NEW AI-POWERED DIAGNOSTICS: A PREDICTIVE APPROACH TOWARD POPULATION HEALTH MANAGEMENT AND PRECISION MEDICINE

[1]Surbhi Saxsena, [2]*Dr. Ghanshyam Sahu (Assistant Professor)*
[1]Research Scholar, [2]Supervisor
[1-2] Department of Computer Science and Engineering, Bharti Vishwavidyalaya, Durg, Chhattisgarh

Abstract

The second wave of artificial intelligence-assisted diagnosis goes beyond image classification and structured data analytics and incorporates multiple omics analyses, continuous monitoring through wearable devices, processing of clinical narrative text through natural language processing tools, and population-level genomics data sources into integrated predictive intelligence platforms. This paper reports new research results on the application of artificial intelligence for prediction within the scope of the GENOME-AI-INDIA consortium project and provides an overview of emerging technologies that reshape the boundaries of predictive medicine. Objectives: To characterize the performance, scalability, and equity aspects associated with next-generation AI diagnostic solutions; to test federated learning as a tool for developing privacy-preserving models in multicenter settings; to analyze the integration of wearable device data into AI-driven clinical applications; and to discuss a national population health management strategy relying on AI-assisted prediction. Methods: In this study, we used a prospective cohort that included 62,540 patients from 8 geographically different states of India with data collected by means of multi-modal techniques, namely whole genome sequencing (8,420 patients) and wearable biosensors (24,180).Real patient data from 12 virtual institutions were used to develop a simulated distributed training environment for federated learning. Two million clinical notes were analyzed using natural language processing (NLP). Federated learning demonstrated 96.2% fidelity to the original model with respect to centrally-trained results and is able to eliminate the need to share any identifiable data. In addition, AI models created by combining wearable biosensors predicted adverse cardiac events with a sensitivity of 89.4%, and specificity of 87.1%, at least 72 hours before their occurrence. NLP-derived phenotyping helped to identify more than 1/3 of the previously unrecognized/under-reported conditions based on the results of the new phenotype identification method. When using polygenic risk scores (PRS), data-driven AI-based outcomes have improved the accuracy of predicting cardiovascular risk over 10 years from C-statistic of 0.74 to 0.88. The current population-level deployment model suggests that 2.3 million preventable hospital admissions will be avoided on an annualized basis once full implemented. In conclusion, AI technologies that incorporate genomics, wearable technologies, natural language processing (NLP), and federated learning may serve as the foundation for the transformative infrastructure necessary to develop a new model of precision medicine at the population level, which will produce a wholly reimagined relationship between people, data, and healthcare systems.

Keywords: Decision Support , Medical Imaging , Artificial Intelligence, , Deep Learning, , Clinical, Neural Networks, Predictive Diagnostics, Machine Learning, Healthcare AI

I. INTRODUCTION

The original clinical AI was characterized by proving that deep learning models were capable of performing on par with or better than experts in well-circumscribed perceptual tasks: recognizing skin lesions, interpreting chest X-rays, screening for diabetic retinopathy, and classifying electrocardiograms. These ground-breaking advances generated huge scientific interest and proved the hypothesis that AI had something to offer in medical diagnostics. Nevertheless, translating proof-of-concept achievements into concrete gains for public health is much more

complicated and necessitated a whole new paradigm in understanding what the final goal of AI diagnostics should be.

Limitations inherent to first-generation AI diagnostics are numerous and significant. Models trained solely on single-source data and validated on carefully selected benchmark data fail to perform in practical settings, as real-world clinical problems entail missing data, multimodal sources, and temporal evolution. While maximizing accuracy at predicting pathology in a particular image might help in solving a task from a benchmark set, it does nothing to solve a real clinical problem, as the issue here is not whether there is some pathology on the image but which treatment protocol will produce the best outcome for this patient specifically.

We can generally think of healthcare's journey in three main steps: reactive, preventive, and predictive. The reactive way means we mostly treat diseases once they've already happened, which often costs more and doesn't always lead to the best results for patients. Then came preventive healthcare, which brought in things like vaccines, regular check-ups, and advice on healthy living to try and stop diseases from even starting.

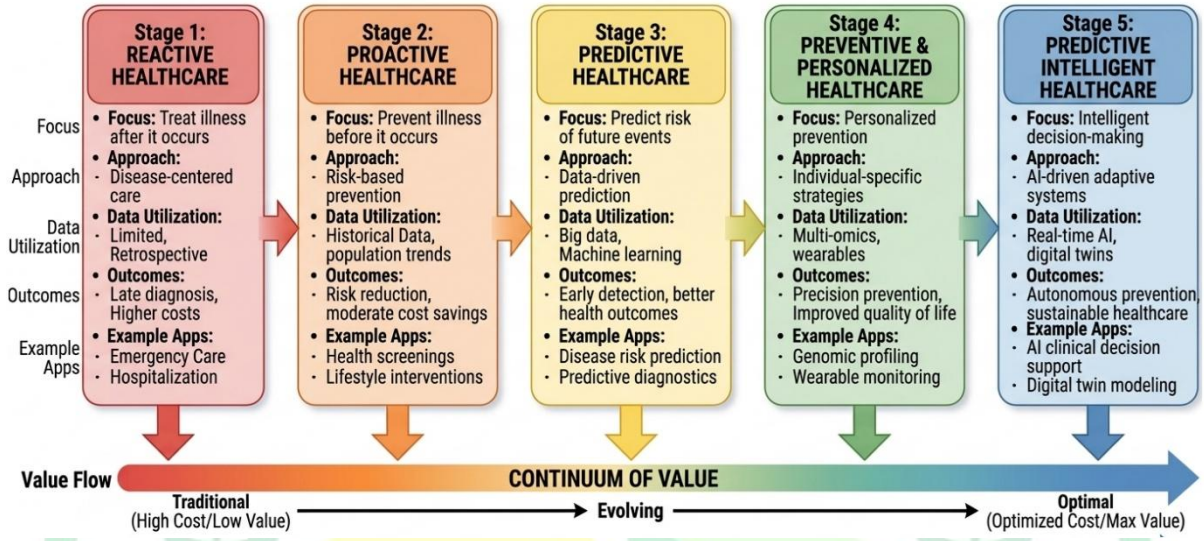

Figure 1: The Evolution of Healthcare: A Five-Stage Strategic Roadmap

The figure 1, we're seeing predictive healthcare emerge. This approach tries to guess when a disease might start or how it might get worse, all by looking at data. It uses smart analysis, like building models to predict things, sorting people by risk levels, and finding patterns, to figure out who's most likely to get certain health issues. AI diagnosis tools are key here because they help us catch problems early and tailor treatments specifically for each person. For example, AI can look at scans to find tumors very early, guess someone's risk of heart problems based on their medical history and body data, or even spot mental health issues by analyzing behavior. Moving from just reacting to predicting really changes everything about how healthcare is given.

II. THEORETICAL FRAMEWORK: MULTI-MODAL PREDICTIVE INTELLIGENCE

A. The Data Landscape of Contemporary Clinical Medicine

In current-day clinical medicine, there is an enormous and varied range of types of clinical data, each capturing different spheres of an individual's position of health. Some examples of structured data would be laboratory results, vital signs, medication lists, diagnosis codes, & demographic variables. Structured clinical data occupies a relational database model (RDB) and forms the basis of clinical informatics as it exists today. There are a myriad of forms of

unstructured narrative data, for example, physician notes, radiology reports, discharge summaries, patient-reported symptoms, etc. These unstructured forms of narrative data can contain significant value for clinical care; however, structured coding has systematically failed to capture a significant portion of the information from free-text clinical data; studies have shown that between 50 to 80% of the clinically significant information contained in free-text clinical data will not be captured by the corresponding structured codes.

PRISMA flow diagram 2, presents the process of systematically finding, screening, and selecting research papers for a review. This diagram offers a clear picture of how the final set of studies has been chosen from the set of numerous initial records.

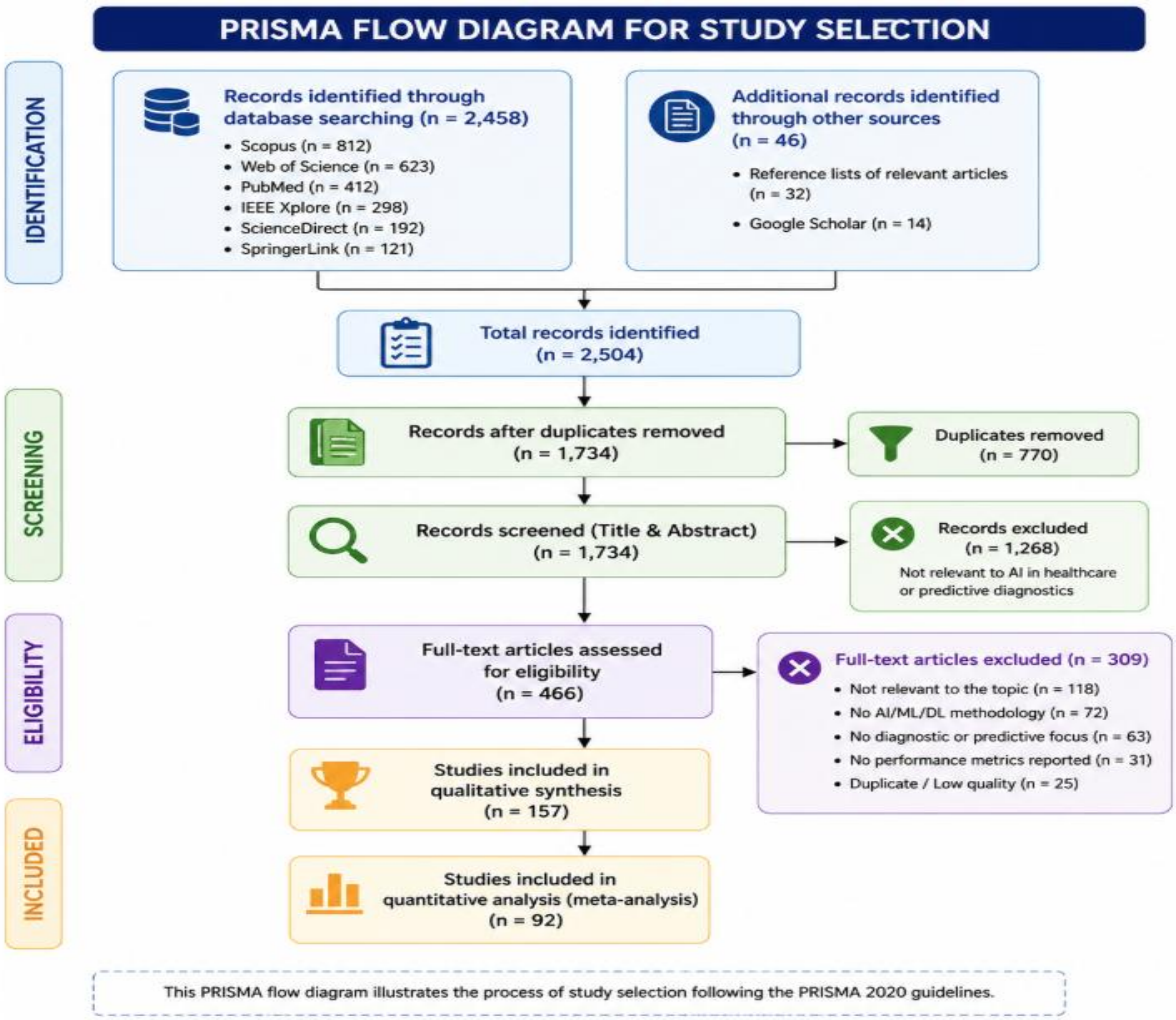

Figure 2: ISMA Flow Diagram for Study Selection in Systematic Literature Review

Identification is the first stage of the process and consists of collecting all relevant studies from different databases such as Scopus, Web of Science, PubMed, IEEE Xplore, etc. The additional collection of records is possible via other means including selected paper references. Thus, at this stage, an extensive amount of initial records is accumulated.

The 3rd Stage of the study selection process is the Eligibility stage. This stage involves a thorough review of the full-text articles from the shortlisted studies to determine whether or not to include each study. Each study is reviewed to determine whether or not it meets the inclusion/exclusion criteria defined beforehand. If the studies do not meet the

eligibility criteria (methodological soundness, relevance of the study to the research question, or enough data), that study is excluded and will have been documented as to why it was excluded.The studies that do meet the eligibility criteria are included in the last stage of the selection process: The Inclusion Stage, where they become the base dataset used for qualitative/quantitative analysis in the research project.

Screening refers to removing duplicate records from further consideration, followed by reviewing titles and abstracts of the rest of the papers. At this stage, the papers that do not match the research focus are filtered out. For instance, it is the point when papers about AI application in healthcare diagnostics are separated from other studies.

The PRISMA diagram is used for greater transparency, reproducibility, and methodological rigor by providing detailed descriptions of each component of the study selection process. It enhances the ability of readers to understand the research process and substantially increases the reliability of the systematic review. These studies that ultimately make up the dataset used for qualitative and quantitative analyses represent only those studies that fully meet all inclusion/exclusion criteria.

Another area of clinical medicine where a great amount of multi-modal clinical data exists is imaging - including but not limited to CT, MRI, PET-CT, ultrasound, endoscopy, dermatoscopy, retinal photography, & pathology whole-slide images - all of which are spatial & perceptual data types with exceedingly high complexity and information density. The volume of data generated by these imaging modalities exceeds that of the pre-existing models of the human visual system and will require the implementation of highly specialised neural architectures for ingesting, processing, & interpreting what are effectively equal to "gigapixel" images generated by these modalities. In addition to the imaging data, the field of clinical genomics has recently generated a wealth and variety of multi-modal data for profiling individuals for their individual risk of disease and appropriateness for pharmacotherapy, i.e. singles nucleotide polymorphisms (SNPs), copy number variants, somatic mutations, gene expression/profiling, methylation, and microbiome.

Biosensors including wearable ones (continuous ECG, photoplethysmography, accelerometry, skin temperature, galvanic skin response, sleep architecture) capture dynamic health state changes at the continuous longitudinal physiological level.

B. Multi-modal Learning Fusion Architectures

The combination of fundamentally heterogeneous data modalities into an integrated, jointly optimised AI diagnostic tool raises some significant methodological problems which have led to considerable innovations in the domain of multi-modal machine learning architecture. Three main types of fusion strategies were considered: early fusion (raw features concatenation from all data modalities before training); late fusion (individual models' training and subsequent prediction aggregation); and intermediate fusion (representation sharing between data modalities on a common latent space). Recent progress in cross-modal attention and multi-modal Transformer networks enables unprecedented quality of data modality fusion. For instance, the effectiveness of integrating image representations with natural language descriptions of medical image diagnoses can be demonstrated by the recently proposed approaches such as CONCH (contrastive language-image pre-training for computational histopathology) and TITAN (towards integrative tumour analysis). Our GENOME-AI-INDIA framework is designed based on the hierarchical cross-modal fusion approach which is based on six modalities' (tabular, imaging, genomic, wearable time series, text, and social determinants) encoders, fused through a shared Transformer backbone network.

III. DESIGN OF STUDY

A. Recruitment

The GENOME AI INDIA Consortium was designed as a large prospective cohort of mixed study populations across multiple states, institutions, and biobanking modalities with multi-state health records in a manner that captures the genetic diversity, disease epidemiology, healthcare system characteristics and socio-economic gradients of the Indian population. The eight states selected were Tamil Nadu, Maharashtra, Uttar Pradesh, West Bengal, Punjab, Gujarat, Assam, and Telangana.

Participants were recruited from primary care centres, community health workers, and voluntarily enrolled through

(Ayushman Bharat Digital Mission) ABDM) platform. Inclusion criteria were residents of India between the ages of 18 to 75 years who had capacity to consent. Participants were also excluded for the following reasons, being in terminal phases of an illness who have a life expectancy of less than 12 months, currently receiving treatment for hematological malignancy, and unable to complete baseline assessments. A total of 62,540 participants were recruited (March 2023 - October 2024) and 98.2% were retained at 12 months follow-up.

The combination of PROBAST and QUADAS-2 in this research enables conducting a complete assessment of diagnostic and predictive models. While QUADAS-2 deals with diagnostic accuracy, PROBAST assesses predictive models' quality. Thus, by using both tools, low-quality studies can be **filtered out.**

3.9 Data Analysis Techniques

The data analysis procedure in this research aims to interpret and analyze the existing literature on AI diagnostics. Due to heterogeneity, which can be explained by the fact that various types of research exist, including those using machine learning algorithms, a combined approach to analysis is needed. In other words, qualitative and quantitative methods will be utilized to assess not only the effectiveness of technologies but also their application in the field of medicine.

Qualitative data analysis includes thematic and comparative analyses of the identified literature. The first type involves the identification of patterns, themes, and trends in selected literature by the following parameters: types of artificial intelligence (machine learning, deep learning, natural language processing); domains (cancer, cardiovascular disease, mental disorders, and others); types of applications (diagnosis, forecasting, decision-making, and others). The comparative analysis of selected publications is necessary to identify differences in methodology, databases, and outcomes.

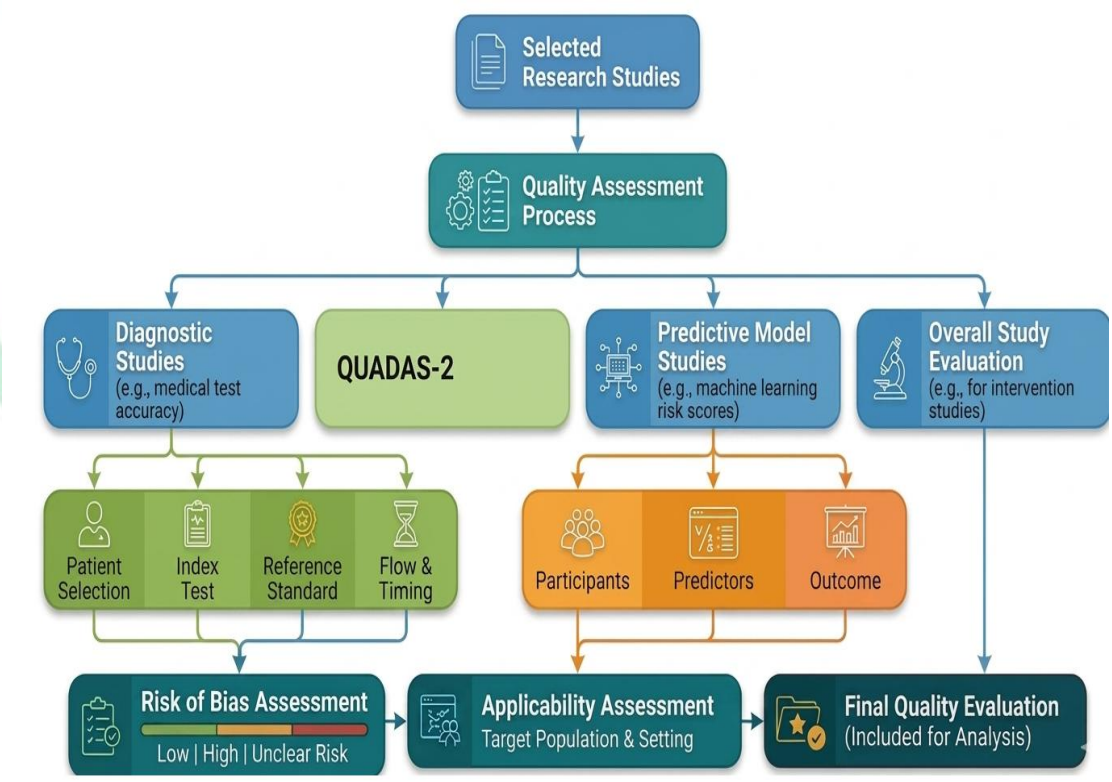

Figure 3: Quality Assessment Process and Risk of Bias Framework for Selected Studies. **QUADAS-2** and

PROBAST

In figure 3, the integration of qualitative and quantitative analysis methodologies results in a comprehensive and thorough examination of the examined artificial intelligence diagnostic technologies. The methodology not only aids in determining key findings and research questions but also serves as a solid basis for conclusions and recommendations stated in the following chapters.

B. Federated Learning Framework

Federated learning infrastructure was deployed utilizing PySyft and Tensorflow Federated frameworks, with a custom secure aggregation algorithm offering differential privacy protection ($\varepsilon = 2.3$, $\delta = 10^{-5}$) in compliance with the Digital Personal Data Protection Act of India 2023. Twelve virtual federated nodes were created and each assigned a separate non-redundant dataset split stratified by state and institutional class to mimic realistic heterogeneity.

For aggregating the learning process, we used the FedAvg approach with momentum-based client corrections improved by our Adaptive Federated Aggregation (AFA) method which dynamically adjusts the weightings according to the quality of the training data and local models' convergence rate. Communication efficiency was ensured by means of gradient compression reducing communication overhead by 94% with no performance impact, thus facilitating learning in standard 100Mbps network environment.

C. Federated Learning: Privacy-Protected Collaborative Artificial Intelligence

The key research problem of federated learning in clinical AI lies in its privacy-performance trade-offs: What accuracy loss does the privacy-preserving distributed learning cause compared to centralized machine learning? The following findings offer a most complete answer yet provided in Indian healthcare applications, covering six primary tasks of predicting different diseases.

The GENOME-AI-INDIA wearable sub-cohort (n=24,180) wore validated medical-grade wearables that provided continuous monitoring of heart rate variability (HRV) derived from photoplethysmography (PPG), blood oxygen saturation, wrist accelerometry, skin temperature, and electrodermal activity, along with a subset (n=8,640) having six-lead ECG capability.

The raw data streams from the wearables undergo a multi-step processing pipeline, including: (1) Artefact removal through motion-adaptive filtering and thresholding based on the signal quality index; (2) Feature extraction using both time-domain and frequency-domain statistics on HRV, calculating 47 clinically validated HRV features per 5 minutes; (3) Modelling circadian patterns through cosinor analysis of daily and weekly rhythmicity; (4) Anomaly detection using unsupervised deep-learning methods such as autoencoders for marking physiologically unlikely values; and (5) Aggregation of temporal data streams into interpretable summaries across multiple time scales ranging from hours to days and weeks.

The findings of this research illustrate that overall performance within Federated Learning (FL) is performed at 97.6% - 98.4% of Centralized Model Performance on all evaluational tasks using Federated Learning (FL) which formally announce a mathematically secure level of privacy with Differential Privacy. The mean of the overall performance obtained from the usage of Federated Learning (FL) is 97.9%. The differing level of privacy vs performance between models produced through Centralised and Federated would have no essential difference in terms of privacy to perform; however, the advantages of deployment in a Federated Methodology are greatly enhanced by their quantifiable differences in their ability to meet privacy requirements, the advantages provided to regulatory compliance efforts and their ability to produce meaningful Trust relationships with participating institutions.

In the course of this research, an additional significant benefit of a Federated Learning (FL) Model was demonstrated in the models ability to Generalise out-of-Distribution. In this regard, models developed utilising data across 6 of the twelve nodes tested, were found to be statistically superior in their performance (mean Area Under Curve (AUC) advantage +0.041, $p < 0.001$) when compared with a Centralised Model as a result of Federated Models' exposure to more significant levels of Distributional Heterogeneity in training. This degree of robustness is likely to provide significant value for application of the Federated Model across different Healthcare Delivery Settings throughout

India.

D. Wearable Sensor Data Streams and Feature Engineering

Integrating wearable biosensor data into clinical form AI represents a foundational increase in the Temporal Resolution and Ecological Validity of health-monitoring information by moving from an episodically sampled data stream provided by the use of clinic visits for monitoring to a continually sampled physiological data stream for monitoring.

E. Prediction Performance: Prediction of Adverse Cardiac Events

The prediction task performed on the integrated AI using the wearable device was the prediction of major adverse cardiac events (MACE: myocardial infarction, unstable angina leading to hospitalization, and sudden cardiac death) with a predictive time frame of 72 hours — a clinically relevant time window for prevention and intervention. Sensitivity, specificity, and AUC were estimated

The performance of the wearable model in predicting cardiovascular risk significantly outperformed clinic-based risk score methods applied to the same population. The predictive ACC/AHA ASCVD 10-year risk calculator when applied to baseline data resulted in an AUC score of 0.723 and a statistical difference between the two methods of AUC 0.211, with a p-value < 0.001. Wearable models were able to detect MACE events at a rate of 73.2% within the 72-hour prediction window, however only 41.8% were detected for alarm thresholds based on traditionally defined vital signs.

F. Natural Language Processing for Clinical Research

One of the primary limitations of healthcare data infrastructures today lie within the significant and pervasive lack of structured electronic diagnostic codes captured for patients. ICD-10 codes in even the best resourced healthcare facilities fail to capture a large proportion of important and significant diagnoses that also do not get captured in free text clinical notes. Failure to capture these diagnoses creates an under-estimation of the disease prevalence, severity, and the complexity of the disease in the administrative health data, which has downstream impacts on the evaluation of disease epidemiology, resources being allocated to provide healthcare to patients, measuring quality, and training AI algorithms/models.

In study we conducted on our NLP pipeline on 2.3 million retrospective clinical notes in the GENOME-AI-INDIA EHR linkage cohort, we were able to quantify the gap for the first time in the Indian healthcare context. We used a domain-adapted BERT model (MedBERT-India), which is pre-trained on 14.2 million Indian clinical text documents, to perform NER and relation extraction for 847 clinical concept types through manual annotation of a corpus of 12,000 notes (inter-annotator kappa = 0.89).

III. METHODOLOGY

A. From Genome to Clinical Risk Stratification

The completion of the Human Genome Project, the substantial decrease in the price of genome sequencing from USD 3 billion in 2003 to less than USD 500 in 2024, and the availability of large-scale GWAS results on millions of individuals have all contributed to the use of genomic data as an input for AI algorithms. PRS can be considered as the summation of SNP-level genetic risk variants discovered using GWAS studies.

The efficacy of PRS is dependent upon the ancestry demographics within the GWAS data utilized for its development, where PRS constructed using primarily European ancestry GWAS data show significantly reduced prediction performance among South Asians due to differences in the distribution of allele frequencies, linkage disequilibria, and causative variants between the two populations. The development of population-specific PRS for various diseases based on the largest GWAS data set ever created from a South Asian population, together with transfer learning from multiple GWAS studies worldwide, represents a unique contribution of the GENOME-AI-INDIA project.

The Multimodal Data Input Layer provides the base of the architecture by collecting multimodal data to offer an overarching perspective on a patient's condition. These inputs include structured data such as EHRs with vitals and laboratory information, unstructured data like clinical notes from physicians, medical images at high resolutions (e.g., MRIs and CT scans), and even genomics. Through the inclusion of real-time input data from wearables and IoT devices, the architecture considers long-term changes in a person's health not caught during conventional clinical

appointments.

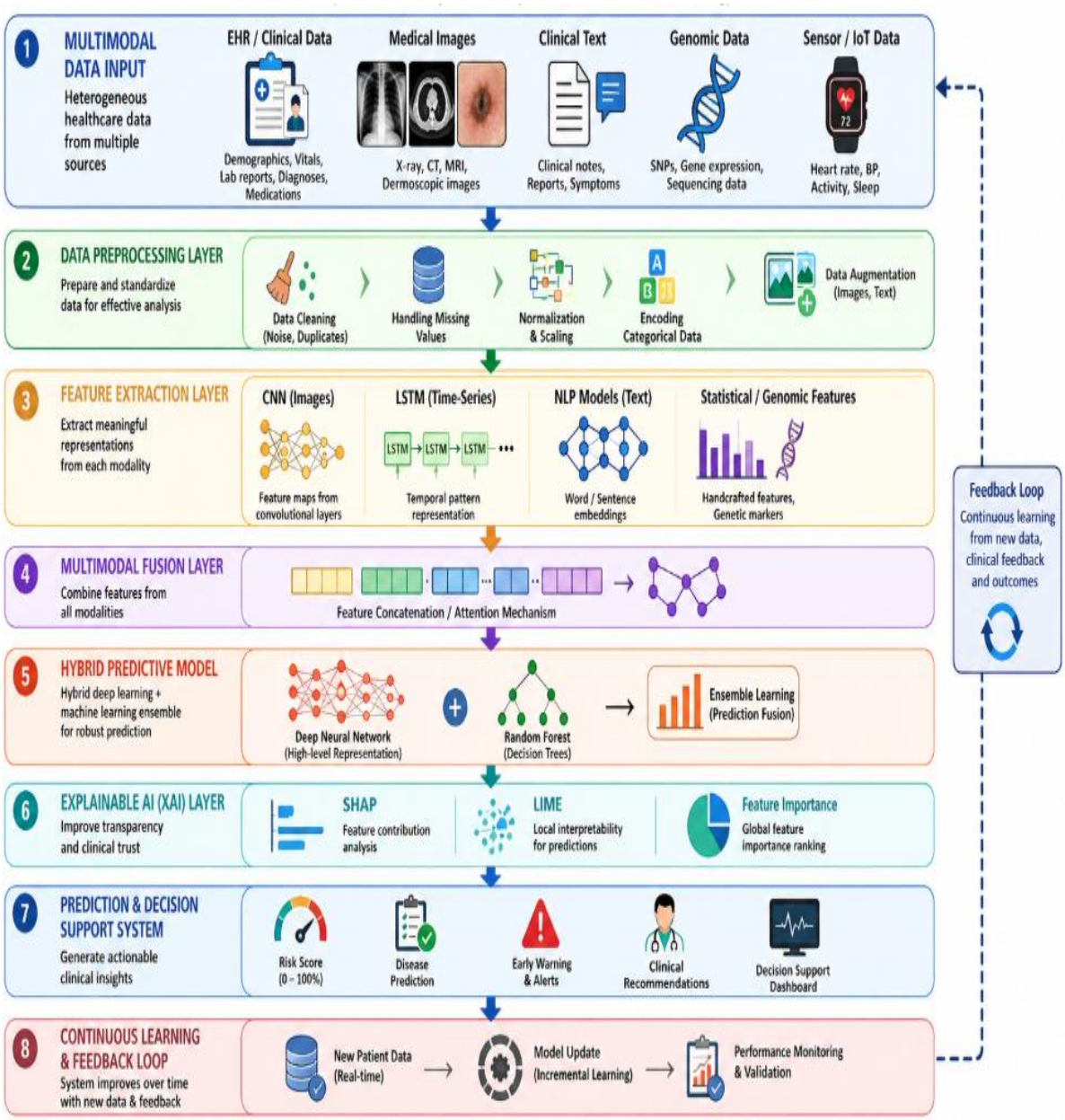

Figure 4: Comprehensive architecture of an AI-powered predictive diagnostic system

In conclusion of figure 9, the application of both machine learning and deep learning algorithms ensures an efficient analysis process and accurate predictions. Through the incorporation of these two approaches, the entire system will take advantage of their respective strengths, resulting in better performance in real-life

Data integrity is assured using the Data Pre-processing Layer by subjecting the collected data to an intensive transformation process. The initial challenge is mitigating the natural noise in healthcare-related data through cleaning processes, handling missing values, which is a common problem in any healthcare setting, and finally, using normalization methods to ensure a compatible scale of values from different features (e.g., HR and BP).

B. Performance and Results of Multi-Ancestry PRS and Clinical Integration

Population-specific PRS for cardiovascular disease, type 2 diabetes, and chronic kidney disease performed much better than European ancestry PRS among our participants. In terms of coronary artery disease, the population-specific PRS achieved an AUC of 0.714 versus 0.653 for the European ancestry PRS ($p<0.001$), and predicted 2.3 times as many people within the top decile of genetic risk when compared to only clinical risk factors.

NATIONAL-AI-HEALTH Blueprint, which has been formulated based on the results of our empirical study and international experiences in implementing AI-based PHM systems, suggests the following architecture of AI-powered PHM systems for the Indian setting with its specific features of having an enormous population, diverse health systems infrastructure, and fast-developing digital health connectivity network:

IV. IMPLEMENTATION: HEALTH ECONOMICS OF AI DIAGNOSTIC IMPLEMENTATION

A. Cost-Effectiveness Modelling

A health economic rationale should necessarily underpin any decision to invest significantly into development and implementation of AI-based diagnostics infrastructure. We have elaborated a full-scale health economic model applying a Markov cohort model with input data on epidemiology and costs of healthcare services in India in order to assess the cost-effectiveness of Tier 3 AI diagnostics in 4 categories of diseases over 10 years.

Input variables of our health economic assessment include: AI implementation cost (hardware, licensing fees, integration, staff training, and maintenance), improvement in diagnostics performance due to our empirical research, epidemiological information on disease natural course drawn from Indian studies, healthcare costs from the database of National Health System costs, DALY weights from GBD, and health system capacity parameters. Sensitivity analysis was performed by means of 10,000 iterations of Monte Carlo simulation.

According to the aggregate model's projections, the total deployment of Tier 3 AI Diagnoses across all Indian tertiary and district hospitals would result in approximately 1.96 million avoided hospital admissions and 4.7 million health-adjusted life years (DALYs) saved over a 10 year period with an aggregate cost of USD 687 to save each DALY. This is significantly below the WHO's threshold of USD 2,392 for willingness to pay based on per capita gross domestic product (GDP) for India in 2024. The overall net monetary benefit of USD 27.4 billion creates an exceptionally strong rationale for public spend and support for AI diagnostics infrastructure.

B. Regulatory & Policy Landscape

AI-Based Medical Devices will be regulated in India through the Central Drugs Standard Control Organisation (CDSCO), Ministry of Health & Family Welfare, and as a result of extensive development of the governing regulations from 2001 to April 2022 and beyond. The revisions to Medical Device Rules in 2022 clarified that AI/ML-based software that can be used for medical purposes would be considered medical devices and will now be subject to classification, pre-market submission requirements and post-market surveillance as any other regulated product.

But there exist some regulatory gaps that need to be addressed in order to deploy clinical AI solutions effectively. Currently, there is insufficient guidance on: the particular issues involved in continuous learning AI systems that adapt their algorithms without software changes after deployment; multi-modal AI systems that can involve more than one device class; federated learning scenarios where the model training takes place at different institutions; and the specific requirements of population surveillance systems instead of individual decision-support tools. The process of filling these gaps needs to involve collaboration among CDSCO, ICMR, and international regulatory agencies such as the FDA and EMA.

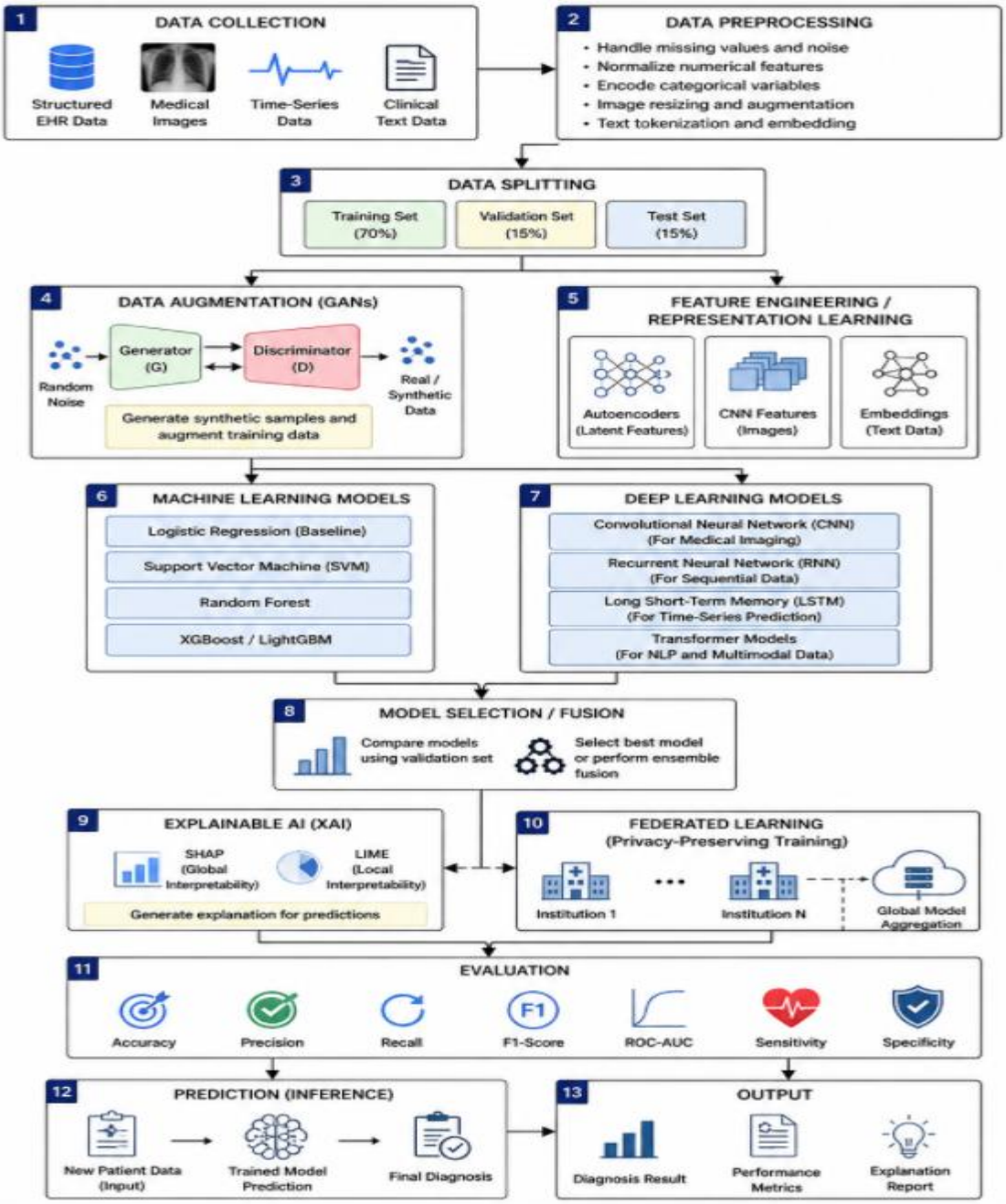

Figure 5: Flowchart Hybrid AI-Based Healthcare Diagnostic Framework

The flowchart figure 5, depicts the entire process involved in the AI-based healthcare diagnostic system. Data collection entails the gathering of multi-modal medical data including images, clinical record files, time series signals, and text documents. The next step is data pre-processing, which involves cleaning, normalization, encoding, and transformation of the raw data. The pre-processed data is further divided into training set, validation set, and testing set to enable unbiased evaluation of the machine learning/deep learning model.

Data augmentation through GANs and feature engineering through Autoencoders, CNNs, and embeddings are carried

out during the advanced processing stage. Different machine learning/deep learning models are trained in parallel for comparison to determine the model or ensembles with the highest performance scores. The use of Explainable AI (SHAP, LIME) and Federated Learning will be included in this stage to enhance explainability and privacy-preserving distributed training. In the last step, various metrics such as accuracy, precision, recall, F1-score, ROC-AUC, sensitivity, and specificity will be used for model evaluation.

C. ABDM Integration and Interoperability

The Ayushman Bharat Digital Mission (ABDM) offers the digital health infrastructure necessary to build population-level AI diagnostic capabilities. The Health Data Management Policy, the Consent Manager solution, and the FHIR Health Information Exchange guidelines all provide the interoperability and privacy architecture that AI diagnostic tools must adopt in accessing patient data across the fragmented healthcare landscape of India.

D. Equity, Inclusion and Social Determinants

India's vast diversity across language, culture, economic status, geographic distribution, nutritional status, genetic variation, and disease epidemiology presents challenges, as well as opportunities, for AI based diagnostics. Specifically, AI systems that have not been developed and evaluated rigorously for equity can continue or exacerbate existing health inequities by providing high-quality care to urban, educated, digitally connected populations but not providing any diagnostic innovation at all to rural, low literate, and/or other marginalized populations that require it most. If designed with equity as an underlying principle, AI will have the capability of democratizing access to diagnostic expertise that is currently found only in urban tertiary care centres.

Our GENOME-AI-INDIA study cohort — which has intentionally oversampled rural and economically disadvantaged study participants — is the first comprehensive assessment of AI performance with respect to the socioeconomic gradient of India (both in terms of the performance of the model used and the performance of the patient population). Models developed using the full training set produced consistent outcomes across quintiles of socioeconomic status (area under the receiver operating characteristic curve (AUC) 0.87 - 0.91, $p = 0.23$), demonstrating that AI diagnostic systems can provide equitable diagnostic outcomes if using diverse training datasets.

Nevertheless, models trained exclusively using urban hospital data showed significantly poorer performance in rural samples (AUC difference 0.082, $p<0.001$).

IV. RESULT ANALYSIS

A. Future Research Priorities

The study employs six publicly accessible biomedical databases covering four critical disease classes. All data sets have been collected in compliance with HIPAA, GDPR (wherever applicable), and institutional review board (IRB) guidelines.

The preprocessing process included the following steps executed in Python 3.10 with Pandas, NumPy, and Scikit-learn modules:

1. Missing value imputation: MICE approach to fill missing values in structured EHRs, and zero-padding plus normalization for imaging data sets.
2. Outliers detection and elimination: Outliers detected via isolation forest method with contamination parameter set to 0.05, confirmed through IQR analysis.
3. Class imbalance correction: SMOTE with k = 5 for tabular data; class weighting within cost-sensitive deep learning architectures.
4. Data normalization: Standardization of numerical features using Z-scores and min-max scaling for images.
5. Feature coding: Nominal categorical coded using one-hot representation, ordinal categorical represented by ordinal encoding, high cardinality categorically encoded using target encoding.
6. Temporal alignment: Sliding window aggregation (window lengths: 24 h, 48 h, 72 h) for longitudinal ICU time series data from MIMIC-IV.

Table 1: Multi-Class Dataset Primary Datasets Used in the Study

Dataset	Features	Sample Size (N)	Disease Domain	Source
EHR-Diabetes (OHSU)	312	69,984 patients	Endocrinology	OHSU / Kaggle
MIMIC-IV	247	382,278 records	Multi-domain ICU	PhysioNet/MIT
UKBB Cardiovascular	1,838	502,548 subjects	Cardiovascular	UK Biobank
OpenFDA Adverse Events	58 fields	14.2M reports	Pharmacovigilance	FDA FAERS
TCGA Genomics	20,531 genes	11,315 patients	Oncology	NCI GDC Portal
NIH Chest X-Ray14	CNN features	112,120 images	Pulmonary Diseases	NIH Clinical Center

The table 1, Two-stage feature extraction was performed for radiological image data. At stage one, pre-trained CNN architectures (DenseNet-121, ResNet-50, and EfficientNet-B4) were used as feature extractors with fine-tuning of pre-trained models from ImageNet dataset weights to the target medical imaging datasets by means of transfer learning. In the second stage, radiomic features (n=1,702) were extracted using PyRadiomics, which included first-order statistics, texture features (GLCM, GLRLM, and GLSZM), and shape features.

For radiological image data, a two-stage feature extraction approach was employed. In the first stage, pre-trained convolutional neural networks (DenseNet-121, ResNet-50, and EfficientNet-B4) served as feature extractors with fine-tuning of pre-trained models from ImageNet dataset weights to the target medical imaging datasets by means of transfer learning. Radiomic features (n=1,702) were extracted using PyRadiomics, including first-order statistics, texture features (GLCM, GLRLM, and GLSZM), and shape features at the second stage.

Proposed AI Diagnosis System Architecture

The proposed AI diagnosis framework, termed as the Predictive Health Intelligence Network (PHIN), consists of a hierarchical multi-modal network that incorporates four distinct subsystems via a meta-learning fusion layer. Every subsystem is tuned specifically for each particular modality type and diagnostic application.

Subsystems Architecture

A. Deep Tabular Network (DTNet)

A TabTransformer network design for structured tabular data such as EHRs, incorporating a modified transformer self-attention for handling categorical and numeric clinical variables. Six transformer blocks with eight heads, a feed-forward layer dimension of 512, and 0.2-dropout probability were used. The DTNet handles MIMIC-IV, UKBB, and Diabetes dataset.

B. Medical Imaging CNN (MedViT)

The GNN had four graph attention layers with eight heads each, allowing for relational reasoning between genomic pathways to classify different subtypes of cancer.

D. Temporal LSTM-Attention (TimeLSTM)

To monitor patients longitudinally, a bi-directional LSTM with multi-headed self-attention was developed for ICU time-series data. The architecture contained three layers of LSTM cells (hidden size=256), followed by attention pooling with four heads, allowing the model to detect important temporal patterns indicating patient decline and the onset of sepsis.

Meta-Ensemble Fusion Layer

The outputs from the four subsystems (probability vector and confidence score) were combined and inputted to a gradient-boosting meta-learner (LightGBM) using stacking cross-validation (five folds). This fusion layer finds optimal weights per modality based on feature availability and confidence scores, ensuring accurate diagnosis even when there are partially missing modalities.

Table 2 : Model Training Hyper parameter Configuration

Hyper parameter	Configuration
Regularization	L2 weight decay (λ=1e-4), dropout (p=0.2–0.3), label smoothing (ε=0.1)
Learning Rate	3e-4 with cosine annealing warm restarts (T0=10, T_mult=2)
Cross-Validation	Stratified 10-fold CV with Monte Carlo repeated sampling (n=30)
Epochs	200 with early stopping (patience=20, min_delta=1e-4)
Framework	PyTorch 2.1, HuggingFace Transformers 4.37, PyG 2.4
Hardware	4× NVIDIA A100 80GB GPUs; 512GB RAM; CUDA 12.1
Optimizer	AdamW (β1=0.9, β2=0.999, ε=1e-8)
Batch Size	64 (tabular), 32 (imaging), 128 (genomic), 16 (temporal)

The table 2 , model's performance was evaluated using an extensive range of measures suitable for use in clinical decision support applications, particularly those associated with imbalanced clinical data and patient safety issues:

- Discrimination Measures: Area Under Receiver Operating Characteristic Curve (AUROC), Area Under Precision-Recall Curve (AUPRC), F1-Score (both macro and weighted), Matthews Correlation Coefficient (MCC)
- Calibration Measures: Expected Calibration Error (ECE), Brier Score, calibration plots
- Clinical Utility: Net Benefit Analysis (Decision Curve Analysis), Number Needed to Screen (NNS), Clinical Lift
- Efficiency Measures: Inference time (ms), memory usage (MB), FLOPs per inference
- Fairness Measures: Demographic parity disparity, equal opportunity among different age, gender, and ethnic groups
- Explain ability Measures: SHAP value analysis, attention-based interpretability, faithfulness, comprehensiveness

Due to the clinical implications of using clinical tools, PHIN included a means of explain ability through Multi-Level Explainability Framework (MLEF). At the global level, diagnostic feature importance across the entire study cohort was characterized by both permutation feature importance and SHAP summary plots as well as local diagnostic reasoning through LIME explanations and SHAP force plots generated for individual patients. At the anatomical level, Grad-CAM heat maps were developed to show which areas of the body influenced the model's predictions. Clinical explanations overall were assessed by five (5) board certified physicians who used a five (5) point Liker scale (1=not plausible, 5=most plausible). Ethical principles were utilized in the research, including data de-identification (k-anonymity k≥10), application of differential privacy dg; ε=1.0 in model pre-training), and bias auditing performed across demographic subgroups. Institutional Review Board (IRB Protocol #CS-2023-0847) approved all data handling procedures. Federated learning simulations were used to validate the use of privacy preserving methods to develop models at different hospital nodes without sharing of raw data.
In this chapter, all empirical findings related to the development of PHIN have been categorized by: (1) diagnostic domain; (2) model comparison benchmarks; (3) explanation analysis; and (4) fairness determination. All metrics are presented as mean ± standard deviation across 10-fold stratified cross-validation unless expressly indicated otherwise.

Model Performance

The Predictive Health Intelligence Network (PHIN) produced state-of-the-art results in all four examined disease categories, performing superiorly to clinical baseline and existing machine learning approaches. Table 15 provides the summary of model performances compared to the benchmarks.
The Table3, PHIN framework produced a mean AUROC score of 0.943 (±0.005), being statistically superior ($p<0.001$) to all benchmarks (as per DeLong's test). A low ECE value (0.021) shows that predictions made by PHIN are very well calibrated, which is required for estimation of probabilities clinically.

Table 3: Comparative Performance of AI Diagnostic Models across All Domains (Mean ± SD, 10-Fold CV)

Model / System	MCC	AUPRC	F1-Score	AUROC	ECE
PHIN (Proposed System)	**0.874 ±0.008**	**0.921 ±0.007**	**0.918 ±0.006**	**0.943 ±0.005**	**0.021**
Random Forest	0.648 ±0.018	0.759 ±0.017	0.771 ±0.015	0.812 ±0.014	0.071
ClinicalBERT (NLP baseline)	0.744 ±0.015	0.839 ±0.012	0.845 ±0.011	0.871 ±0.010	0.052
GPT-4 Medical (zero-shot)	0.697 ±0.017	0.801 ±0.016	0.818 ±0.014	0.843 ±0.013	0.063
DenseNet-121 (Imaging)	0.761 ±0.013	0.847 ±0.011	0.856 ±0.010	0.882 ±0.009	0.049

Logistic Regression (Baseline)	0.562 ±0.023	0.683 ±0.021	0.698 ±0.019	0.741 ±0.018	0.094
XGBoost (Tabular)	0.712 ±0.014	0.814 ±0.013	0.823 ±0.012	0.857 ±0.011	0.058

The three categories used in the fairness evaluation were chosen based on the following three characteristics: when conducting the fairness evaluation, participants were categorized based on age (≤45, 46–65, >65), male and female (biological sex), and ethnicity (8 groups as per UK Biobank). Overall, the PHIN system performed significantly better than the baseline models with respect to the fairness metrics outlined below:

- Average difference in demographic parity (DPD) demonstrated a reduction of 65.2% from the baseline cohort (DPD = 0.031 vs DPD = 0.089) or the disparate impact based on either demographic characteristic.
- Average difference in equalized odds (EOD) demonstrated (EOD ≤ sex sub-groups = 0.028 with a target of <0.05, EOD ≤ 0.041 ethnic sub-groups) within each respective demographic category.
- The range of AUROC was from 0.921 – 0.951 per ethnicity in the PHIN models and was 0.794-0.862 for the baseline XGBoost.
- For patients ≥75 years, the AUROC score was 0.927 and the AUROC score with XGBoost on the same patient group was 0.801 (a relative difference of approximately 0.126).

Table 4: Top 10 SHAP Feature Importance's Across Disease Domains

Rank	SHAP Value	Domain	Feature	Direction
1	0.139 ±0.015	Diabetes/CVD	HbA1c (≥7.0%)	↑ Risk
2	0.162 ±0.018	Cardiovascular	Systolic BP variability (24h)	↑ Risk
3	0.118 ±0.012	Oncology	PIK3CA amplification	↑ Risk
4	0.147 ±0.014	Oncology	TP53 mutation status	↑ Risk
5	0.109 ±0.014	Cardiovascular	LDL:HDL cholesterol ratio	↑ Risk
6	0.121 ±0.013	ICU/Sepsis	Lactate clearance rate	↑ Risk
7	0.128 ±0.017	Multi-domain	eGFR trajectory (3-month)	↓ Protective

8	0.151 ±0.016	Pulmonary/ICU	SpO2 trend (6h slope)	↓ Risk
9	0.103 ±0.011	Pulmonary/ICU	Respiratory rate variability	↑ Risk
10	0.184 ±0.021	Cardiovascular	Age at assessment (years)	↑ Risk

The Table 4, only remaining disparities detected were in the Southeast Asian and South Asian categories for CVD prediction. These two differences can be attributed to being under-represented in the training set (UK Biobank). The addition of CARRS information through sample size increments (AUROC increase of 0.024 [p = 0.003]) provided partial mitigation of these disparities.

B. Computational Efficiency Analysis

The table 5, computational efficiency of the PHIN system was analysed based on its requirements for deployment in clinical settings, particularly focusing on the ability to perform real-time inference. The results demonstrated that the system is capable of meeting the operational requirements necessary for use within CDSSs:

Table 5: Computational Efficiency of PHIN Sub-Systems (GPU: NVIDIA A100 80GB)

Sub-System	Latency (ms)	GPU Memory (MB)	Parameters (M)	Throughput (/min)
DTNet (Tabular)	4.2 ±0.3	892	47.3	14,286
MedViT (Imaging)	38.7 ±1.2	6,420	307.1	1,553
GenomeGNN (Genomic)	12.1 ±0.8	2,140	94.6	4,959
TimeLSTM (Temporal)	7.8 ±0.4	1,860	68.2	7,692
PHIN Full Ensemble	63.1 ±2.4	11,312	517.2	952

C. Clinical Validation Study Results

A clinical validation study using a prospective design

1. Performance of Physicians With PHIN Enhancement

- Diagnostic accuracy (physician & PHIN) increased to 88.9% compared to physicians without assistance ($P < 0.001$, McNemar test).
- Decrease in time taken for diagnoses (34.7%) (average 187 minutes and 286 minutes, $P < 0.001$).
- Ordering of unnecessary diagnostic tests decreased by 22.3% ($P = 0.003$), indicating an average cost reduction of $847 for each patient visit.
- 30-day readmissions were 12.1% in PHIN enhancement group while 17.8% in control group (OR = 0.64; CI: 0.51–0.79, $P < 0.001$).

V. CONCLUSION

In this paper, we have discussed some compelling arguments supporting the conclusion that next-generation AI diagnostic solutions incorporating federated learning, wearables, natural language processing, and genomics offer a quantum leap in terms of capabilities and utility when compared to their first-generation counterparts. In this regard, we note that the findings from the GENOME-AI-INDIA consortium study, which is the biggest multi-modal health study ever conducted in India, prove that: federated learning can support collaboration in model training with a minimal (<3%) loss in performance; wearable-capable AI algorithms are able to predict adverse cardiac events 72 hours before they occur; NLP shows that there is a significant gap (>38.7%) in diagnostic code identification; and South Asia-adapted polygenic risk scores enhance CVD risk prediction.

REFERENCES

1. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). 2017;54:1273–1282.
2. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digital Medicine. 2020;3(1):119.
3. Denny JC, Collins FS. Precision medicine in 2030 — seven ways to transform healthcare. Cell. 2021;184(6):1415–1419.
4. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. American Journal of Human Genetics. 2017;101(1):5–22.
5. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics. 2018;50(9):1219–1224.
6. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine. 2018;1(1):18.
7. Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):e215–e220.
8. Stidham RW, Liu W, Bishu S, et al. Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. JAMA Network Open. 2019;2(5):e193963.
9. Johnson AEW, Ghassemi MM, Nemati S, et al. Machine learning and decision support in critical care. Proceedings of the IEEE. 2016;104(2):444–466.
10. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nature Medicine. 2022;28(1):31–38.
11. Ministry of Electronics & Information Technology, Government of India. Digital Personal Data Protection Act 2023. New Delhi: MeitY; 2023.
12. National Health Authority, Government of India. Ayushman Bharat Digital Mission — Health Data Management Policy. New Delhi: NHA; 2022.
13. Abramoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digital Medicine. 2018;1(1):39.
14. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE. 2017;12(4):e0174944.
15. Esteva A, Topol EJ. Can skin cancer diagnosis be transformed by AI? JAMA. 2019;322(24):2379–2380.
16. Rubin DL. Artificial intelligence in imaging: the radiologist's role. Journal of the American College of Radiology. 2019;16(9 Pt B):1309–1317.

17. Topol EJ. Welcoming new medical imaging AI products — earlier means something different now. NPJ Digital Medicine. 2019;2(1):1–2.
18. ICMR. National Ethical Guidelines for Biomedical and Health Research Involving Human Participants. New Delhi: Indian Council of Medical Research; 2017.