaluation Strategy: To ensure robust performance assessment, we adopted iterative stratified k-fold cross-validation [26], which is specifically tailored for multi-label tasks. This procedure preserves the relative frequency of each comorbidity label across folds, preventing label imbalance and yielding fair comparisons across different ML and DL models. To avoid patient-level data leakage, all cross-validation splits were performed at the patient level. That is, all records and labels belonging to the same patient were assigned exclusively to either the training or test fold within each cross-validation iteration, and no patient appeared in both sets. This ensures that model performance reflects generalization to unseen patients rather than memorization of patient-specific laboratory or demographic patterns.

Table 5. Distribution of High-Level ICD Diagnostic Categories Among HIV Patients.

ICD-10 Category	Proportion (%)
Z00–Z99: Factors influencing health status and contact with health services	87%
A00–B99: Certain infectious and parasitic diseases	46%
I00–I99: Diseases of the circulatory system	41%
K00–K95: Diseases of the digestive system	37%
E00–E89: Endocrine, nutritional and metabolic diseases	29%
F01–F99: Mental, behavioral and neurodevelopmental disorders	28%
N00–N99: Diseases of the genitourinary system	28%
J00–J99: Diseases of the respiratory system	27%
M00–M99: Diseases of the musculoskeletal system and connective tissue	20%
D50–D89: Blood and immune-related diseases	19%
G00–G99: Diseases of the nervous system	15%
C00–D49: Neoplasms	12%

3.4. Metrics

In this study, we evaluate the performance of our ML and DL models using metrics specifically suited to the multi-label classification nature of our comorbidity prediction task. Additionally, we employ permutation feature importance to quantify the contribution of each input feature to the models' predictive performance.

For multi-label classification problems, the standard performance metric we adopt is the macro F_1 score [27]. Macro- F_1 was selected as the primary evaluation metric because the dataset exhibits substantial label imbalance and our objective was to evaluate performance across all comorbidity categories equally, irrespective of prevalence. Unlike micro-averaged metrics, which are dominated by more prevalent labels, Macro- F_1 assigns equal weight to each comorbidity category and is therefore widely used in imbalanced multi-label classification settings. This metric is computed independently for each label and then averaged, treating all comorbid conditions equally regardless of their frequency in the dataset. The F_1 score itself is the harmonic mean of precision and recall, where precision quantifies the proportion of correctly predicted positive cases among all predicted positives and recall reflects the proportion of true positives that were correctly identified. The F_1 score ranges from 0 to 1, with higher values indicating better performance.

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

To account for all labels (i.e., multiple comorbid conditions), we compute the macro-averaged F_1 score as follows:

$$\mathcal{P} = \frac{1}{m} \sum_{j=1}^m F_1^j \quad (2)$$

where m is the number of comorbidity labels and F_1^j is the F_1 score computed for the j th label. This metric is especially useful in our setting, as it ensures that rare and common comorbidities are weighted equally in the overall performance evaluation.

In addition to classification performance, we apply permutation feature importance [28] to estimate the relative importance of each input variable (e.g., lab test result or demographic attribute) to the model's predictions. This model-agnostic method helps identify features that contribute to model performance and supports model interpretability. Permutation feature importance should be interpreted as a descriptive measure of model reliance on a feature rather than a formal statistical test of feature significance or causality.

The procedure involves the following steps: (i) after training the model, one feature is randomly permuted across all instances, breaking its relationship with the output labels; (ii) the model is then re-evaluated using the same performance metric (macro F_1); (iii) the

drop in performance, if any, quantifies how much the model relied on that feature; and (iv) the process is repeated for each feature. A larger decrease in the F_1 score indicates a more important feature, whereas a minimal change suggests low relevance to the model.

This approach enables us to analyze and compare the predictive value of both clinical (lab-based) and demographic and social features, guiding the refinement of input selection and improving model transparency.

3.5. Pre-Processing and Implementation Details

In terms of pre-processing, we applied min–max normalization [29] to each input variable independently. In terms of implementation details, we experimented with different hyperparameters for the ML and DL models.

In the case of RF, we used an ensemble of 10 decision trees. Each tree was grown to full depth unless all leaves became pure (based on Gini impurity) or contained one or no samples. The minimum number of samples required at a leaf node was set to one. For LR, we included a bias term in the decision function and applied L2 regularization to prevent overfitting. The inverse regularization strength was set to 1.0, and the optimization stopped when the tolerance for the stopping criterion reached 0.0001. LR was implemented independently for each label using the one-vs-rest strategy, a standard approach in multi-label classification tasks. In the case of XGB and LGBM, both models were configured with default tree-based boosting parameters unless otherwise stated. These included early stopping criteria and learning rate adjustments, tailored to multi-label performance using binary relevance wrappers.

For the MLP and ANN, we implemented custom architectures consisting of multiple fully connected layers with ReLU or LeakyReLU activations, dropout for regularization, and sigmoid output activation to handle the multi-label classification setting. The models were trained using binary cross-entropy loss and optimized using the Adam optimizer (we evaluated learning rates in range $[1 \times 10^{-4}, 1 \times 10^{-2}]$, batch sizes in $\{16, 32, 64, 128\}$, and L2 weight decay values in range $[0, 1 \times 10^{-4}]$). Training was performed for up to 100 epochs, with early stopping based on validation macro- F_1 to reduce overfitting. We evaluated several architectures, including deeper models with skip connections, and selected the best performing variant based on macro F_1 . For TabNet, we used the PyTorch TabNet implementation adapted for multi-label classification with sigmoid outputs and binary cross-entropy loss. We tuned the number of decision steps, feature transformation dimensions, attention dimensions, sparsity regularization coefficient, learning rate and batch size. Candidate values included decision steps in $\{3, 5, 7\}$, attention/decision dimensions in $\{8, 16, 32\}$, learning rates in range $[1 \times 10^{-4}, 1 \times 10^{-2}]$, batch sizes in $\{64, 128, 256\}$ and sparsity regularization values in range $[1 \times 10^{-5}, 1 \times 10^{-3}]$. Early stopping was again applied using validation macro- F_1 .

Hyperparameter tuning was performed using nested cross-validation. Specifically, we used an outer 5-fold patient-level cross-validation loop for final model evaluation and an inner 3-fold cross-validation loop within each training split for hyperparameter selection. The random seed was fixed to 42 for reproducibility.

Statistical comparisons were performed using paired t-tests. All presented results showed that the best performing models achieved statistically significant improvements over the compared models (2nd and 3rd best performing ones), with p -values below 0.05.

In the case of permutation feature importance, we applied a model-agnostic approach to assess the impact of each input feature. Each feature was permuted 10 times, and the resulting drop in macro F_1 score was used to quantify its importance. All model implementations, training, and evaluations were conducted using the scikit-learn [30] and PyTorch libraries within the Python environment.

4. Experimental Results

In the following, we present a comprehensive experimental evaluation in which we compare the performance of all employed ML and DL models in predicting comorbidities among HIV patients, using the macro F_1 score as the evaluation metric. We assess each model under two settings: the demographic-aware approach, where both laboratory, demographic and social features are included as inputs, and the demographic-unaware approach, which relies solely on laboratory and social features. This comparison allows us to evaluate the contribution of demographic information to model performance. Finally, we report the results of the permutation feature importance analysis for each model under both configurations, providing insights into the most influential clinical and demographic features driving model predictions.

4.1. Performance Comparison Across Demographic-(Un)Aware ML and DL Models

Table 6 presents a performance comparison in terms of the overall F_1 Score (shown in %), averaged across all folds, between the ML and DL methods (described in Section 3), under the two studied settings (demographic-aware vs demographic-unaware approach).

Table 6. Macro F_1 Score (%) comparison between demographic-aware and demographic-unaware approaches across ML and DL models. Best scores are in **bold**.

Model	Demographic-Aware	Demographic-Unaware
LR	37.7	33.9
RF	40.7	38.7
XGB	45.8	43.5
LGBM	44.7	43.0
MLP	44.5	40.8
TabNet	32.6	28.0

These results demonstrate that among all evaluated models, XGBoost (XGB) achieved the highest overall macro F_1 score, establishing itself as the best performing model in both demographic-aware vs demographic-unaware approaches, while TabNet yielded the lowest performance in both demographic-aware vs demographic-unaware approaches. This outcome aligns with expectations, as XGB is well known for its robustness in handling imbalanced datasets, capturing complex feature interactions, and offering strong generalization, particularly in structured medical data. In contrast, TabNet, although designed for tabular data, often requires extensive hyperparameter tuning and larger datasets to realize its full potential, which may explain its underperformance in our setting. Logistic Regression (LR) and Random Forest (RF) showed moderate results (being the second and third worst methods in both demographic-aware and demographic-unaware approaches). RF outperformed LR due to its capacity to model non-linearities and feature dependencies. The LGBM performed strongly, being the second best performing method in both the demographic-unaware and demographic-aware approaches (closely matching XGB). The MLP performed strongly, being the third best performing method in both the demographic-unaware and demographic-aware approaches.

Our findings consistently show that the demographic-aware approach outperformed the demographic-unaware approach across all models. This highlights the value of incorporating demographic features, such as age group, gender identity, race and continent of birth into the prediction pipeline. These variables are likely to capture underlying biological factors that influence the manifestation and progression of comorbidities in HIV patients. For example, certain comorbidities may be more prevalent in specific age groups due to behavioral, clinical, or socioeconomic patterns. By excluding these variables,

the demographic-unaware models lose potentially critical context, resulting in lower performance. This performance gap was particularly noticeable in models like LR, MLP and TabNet, where the inclusion of demographic information substantially boosted macro F_1 scores. Even in high-performing models like XGB and LGBM, the demographic-aware version consistently outperformed the unaware variant. This suggests that demographic variables not only offer complementary information to clinical data but may also help the model better generalize, especially for less prevalent comorbidities. Therefore, while there may be ethical considerations in how sensitive demographic data are used, from a technical standpoint, our results demonstrate that demographic awareness enhances predictive accuracy and reliability in multi-label comorbidity classification for HIV patients.

Table 7 presents a performance comparison in terms of the F_1 Score (shown in %) per each comorbidity (i.e., per class), averaged across all folds, between the ML and DL methods, under the two studied settings (demographic-aware vs. demographic-unaware approach).

Table 7. Mean F_1 score (%) per ICD-10 block between demographic-aware (denoted as ‘-A’) and demographic-unaware (denoted as ‘-U’) approaches across ML and DL models. Best scores are in bold.

Model	A00–B99	C00–D49	D50–D89	E00–E89	F01–F99	G00–G99	M00–M99	I00–I99	J00–J99	K00–K95	N00–N99	Z00–Z99
LR-A	56.8	0.0	38.5	30.2	37.2	1.1	7.1	60.5	34.5	39.2	54.0	92.8
LR-U	56.8	0.0	37.8	23.3	18.4	0.0	3.0	53.6	31.3	38.4	51.6	92.8
RF-A	60.0	1.5	42.6	35.7	34.3	5.0	11.8	63.1	39.1	45.3	56.6	92.8
RF-U	58.7	1.0	42.3	33.1	25.6	3.6	10.3	59.7	37.3	44.2	55.8	92.8
XGB-A	61.4	13.8	44.0	44.8	41.4	13.1	23.3	63.6	48.0	48.1	57.0	91.3
XGB-U	60.3	9.8	41.9	41.3	38.1	11.0	22.5	61.1	44.7	46.9	53.2	91.3
LGBM-A	61.0	8.6	42.4	45.4	43.5	12.7	20.1	61.9	45.9	48.9	54.0	91.4
LGBM-U	59.5	8.0	41.1	43.0	36.4	10.9	18.9	60.6	44.3	48.5	53.4	91.4
MLP-A	57.8	9.3	43.9	40.5	47.7	16.2	19.1	64.5	43.8	44.4	53.2	92.8
MLP-U	55.3	4.7	42.8	37.2	33.3	9.7	16.6	58.8	43.5	43.7	50.8	92.8
TabNet-A	52.2	0.0	40.8	21.8	18.2	3.7	0.0	59.8	16.5	38.6	46.6	92.8
TabNet-U	43.3	0.0	39.0	14.7	8.8	2.8	0.0	51.2	13.9	25.0	45.0	92.8

Looking at the F_1 scores per ICD-10 block, we observe a strong alignment between model performance and label distribution. The Z00–Z99 category (e.g., routine visits or conditions not attributable to specific diseases), being the most prevalent class (87%), was generally predicted with the highest accuracy across all models. In contrast, rare comorbidities such as neoplasms (C00–D49, 12%) and nervous system disorders (G00–G99, 15%) had much lower F_1 scores. This discrepancy is not surprising: ML and DL models often underperform on underrepresented classes due to data imbalance and a lack of sufficient patterns to learn from. Furthermore, for mid-prevalence conditions like mental health (F01–F99) or endocrine disorders (E00–E89), the performance varied depending on the model’s ability to leverage complex feature interactions. Overall, the results are both logical and expected. Models with greater capacity to model non-linear interactions and integrate demographic context performed better.

Additionally, one can notice a consistent trend: the demographic-aware approach outperforms the demographic-unaware counterpart across nearly all ICD-10 diagnostic categories. This superiority spans a wide range of comorbidities, from common conditions like circulatory and digestive disorders to rarer categories such as neoplasms and nervous system diseases. These results strongly support the hypothesis that demographic context (e.g., age group, race, gender) provides valuable complementary information that aids in the accurate classification of comorbidities in HIV patients. The only exception to this pattern is the Z00–Z99 category, where both approaches achieved identical performance across all models. This is a particularly interesting finding, likely attributable to two factors.

The first factor is the nature of this ICD-10 category. Z00–Z99 includes routine medical visits and general health-related encounters, which are overwhelmingly frequent in the dataset (87%). Since such visits are common across all demographic groups and not associated with disease-specific clinical or demographic patterns, the inclusion of demographic features did not offer any significant discriminative advantage. In other words, the abundance and generality of these cases make them equally easy to predict regardless of demographic context. Overall, this finding underscores the importance of tailoring model input features based on the target label. For disease categories where social determinants play a role (e.g., mental health or endocrine disorders), demographic information is highly useful. However, for generic categories such as Z00–Z99, which are driven more by healthcare utilization patterns than by demographic variation, such features offer limited added value.

The second factor is its very high distribution (87%) in the dataset. When a class is so dominant, models are more likely to learn its patterns well, even without additional demographic context, simply because they are exposed to it frequently during training. By contrast, most of the other ICD-10 categories have much lower prevalence (often below 30%), which means the model has fewer samples to learn from. In these cases, demographic features act as important side information that help the model identify hidden patterns and better compensate for the limited number of training examples. For example, mental health disorders or musculoskeletal conditions may manifest differently across age groups or vary based on gender or sexual orientation, making demographic data especially valuable in those cases. This may suggest that if other comorbidities were as frequent as Z00–Z99, the performance gap between demographic-aware and demographic-unaware models might also shrink. However, it is also important to note that not all comorbidities are equally predictable from lab tests, regardless of distribution. So while class frequency helps, the nature of the condition and its relationship to the input features (e.g., lab values, demographics) also plays a critical role in prediction success.

4.2. Feature Importance

Figures 1 and 2 show the permutation feature importance for the XGB model. The feature importance analysis of XGBoost models across individual ICD-10 diagnostic blocks provides valuable insights into which clinical and demographic variables contribute most significantly to disease prediction in HIV patients. Overall, the findings appear consistent with clinical expectations and the nature of each disease category. The involvement of such demographic features explains the performance gap between demographic-aware and unaware models in this category.

A00–B99 (Infectious and Parasitic Diseases): The ranking is led by HIV RNA viral load. The following predictors (Race, Gender, Route Of Infection, Albumin, RBC, Bilirubin-Total, AST) cluster reasonably close. This pattern is medically coherent. Viral load is a direct proxy for virologic control; higher loads are associated with immune activation, CD4 decline and greater susceptibility to opportunistic and intercurrent infections. Liver enzymes and bilirubin reflect hepatic inflammation or cholestasis that often accompany systemic infection, drug toxicity, or coinfections (e.g., viral hepatitis). Albumin and RBC/hematologic indices capture catabolic state and anemia of chronic disease, both common in infectious morbidity. The presence of demographics and exposure routes among the top features aligns with known epidemiology—parenteral exposure and male sex correlate with distinct infection risk profiles and care pathways. Because several demographic variables carry non-trivial importance (not just noise), it is expected that the demographic-aware model outperforms the unaware variant for A00–B99. Importantly, since the drop from the top

feature to the rest is moderate (not orders of magnitude), the model appears to rely on a multifactorial signal rather than a single driver.

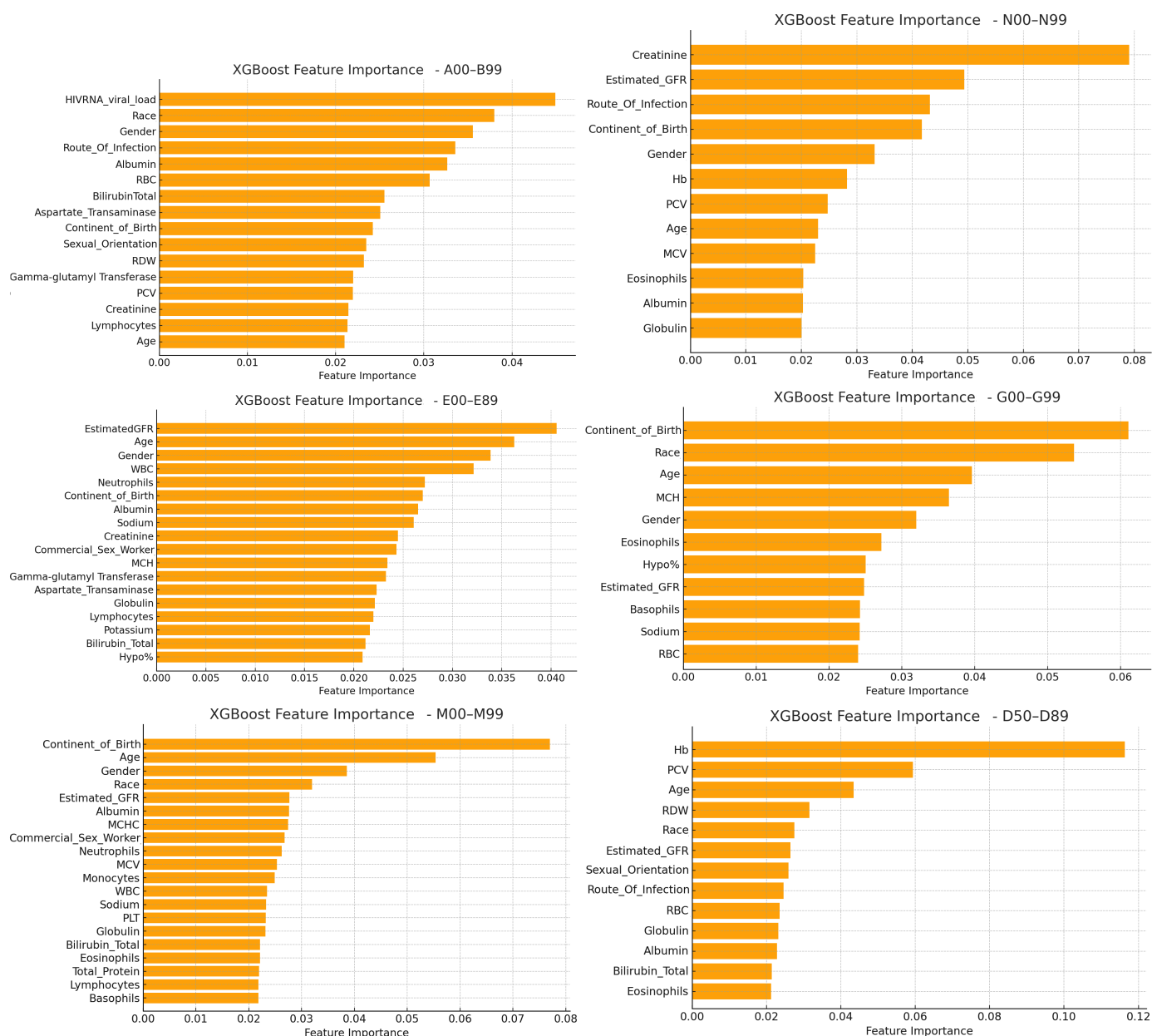


Figure 1. Permutation feature importance for the XGB.

N00–N99 (Genitourinary): Here the dominance of Creatinine (with Estimated GFR second) is pronounced; they exceed the remaining (non-demographic) features by a clear margin. This indicates the model’s predictions for renal/genitourinary disease are driven primarily by canonical kidney function measures—exactly what one would expect clinically. Creatinine and eGFR define chronic kidney disease (CKD) staging and are sensitive to acute kidney injury; they also capture HIV-associated nephropathy and treatment-related nephrotoxicity. The next tier (Route Of Infection, Continent of Birth, Gender) signals demographic/epidemiologic stratification of renal risk (e.g., sex differences in CKD prevalence and the well-documented ancestry-linked susceptibility in certain populations). Hematologic features (Hb, PCV, MCV) plausibly reflect anemia of CKD and altered erythropoiesis. Because a few features (Creatinine/eGFR) carry much larger importance than the rest, the task exhibits a strong physiologic anchor with secondary modifiers;

adding demographics that rank meaningfully high explains the consistent gain of the demographic-aware model.

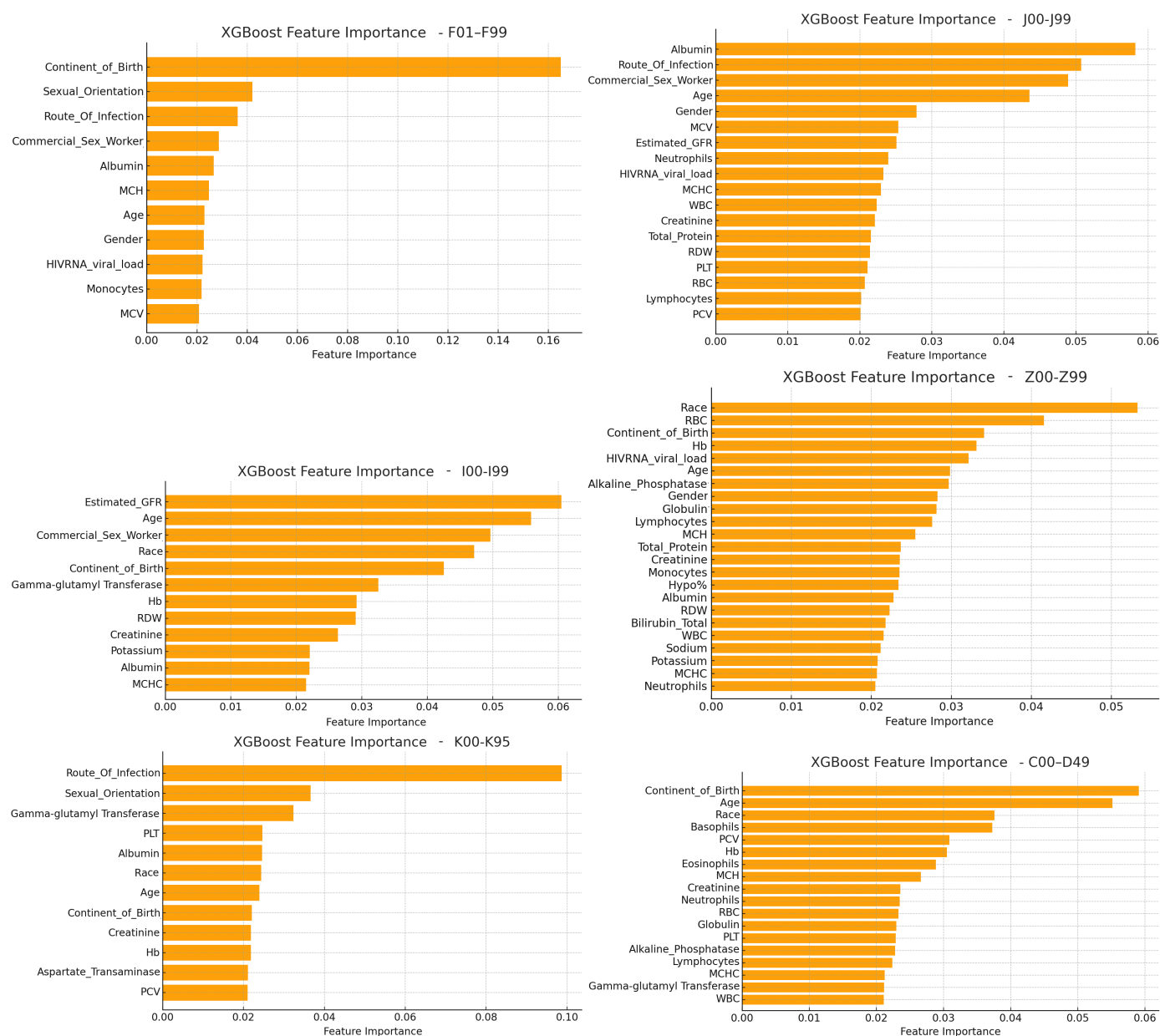


Figure 2. Permutation feature importance for the XGB.

E00–E89 (Endocrine, Nutritional and Metabolic): The top feature is EstimatedGFR, followed by Age and Gender, with a gentle gradient to WBC, Neutrophils, Albumin, Sodium, Creatinine and others. This spread indicates that endocrine/metabolic outcomes in this cohort are polygenic/multifactorial from the model’s perspective—age and sex (well-known risk modifiers for diabetes, thyroid disease, osteoporosis, dyslipidemia) matter alongside renal function and systemic inflammation/nutritional status. The prominence of eGFR/Creatinine is logical given the intimate bidirectional links between metabolic disease and renal impairment (e.g., diabetic nephropathy, electrolyte disturbances). Albumin and inflammatory counts (WBC, neutrophils) capture malnutrition/inflammation syndromes commonly co-occurring with metabolic derangements in chronic HIV. Because demographic variables occupy several top positions, the demographic-aware setting reasonably

yields higher F1: age/sex strata encode baseline risk and treatment patterns that lab values alone cannot fully explain.

G00–G99 (Nervous System): This block is strikingly demographics-led: Continent of Birth and Race top the list, followed by Age, then physiologic markers (MCH, RBC, Sodium, Eosinophils, etc.). Clinically, neurological complications in HIV (from HIV-associated neurocognitive disorder to neuropathies) show clear age effects, ancestry-linked risk and contextual differences in exposure, comorbidity burden and access to care. Hematologic parameters (e.g., MCH, RBC) and Sodium map to anemia and electrolyte disturbances that can worsen cognitive function and seizure thresholds. The large importance of demographic features relative to the rest indicate that population structure and social determinants are major drivers in the model’s decision process here, directly explaining why demographic-aware approach outperforms demographic-unaware for G00–G99. The absence of a single biochemical outlier suggests the neurologic label relies more on contextual risk profiling than on a single hallmark lab test.

M00–M99 (Musculoskeletal and Connective Tissue): The top feature (Continent of Birth) is clearly larger than the rest, followed by Age and Gender, then a descending mix of eGFR, Albumin, MCHC, Neutrophils, etc. This pattern signals strong demographic stratification—sex hormones, body composition, and ancestry-linked bone density patterns affect risks of osteopenia/osteoporosis, osteonecrosis and inflammatory arthropathies seen in chronic HIV and ART exposure. Renal function and albumin reflect bone-mineral metabolism and nutritional status (e.g., CKD-mineral bone disorder, protein-energy wasting). Inflammatory markers (neutrophils/monocytes) resonate with immune activation that fuels musculoskeletal pain and connective tissue disease. Because the top few demographic variables dominate (with a noticeable drop to physiologic features), demographic context carries decisive information, making the demographic-aware model expectedly stronger.

D50–D89 (Blood and Immune): This panel shows the clearest single-feature dominance: Hemoglobin (Hb) sits far above all others; PCV is next but notably lower, followed by Age, RDW, Race, eGFR and additional hematologic/biochemical features (RBC, Globulin, Albumin, Bilirubin Total, Eosinophils). The physiology is textbook: anemia is the most common hematologic abnormality in HIV, driven by chronic inflammation, marrow suppression, micronutrient deficiency, renal disease, or ART effects. Hb and PCV directly quantify anemia; RDW captures anisocytosis typical in mixed/nutritional anemias; RBC corroborates oxygen-carrying capacity; Globulin/Albumin reflect immune activation and nutritional state. The steep drop from Hb to the rest means the model relies primarily on Hb, with others as supporting context. Demographics still contribute and sensibly so—ageing increases anemia risk; ancestry can influence hemoglobinopathies and baseline Hb; social factors link to nutrition and care access, hence the demographic-aware advantage.

F01–F99: Mental, Behavioral and Neurodevelopmental Disorders: The most influential feature is Continent of Birth, which is markedly higher than all other features. This suggests that geographic and cultural background strongly affects the likelihood of mental health comorbidities in HIV patients. This aligns with known disparities: African-born patients may face migration stress, trauma exposure, stigma and barriers to accessing mental health services, all of which elevate risk. The second most important feature is Sexual Orientation, which is medically logical given that men who have sex with men (MSM) populations often report higher rates of depression and anxiety due to stigma, discrimination and minority stress. Other moderately important factors include Commercial Sex Worker status, tied to psychosocial risk and stress-related vulnerability. Laboratory markers such as Albumin and MCH rank much lower, indicating that biochemical data play a minimal role compared to demographics and social factors. This distribution clearly

shows why the demographic-aware model outperforms the unaware one; demographics carry the dominant predictive signal here.

J00–J99: Respiratory Diseases: For respiratory diseases, the leading predictor is Albumin. Low serum albumin is strongly linked to poor nutritional and immune status, both of which increase susceptibility to respiratory infections (e.g., pneumonia, bronchitis) in HIV patients. The next most important predictors are Route of Infection and Commercial Sex Worker status, both of which may reflect behavioral exposures that increase the risk of respiratory infections such as tuberculosis. Age and Gender are also moderately important, consistent with known increased vulnerability in elderly and some gender-specific patterns in respiratory comorbidity. Several immune and hematological markers (MCV, neutrophils, WBC) appear lower down but remain relevant, as impaired immunity directly affects respiratory outcomes. Importantly, here we see a mix of lab and demographic predictors, confirming that both pathways, clinical vulnerability and demographic/social risk, jointly explain outcomes.

I00–I99: Circulatory System Diseases: The strongest predictor here is Estimated GFR, highlighting the close link between renal function and cardiovascular risk in HIV patients. Chronic kidney disease is a well-established risk factor for cardiovascular disease, which explains its dominance. Age follows closely, consistent with cardiovascular disease being strongly age-dependent. Commercial Sex Worker status also appears, suggesting that behavioral and social risk factors play some secondary role. Race and continental origins also carry modest importance, reflecting known ethnic and geographic disparities in hypertension, ischemic heart disease, and stroke risk. Hematological and biochemical markers (Gamma-GT, Hb, RDW, potassium) are further down, reinforcing that kidney function and demographics dominate. This makes sense clinically: cardiovascular risk in HIV is a multifactorial interplay of aging, renal impairment and social/behavioral determinants. The clear strength of demographic features again justifies the advantage of the demographic-aware model.

Z00–Z99: Factors Influencing Health Status and Contact with Health Services: Here, Race is the top feature, followed by RBC counts. This reflects that health service use is not purely clinical but also mediated by ethnicity and access disparities. For instance, non-White patients in the UK often face structural barriers to healthcare, leading to distinct utilization patterns. The contribution of viral load makes sense: patients with poor viral suppression are more likely to have frequent medical visits. Other demographic features (Age, Gender, Continent of Birth) also appear prominently, reflecting different healthcare-seeking patterns. Unlike organ-specific disease categories, here the outcome is more strongly driven by social and demographic context rather than biochemical markers, explaining why the demographic-aware pipeline is especially powerful in this category.

K00–K95: Digestive Diseases: The most dominant feature is Route of Infection, far surpassing others. This suggests that heterosexual transmission is highly associated with digestive comorbidities in this cohort. This may be due to co-infections (such as hepatitis B/C, more common in heterosexual HIV-positive individuals in some regions) that directly affect the liver and gastrointestinal system. The next two important features are Sexual Orientation and Gamma-GT, the latter being a classic biomarker of hepatic and bile duct disease. Platelet count (PLT) and Albumin also show relevance, consistent with liver disease pathology. Demographic features such as Race, Continent of Birth and Age provide additional predictive value. Taken together, this indicates that both infection route and liver biomarkers explain digestive comorbidity risk. The dominance of demographic and infection-route features helps clarify why demographic-aware models outperform here.

C00–D49: Neoplasms (Cancers): The leading feature is Continent of Birth, followed closely by Age. This is logical: cancer incidence is strongly age-dependent, and there may

be geographic differences in cancer prevalence due to genetic, lifestyle, or environmental exposures. Basophils is next, reflecting immune dysregulation which may be linked to hematological malignancies. Hematological and biochemical markers (Hb, PCV, albumin, creatinine, neutrophils) appear in the mid-range, consistent with their role in cancer-related anemia, inflammation, and organ impairment. The pattern is clear: demographic predictors (continent, age, race) dominate, while laboratory markers capture secondary physiological manifestations. This strong demographic dependence again highlights the performance gap between demographic-aware and -unaware models.

Take-home about magnitudes: permutation importance reflects how much performance drops when a feature is shuffled. When the top feature's value is much larger than the rest (as in D50–D89 with Hb), the model's prediction relies heavily on that feature; the others are secondary. When several features are close (as in A00–B99), the signal is distributed across multiple clinically plausible drivers. In blocks where demographics appear among the top features (A00–B99, N00–N99, E00–E89, G00–G99, M00–M99), their non-trivial importances explain why the demographic-aware model consistently outperforms the demographic-unaware variant.

5. Clinical and Fairness Implications

The results of this study offer several actionable insights for clinical deployment and fairness-aware model design in HIV care. First, the feature importance analysis highlights that the predictive drivers of comorbidities are largely clinically interpretable and label-specific. For instance, creatinine and eGFR dominate predictions of renal and circulatory disorders, while hemoglobin and hematological indices govern hematologic outcomes, reflecting well-established pathophysiological mechanisms. This interpretability suggests that such models could be integrated into electronic health record systems to prioritize patients for targeted screening and early intervention, especially where routine markers already signal elevated risk.

Second, demographic features emerge as context-dependent enhancers rather than universal predictors. Their contribution is substantial for conditions influenced by social and biological heterogeneity (such as mental health, neurological, and musculoskeletal disorders) yet negligible for highly prevalent, non-specific categories like Z00–Z99. This distinction underscores that demographic variables can enrich prediction when they encode meaningful epidemiological signal, but their inclusion should be carefully justified and regularly audited to avoid reinforcing structural biases.

Finally, the consistent superiority of demographic-aware models indicates that fairness and accuracy need not be mutually exclusive when demographic inputs are used transparently and interpreted within clinical context. In practice, this means models should accompany clear documentation of input rationale, variable sensitivity, and population representativeness, allowing clinicians to understand not only who is at risk, but why.

6. Conclusions

We demonstrated that multi-label comorbidity prediction for people living with HIV is feasible using routine EHR data and modern ML/DL methods. Across models, XGBoost delivered the best macro- F_1 performance on our 2200-patient cohort, with demographic-aware configurations consistently outperforming demographic-unaware ones. Disease-specific permutation importance profiles aligned closely with medical knowledge (eGFR/creatinine for renal and cardiometabolic risk, hemoglobin/PCV for hematologic disease, albumin for respiratory outcomes and viral load for infectious labels) while also revealing categories where demographic context (age, gender, race, continent of birth, sexual orientation, route of infection, commercial sex worker status) materially improves prediction (e.g., mental health,

neurologic, musculoskeletal). These insights suggest a pragmatic pathway for deployment: use models to prioritize screening and tailor monitoring based on label-specific drivers while transparently governing the use of sensitive attributes and auditing models for bias.

Our study has a few limitations. First, the best performing model achieved an overall macro-F1 score of 45.8%, with substantial variation between ICD-10 categories. Performance was particularly limited for underrepresented outcomes, including neoplasms, nervous-system disorders and musculoskeletal disorders. Although macro-F1 is a stringent metric that assigns equal importance to rare and common labels, the observed performance remains insufficient for autonomous clinical application. Second, the data originated from a single centre and included approximately 2200 patients, which may constrain generalizability and reduce the number of positive examples available for less prevalent comorbidities. Third, the study focused on a restricted set of structured laboratory and demographic/social variables. Medication histories, longitudinal laboratory trajectories, clinical notes, prior healthcare utilization, treatment exposure, imaging findings, and additional lifestyle factors were not incorporated. Fourth, the broad ICD-10 categories contain clinically heterogeneous conditions that may have different underlying risk factors and laboratory manifestations. Future work should therefore incorporate richer longitudinal data, investigate imbalance-aware and label-dependency modeling, report calibration and clinically meaningful operating thresholds and conduct external, subgroup-specific and prospective validation before considering clinical implementation.

In summary, routinely collected laboratory and demographic information contains measurable signals associated with HIV comorbidity profiles. The results establish a retrospective proof of concept for multi-label comorbidity risk stratification and demonstrate the additional information contributed by the demographic context. However, the moderate overall performance and poor results for several rare categories mean that the proposed models are not currently suitable for autonomous clinical use. Further methodological development and prospective multi-centre validation are required to determine whether these models can ultimately provide clinically useful decision support.

Author Contributions: Conceptualization and methodology, S.R., D.K., S.P.A. and Q.Z.; software, validation, investigation and formal analysis, S.R. and D.K.; writing (original draft preparation, review, editing), S.R., D.K., S.P.A. and Q.Z.; supervision and project administration, D.K. and Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The work of Solomon Russom had been supported by King's College Hospital NHS Foundation Trust, under study reference KCH23-246.

Institutional Review Board Statement: This study was conducted using existing anonymised and non-patient-identifiable data provided by King's College Hospital NHS Foundation Trust under a formal collaboration/data-sharing agreement with Queen Mary University of London for the project 'Leveraging Healthcare Data for AI-Enabled, Effective and Improved HIV Care' (IRAS ID: 337179).

Informed Consent Statement: The data were anonymized and non-patient-identifiable; therefore, individual informed consent was not required.

Data Availability Statement: Data can be available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Deeks, S.G.; Lewin, S.R.; Havlir, D.V. The end of AIDS: HIV infection as a chronic disease. *Lancet* **2013**, *382*, 1525–1533. [https://doi.org/10.1016/S0140-6736\(13\)61809-7](https://doi.org/10.1016/S0140-6736(13)61809-7).
2. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. <https://doi.org/10.1038/s41591-018-0300-7>.

3. Shickel, B.; Tighe, P.J.; Bihorac, A.; Rashidi, P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 1589–1604. <https://doi.org/10.1109/JBHI.2017.2767063>.
4. Miotto, R.; Li, F.; Kidd, B.A.; Dudley, J.T. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **2016**, *6*, 26094. <https://doi.org/10.1038/srep26094>.
5. Rajkomar, A.; Oren, E.; Chen, K.; Dai, A.M.; Hajaj, N.; Hardt, M.; Liu, P.J.; Liu, X.; Marcus, J.; Sun, M.; et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **2018**, *1*, 18.
6. Swinckels, L.; Bennis, F.C.; Ziesemer, K.A.; Scheerman, J.F.; Bijwaard, H.; de Keijzer, A.; Bruers, J.J. The use of deep learning and machine learning on longitudinal electronic health records for disease prediction: A systematic review. *J. Med. Internet Res.* **2024**, *26*, e48320.
7. Amirahmadi, A.; Ohlsson, M.; Etminani, K. Deep learning prediction models based on EHR trajectories: A systematic review. *J. Biomed. Inform.* **2023**, *144*, 104430.
8. Islam, K.R.; Prithula, J.; Kumar, J.; Tan, T.L.; Reaz, M.B.I.; Sumon, M.S.I.; Chowdhury, M.E. Machine learning-based early prediction of sepsis using electronic health records: A systematic review. *J. Clin. Med.* **2023**, *12*, 5658.
9. Schwab, P.; Mehrjou, A.; Parbhoo, S.; Celi, L.A.; Hetzel, J.; Hofer, M.; Schölkopf, B.; Bauer, S. Real-time prediction of COVID-19 related mortality using electronic health records. *Nat. Commun.* **2021**, *12*, 1058.
10. Pourhomayoun, M.; Shakibi, M. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health* **2021**, *20*, 100178.
11. Peetluk, L.S.; Ridolfi, F.M.; Rebeiro, P.F.; Liu, D.; Rolla, V.C.; Sterling, T.R. Systematic review of prediction models for pulmonary tuberculosis treatment outcomes. *BMJ Open* **2021**, *11*, e044687.
12. Peng, A.Z.; Kong, X.H.; Liu, S.T.; Zhang, H.F.; Xie, L.L.; Ma, L.J.; Zhang, Q.; Chen, Y. Explainable machine learning for early predicting treatment failure in tuberculosis patients with diabetes. *Sci. Rep.* **2024**, *14*, 6150.
13. Ngcobo, S.; Mntla, E.M.; Shock, J.; Louw, M.; Mbonambi, L.; Serite, T.; Rossouw, T. Artificial intelligence for HIV care: A global systematic review of current studies and emerging trends. *J. Int. AIDS Soc.* **2025**, *28*, e70045.
14. Boudra, T.; Idrissou, A.; Barakat, O.; Davani, S.; Valnet Rabier, M.B.; Lagoutte-Renosi, J. Machine Learning in HIV Care and Antiretroviral Therapy: Systematic Review. *J. Med. Internet Res.* **2026**, *28*, e79219.
15. Yang, X.; Zhang, J.; Chen, S.; Weissman, S.; Olatosi, B.; Li, X. Utilizing electronic health record data to understand comorbidity burden among people living with HIV: A machine learning approach. *AIDS* **2021**, *35*, 105–113.
16. Roth, J.A.; Radevski, G.; Marzolini, C.; Rauch, A.; Günthard, H.F.; Kouyos, R.D.; Fux, C.A.; Scherrer, A.U.; Calmy, A.; Cavassini, M.; et al. Cohort-derived machine learning models for individual prediction of chronic kidney disease in people living with HIV: A prospective multicentre cohort study. *J. Int. AIDS Soc.* **2021**, *24*, e25723.
17. Liu, Z.; Meng, Z.; Wei, D.; Qin, Y.; Lv, Y.; Xie, L.; Qiu, H.; Xie, B.; Li, L.; Wei, X.; et al. Predictive model and risk analysis for coronary heart disease in people living with HIV based on machine learning and electronic medical records. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 160.
18. Li, Y.; Huang, H.; Li, C.; Li, X.; Zhang, J.; Pan, X.; Xu, Y.; Li, S.; Zhao, Y.; Gong, B.; et al. Development of a machine learning model for cardiovascular diseases risk prediction in older HIV / AIDS patients in Southern China from an eleven-year cohort study. *BMC Infect. Dis.* **2026**, *26*, 52.
19. Paulus, J.K.; Kent, D.M. Predictably unequal: Understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *npj Digit. Med.* **2020**, *3*, 99.
20. Wang, Y.; Pillai, M.; Zhao, Y.; Curtin, C.; Hernandez-Boussard, T. Fairehr-clp: Towards fairness-aware clinical predictions with contrastive learning in multimodal electronic health records. *arXiv* **2024**, arXiv:2402.00955.
21. Menard, S. Applied Logistic Regression Analysis. In *Quantitative Applications in the Social Sciences*; SAGE Publications: Thousand Oaks, CA, USA, 2002; Volume 106.
22. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
23. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
24. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017*; NeurIPS: Sydney, Australia, 2017; Volume 30, pp. 3146–3154.
25. Arik, S.O.; Pfister, T. TabNet: Attentive interpretable tabular learning. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 6679–6687.
26. Sechidis, K.; Tsoumakas, G.; Vlahavas, I. Stratification for multi-label data. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD) Workshop on Knowledge Discovery and Data Mining, Athens, Greece, 5–9 September 2011*; ECML PKDD: Athens, Greece, 2011; pp. 145–158.
27. Grandini, M.; Bagli, E.; Visani, G. Metrics for multi-class classification: An overview. *arXiv* **2020**, arXiv:2008.05756.

28. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347.
29. Singh, D.; Singh, B. Investigating the impact of data normalization on classification performance. *Appl. Soft Comput.* **2020**, *97*, 105524.
30. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.