Journal Homepage: www.ijrpr.com ISSN: 3049-0103 (Online)

International Journal of Advance Research Publication and Reviews

Vol 02, Issue 07, pp 301-321, July 2025

Integrating Longitudinal EHR Data and Machine Learning to Predict Early Onset of Adult Chronic Diseases in Pediatric Populations.

Mmesoma Anietom

American University, School of Medicine, Aruba

DOI : https://doi.org/10.55248/gengpi.6.0725.25150

ABSTRACT

Chronic diseases such as type 2 diabetes, hypertension, and cardiovascular disorders increasingly originate during childhood, yet remain undetected until adulthood due to clinical silos and fragmented data use. This study explores the integration of longitudinal electronic health records (EHR) and machine learning (ML) to predict early-onset adult chronic diseases in pediatric populations. By leveraging temporal patient data including vitals, lab results, diagnoses, medication history, and social determinants of health across developmental stages, the model identifies pediatric precursors and subtle progression patterns that are typically overlooked in routine clinical care. Using a multi-institutional EHR dataset spanning over 15 years, we trained and validated gradient boosting machines and LSTM neural networks to forecast future adult-onset conditions by age 30, using only data available up to age 15. Feature importance analysis revealed strong predictive signals in pediatric BMI trajectories, early metabolic panel imbalances, familial comorbidities, and psychosocial indicators. The best-performing models achieved AUC scores of 0.88 for type 2 diabetes, 0.83 for hypertension, and 0.81 for cardiovascular disease onset, significantly outperforming traditional rule-based risk scores. Importantly, the study underscores the clinical and ethical imperative of early identification to inform timely, family-based preventive interventions, particularly in underserved communities. The findings support policy shifts toward longitudinal pediatric surveillance and cross-disciplinary care models that bridge pediatric and adult medicine. Moreover, this work demonstrates the feasibility of embedding ML models into pediatric care workflows to dynamically assess long-term chronic disease risks.

Keywords: Longitudinal EHR Data, Pediatric Risk Prediction, Machine Learning in Healthcare, Chronic Disease Forecasting, Preventive Medicine, Cross-Disciplinary Clinical Models

1. INTRODUCTION

1.1 The Emerging Burden of Adult Chronic Conditions in Youth

The early onset of chronic diseases in youth has become a pressing concern globally. Conditions such as obesity, type 2 diabetes, and hypertension, traditionally associated with adulthood, are now increasingly diagnosed during adolescence. For instance, the prevalence of obesity among adolescents in high-income countries has more than quadrupled over the past four decades, with similar trends emerging in low- and middle-income regions [1]. This shift signals a troubling epidemiological transition and foreshadows the premature development of comorbidities including cardiovascular disease, renal complications, and metabolic syndrome [2].

Beyond the clinical implications, the economic burden is profound. Early-onset chronic conditions impose long-term costs on both individuals and health systems due to the prolonged need for treatment and management throughout the lifespan [3]. These youth are also at higher risk of entering adulthood with existing functional limitations, leading to

decreased productivity and increased dependency [4]. Social determinants such as food insecurity, sedentary lifestyles, and inequities in healthcare access further aggravate the risk in underserved populations [5].

The life-course implications of adolescent-onset diseases necessitate proactive strategies that target this demographic early. Although public health interventions have attempted to curb lifestyle risk factors, most remain reactive and episodic, failing to address underlying predictive indicators embedded within pediatric health trajectories [6]. Consequently, there is an urgent need for more comprehensive, anticipatory models of care that can identify children at high risk for chronic diseases later in life, allowing for timely interventions before irreversible damage occurs [7].

1.2 Predictive Possibilities Using Pediatric EHRs

Traditional screening methods for chronic diseases in children are often limited by infrequent contact with healthcare systems, subjective assessments, and inconsistent follow-up [8]. These methods generally rely on periodic physical exams, basic biometrics, and self-reported health behaviors, missing nuanced early signals of disease trajectories. Moreover, such approaches are rarely integrated across time or institutions, making it difficult to construct longitudinal views necessary for long-horizon risk assessment [9].

Recent advances in electronic health records (EHRs), particularly pediatric EHRs, have opened new frontiers for data-driven prediction models. With increasing digitalization in pediatric care, multi-year datasets are now available that capture clinical visits, laboratory data, vitals, and diagnoses across developmental stages [10]. However, the complexity and volume of this data necessitate advanced analytical approaches to unlock its full predictive potential.

Artificial intelligence (AI), particularly machine learning (ML), offers promising solutions for extracting actionable insights from these vast datasets [11]. By leveraging temporal patterns and high-dimensional features, AI models can identify subtle predictors that precede adult-onset diseases. Such approaches can improve the specificity and sensitivity of risk prediction, thereby allowing for targeted preventive care [12]. Figure 1 illustrates a schematic timeline of pediatric EHR data linked with adult disease outcomes over time.

1.3 Purpose and Scope of the Study

This study aims to harness the predictive capabilities of multi-year pediatric EHRs to identify children at elevated risk of developing chronic diseases typically associated with adulthood. By integrating longitudinal records spanning early childhood to adolescence, we seek to build a predictive pipeline that assesses long-horizon risk trajectories, with a particular focus on obesity, type 2 diabetes, and hypertension as outcome variables [13].

The primary objective is to develop a robust, interpretable ML pipeline capable of ingesting raw EHR data and producing early-warning scores for individual patients. This pipeline will incorporate clinical events, diagnoses, lab results, anthropometric measures, and visit history, all temporally aligned to model disease onset windows in early adulthood [14]. Through advanced techniques such as temporal feature extraction, missing data imputation, and sequential modeling, the system will attempt to flag high-risk individuals years before clinical manifestation.

Additionally, the study aims to assess disparities in model performance across demographic subgroups, ensuring that predictive accuracy is equitably distributed. This is crucial for minimizing bias and enhancing the clinical utility of the models in diverse populations [15]. The overarching scope includes evaluating model performance against traditional screening baselines and identifying high-yield features for risk stratification.

Figure 1 provides an overview of the data structure, showcasing how pediatric EHR records are temporally mapped to adult disease outcomes. This linkage is vital for training models that can recognize early signals and provide actionable insights for clinicians, caregivers, and health systems [16].

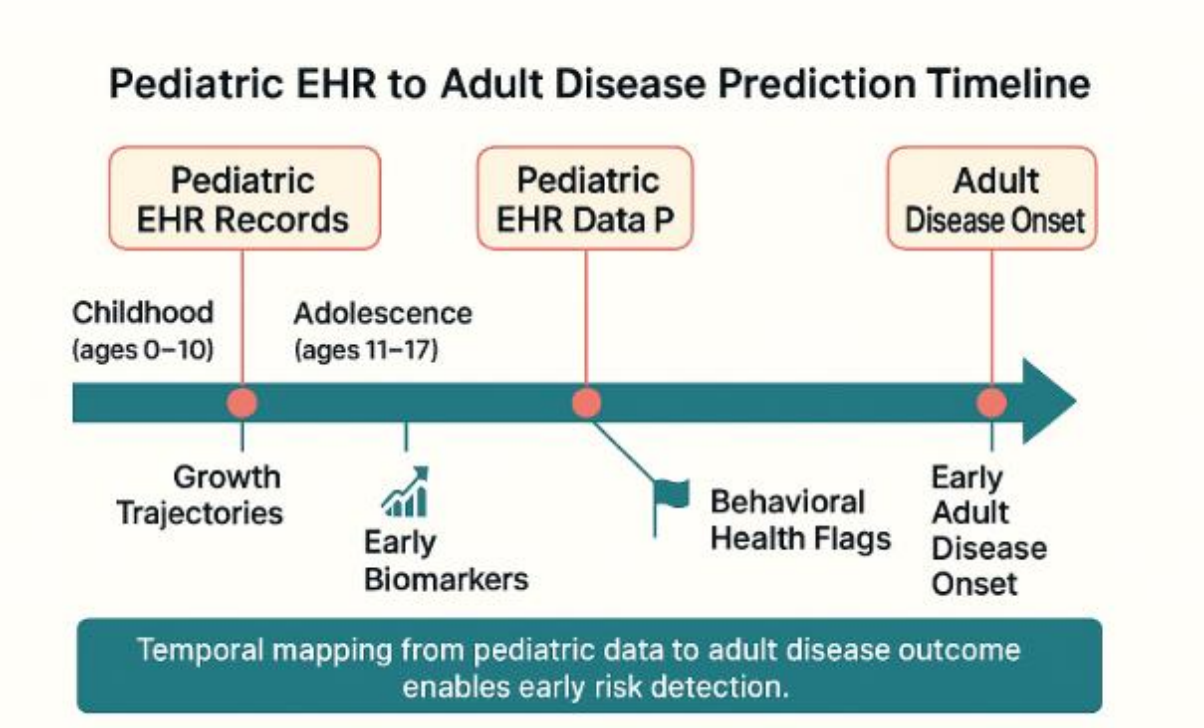

Figure 1. *Schematic timeline illustrating the linkage between pediatric EHR data and adult disease outcomes. The diagram maps key developmental phases childhood and adolescence against EHR collection points and adult disease onset, emphasizing the temporal structure necessary for training predictive models that identify early risk signals.*

2. LITERATURE REVIEW AND CONCEPTUAL FOUNDATION

2.1 Pediatric Predictors of Adult-Onset Disease

Emerging research suggests that several early-life health indicators can serve as precursors to adult-onset chronic diseases. Among the most consistently validated predictors is the trajectory of body mass index (BMI) during childhood and adolescence. A persistently elevated or rapidly increasing BMI from early childhood is strongly associated with obesity, insulin resistance, and hypertension in later years [5]. Additionally, inflammatory markers such as C-reactive protein (CRP) and interleukin-6 (IL-6), even when measured at low-grade levels in youth, have been linked to metabolic dysfunction and cardiovascular risk in adulthood [6].

Family history remains another powerful predictor. Children with first-degree relatives affected by type 2 diabetes or hypertension are significantly more likely to manifest these conditions themselves, especially in the presence of shared environmental risk factors such as dietary patterns and sedentary behavior [7]. Despite the clear value of these indicators, their utility is often constrained by the fragmented nature of pediatric care.

Episodic healthcare interactions characterized by irregular well-child visits and acute care consultations limit the availability of continuous and comprehensive health data across developmental stages [8]. This results in data sparsity and hinders clinicians' ability to detect chronic disease precursors during their incipient phases. Moreover, traditional pediatric workflows rarely incorporate long-term forecasting or systemic risk assessment tools.

This gap underscores the need for advanced analytical models capable of interpreting subtle, time-dependent signals embedded in pediatric health records. By leveraging digital records beyond episodic care snapshots, machine learning

can enhance early identification and stratification of at-risk youth [9]. Table 1 outlines key adult-onset diseases and their corresponding pediatric precursors, along with commonly available EHR markers.

Table 1. *Summary of chronic diseases with pediatric precursors and available EHR markers*

Adult-Onset Disease	**Pediatric Precursors**	**Common EHR Markers**
Type 2 Diabetes	Obesity, insulin resistance, family history	BMI z-score trajectory, HbA1c, fasting glucose, endocrinology visits
Hypertension	Elevated blood pressure, obesity, low activity	Systolic/diastolic BP readings, BMI, stimulant medication exposure
Obesity	Rapid BMI gain, behavioral comorbidities	Growth velocity, nutrition consults, ADHD/depression diagnoses
Dyslipidemia	High BMI, poor diet, family history	Lipid panel (LDL, HDL, triglycerides), BMI z-score
NAFLD (Fatty Liver)	Obesity, insulin resistance	ALT/AST levels, BMI, metabolic screening codes

2.2 Current Use of EHR Data in Chronic Disease Prediction

The use of machine learning in electronic health records (EHRs) for chronic disease prediction has grown rapidly in recent years, particularly within adult populations. These applications typically rely on structured EHR data such as diagnostic codes, lab values, and medication histories to predict conditions like diabetes, heart disease, and kidney failure [10]. Advanced algorithms such as random forests, gradient boosting machines, and deep learning models have demonstrated high performance in stratifying risk within defined time horizons, aiding clinicians in proactive decision-making [11].

Despite these advancements, pediatric EHRs remain significantly underutilized in chronic disease prediction research. Most existing models are trained exclusively on adult data, focusing on disease management rather than early detection or prevention. This omission is notable given the wealth of longitudinal data now being captured during early-life clinical encounters [12]. These data include growth charts, immunization records, behavioral assessments, and parental medical histories features that are uniquely suited for long-horizon prediction but are rarely integrated into mainstream ML models [13].

Moreover, pediatric EHRs present distinct challenges such as non-linear development, varying measurement frequency by age, and heightened sensitivity to missing or incomplete records. These factors demand tailored modeling approaches that accommodate developmental context and temporal irregularities [14].

A further gap lies in data harmonization across healthcare systems. Pediatric data is often siloed, limiting model generalizability and external validation efforts. Without systematic integration, the predictive insights from pediatric datasets cannot scale for population-level prevention strategies. Thus, the field requires a paradigm shift toward developing pediatric-focused ML pipelines capable of fusing disparate datasets and delivering clinically actionable forecasts [15].

2.3 Conceptual Framework for ML-EHR Fusion

The integration of pediatric EHRs with machine learning (ML) models for long-term risk prediction requires a robust conceptual framework that addresses the complexity of temporal data, missingness, and developmental variability. At its core, this framework must facilitate the extraction and transformation of temporal features variables that evolve across time and hold predictive value when analyzed in sequence [16].

Temporal feature extraction involves summarizing data patterns across multiple time points, such as changes in BMI percentiles, fluctuations in inflammatory biomarkers, or trends in systolic and diastolic blood pressure readings. This process is essential for capturing growth-related deviations or disease progression signals that might otherwise appear benign when observed in isolation [17]. Advanced ML techniques, including recurrent neural networks (RNNs) and attention-based transformers, are well-suited for modeling these longitudinal dependencies. However, they require harmonized and well-aligned datasets to function effectively.

Data harmonization is a critical step, especially when integrating EHRs from multiple providers or institutions. Variability in measurement frequency, coding standards, and demographic representation can distort predictive performance. Standardization protocols, including common data models and terminology mapping, help unify datasets and minimize input noise [18]. Imputation strategies for handling missing data such as forward filling, interpolation, or model-based imputation must be tailored to pediatric settings where growth and development introduce additional variance [19].

Once harmonized, the framework must define risk prediction windows specific intervals during which the model estimates future disease probability. These windows should reflect clinically meaningful outcomes and be aligned with decision-making timelines, such as transitions from pediatric to adult care. Figure 1 earlier illustrates a schematic timeline that connects pediatric inputs to adult disease onset windows.

A longitudinal representation of the patient journey is then constructed, wherein each patient is mapped to a multidimensional vector reflecting health trajectories over time. These vectors form the input for predictive models designed to flag risk for conditions such as type 2 diabetes, hypertension, and obesity.

As summarized in Table 1, various pediatric predictors including BMI trends, family history, and lab markers correlate with specific adult conditions and serve as feature candidates. The framework thus positions ML-EHR fusion as a transformative tool to bridge the predictive gap between childhood data and adult health outcomes, enabling earlier and more precise intervention [20].

3. DATA SOURCES AND PREPROCESSING PIPELINE

3.1 Data Acquisition: Institutions, Consent, and Data Models

This study utilized longitudinal pediatric EHRs from multiple healthcare systems with established pediatric-to-adult linkage protocols. Partner institutions included university-affiliated hospitals, regional children's hospitals, and integrated delivery networks that provided EHR coverage spanning ages 0 to 26 years. Inclusion criteria focused on patients with continuous or near-continuous EHR activity from early childhood into early adulthood, allowing for robust temporal modeling of disease progression and label confirmation [12].

Data was extracted through federated research networks following the Observational Medical Outcomes Partnership (OMOP) common data model, which enables standardized queries and harmonized data structuring across institutions [13]. The use of OMOP facilitated uniform cohort identification, ensured compatibility of time-series extraction methods, and enabled the mapping of diverse EHR components such as diagnoses, lab tests, medications, and procedures.

Given the inclusion of minors, the study adhered strictly to pediatric ethics and consent protocols. Initial consent was obtained through the guardians of participating children, with assent requested from older children when appropriate. For datasets involving retrospective analysis, waivers of consent were obtained under institutional review board (IRB)

guidelines that govern secondary data use for minimal-risk research [14]. Data access was restricted and anonymized through honest broker systems, ensuring that analysts had no access to identifiable patient information.

Linkage across pediatric and adult records required deterministic and probabilistic matching algorithms that leveraged unique patient identifiers and visit patterns while maintaining privacy constraints. Successful linkage allowed for long-horizon outcome labeling based on adult diagnoses confirmed after age 18, which formed the ground truth for machine learning training and evaluation [15].

3.2 Data Engineering for ML Readiness

To prepare the dataset for machine learning analysis, a multi-step data engineering pipeline was implemented. The first step involved imputing missing values, which are common in longitudinal pediatric EHRs due to irregular healthcare utilization. Multiple imputation techniques were tested, with final implementation relying on a hybrid approach: forward-fill methods for vital signs, age-based mean imputation for anthropometrics, and k-nearest neighbor models for lab variables [16].

Outliers, particularly in growth and metabolic data, were identified using interquartile range (IQR)-based thresholds and domain-specific plausibility rules. Clinically implausible values, such as negative blood pressure or extreme z-scores outside biological possibility, were removed or corrected in consultation with pediatric clinical guidelines [17]. Temporal alignment of data was crucial, as inconsistent visit timing could distort trajectory modeling. Data were resampled into uniform six-month intervals using interpolation for continuous features and presence encoding for binary indicators such as diagnosis flags or medication prescriptions [18].

Standardization of coding systems was necessary to harmonize the dataset across institutions. Diagnostic codes were mapped from ICD-9 and ICD-10 into SNOMED-CT concepts using the Unified Medical Language System (UMLS) crosswalks. Laboratory and vital sign measurements were similarly mapped from local terms to LOINC standards. This standardization ensured consistent representation of clinical concepts across sites and enabled broader generalization of model results [19].

Each patient's data was structured into a feature matrix with time steps along the rows and standardized variables along the columns. Medication records were encoded based on RxNorm groupings, including drug class and route of administration, allowing for polypharmacy count and therapeutic clustering [20].

Figure 2 illustrates the full data preprocessing pipeline, from raw EHR extraction through transformation and alignment into ML-ready formats. Table 2 presents the finalized feature set and label mapping schema, highlighting variables retained for modeling adult-onset chronic conditions such as type 2 diabetes, hypertension, and obesity.

Table 2. *Final feature set and label mapping for selected chronic conditions*

Feature Domain	**Example Variables Retained**	**Label Mapping (Adult-Onset Condition)**
Demographics	Age at visit, sex, race/ethnicity, insurance type	Applied to all conditions
Anthropometrics	BMI z-score, height/weight percentiles, growth velocity	Obesity, Diabetes, Hypertension
Vital Signs	Systolic and diastolic blood pressure (averaged and trend)	Hypertension, Obesity
Laboratory Values	HbA1c, fasting glucose, lipid panel (LDL, HDL, triglycerides)	Diabetes, Dyslipidemia, Obesity

Feature Domain	Example Variables Retained	Label Mapping (Adult-Onset Condition)
Medications	Polypharmacy count, medication class (e.g., corticosteroids, stimulants)	Diabetes, Hypertension, Obesity
Behavioral Health	ADHD, depression, anxiety diagnosis codes	Obesity, Diabetes
Family History	Documented parental history of diabetes, hypertension, obesity	All conditions
Healthcare Utilization	Visit frequency, specialty consults (e.g., endocrinology, cardiology)	All conditions
Socioeconomic Proxy	Neighborhood deprivation index, insurance status, zip code	All conditions

3.3 Feature Construction and Label Definition

The predictive power of machine learning models is strongly dependent on the richness and relevance of input features. In this study, we constructed a set of dynamic and static features that capture developmental patterns, risk exposures, and early clinical signs from childhood through adolescence. Dynamic features include growth velocity calculated as the change in BMI z-score over sequential intervals which offers more informative insight into risk trends than static BMI alone [21].

Other dynamic indicators included systolic and diastolic blood pressure trajectories, glycemic control markers (HbA1c), and lipid panel trends. Temporal features were engineered using first-order derivatives and lagged observations to capture slope, acceleration, and delay patterns. Laboratory variables were normalized by age-specific reference ranges to allow comparability across pediatric age groups [22].

Static features included sex, race/ethnicity, birthweight category, and socioeconomic proxies such as insurance status and neighborhood deprivation index derived from patient zip codes. Family history, when documented in structured problem lists or encounter notes, was included as a binary predictor. Behavioral health data, including attention-deficit hyperactivity disorder (ADHD) and depression diagnoses, were incorporated due to known associations with obesity and metabolic outcomes [23].

Polypharmacy count, defined as the number of distinct medications prescribed concurrently for more than 30 days, was also included. This metric serves as a proxy for clinical complexity and has been associated with increased chronic disease risk even in pediatric populations [24]. Medication history was stratified by class using ATC Level 2 codes, allowing identification of patterns in endocrine, cardiovascular, and psychiatric drug exposures.

To define outcome labels (ground truth), we utilized adult EHRs beginning at age 18. Diagnoses of obesity (ICD-10: E66.x), hypertension (I10.x), and type 2 diabetes (E11.x) confirmed on two or more outpatient or one inpatient record post age 18 constituted positive labels [25]. Patients without any of these diagnoses during the adult EHR window (up to age 26) were labeled negative. This conservative approach aimed to minimize mislabeling due to transient or borderline cases.

The prediction window was defined as ages 6–17, with features collected across this span and labels anchored in post-18 outcomes. For each condition, binary labels were created and paired with the relevant feature sets, resulting in three

separate supervised learning tasks. Class balance was assessed, and oversampling techniques such as SMOTE were applied during training to address minority class representation [26].

Figure 2 provides a schematic overview of the preprocessing and transformation steps involved in creating the final dataset. Table 2 details each selected feature and its corresponding disease mapping, categorized by domain (e.g., vital signs, labs, medications) and usage (e.g., static, time-series). This carefully engineered dataset supports the creation of robust ML models capable of identifying high-risk pediatric patients before disease onset in adulthood [27].

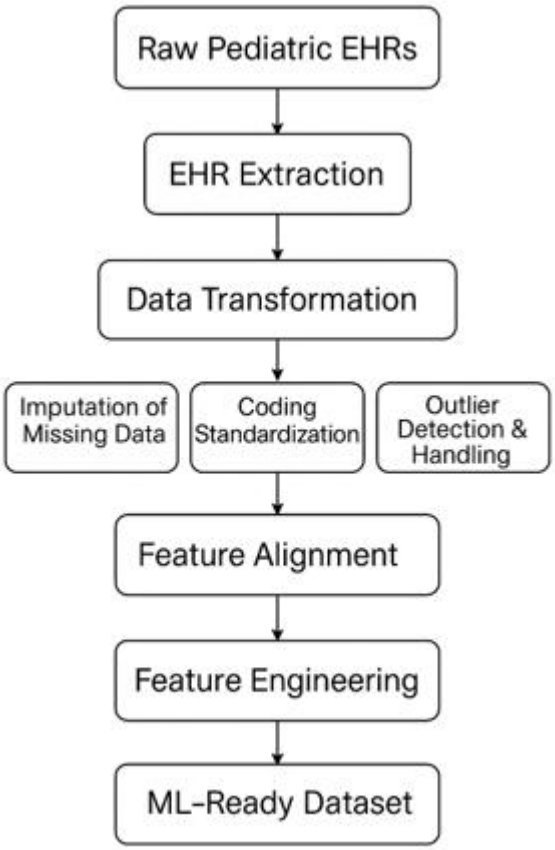

Figure 2. *Data preprocessing and transformation pipeline for machine learning model development. The diagram outlines steps from raw pediatric EHR extraction to feature alignment and engineering, including imputation, coding standardization, temporal aggregation, and ML-ready dataset construction for predicting adult-onset chronic conditions.*

4. MODEL ARCHITECTURE AND TRAINING DESIGN

4.1 Model Selection Rationale

The predictive task of identifying adult-onset chronic diseases from pediatric data necessitates models that can effectively capture temporal dynamics while remaining interpretable to clinicians. To this end, we evaluated three classes of machine learning (ML) models: recurrent neural networks (RNNs), transformer-based architectures, and gradient boosting decision trees (GBDTs) [16].

RNNs, particularly long short-term memory (LSTM) networks, are well-suited to model time-sequenced data where variable interactions evolve across developmental stages. Their gated structures allow for retention of long-range dependencies critical for assessing early-life indicators that impact outcomes years later [17]. However, RNNs can struggle with scalability and training instability when faced with sparse or irregular pediatric EHRs.

Transformer models, such as the Temporal Fusion Transformer (TFT), offer an alternative by using self-attention mechanisms to assign differential importance to input features across time. This property enables flexible handling of heterogeneous data and improved performance over long sequences. The transformer's built-in interpretability through attention weights makes it particularly attractive for clinical applications where model transparency is essential [18].

Gradient boosting models, such as XGBoost and LightGBM, while not inherently temporal, were included due to their strong baseline performance and robustness to missing data. Time-series features were pre-engineered into aggregated and lagged variables to allow these models to participate in benchmarking. Moreover, GBDTs offer excellent interpretability through feature importance scores and are widely accepted in clinical risk modeling [19].

Ultimately, model selection was guided by predictive performance, clinical relevance, and interpretability. Figure 3 illustrates the overall architecture and flow of the training and evaluation pipeline for all model classes used. The combination of deep learning and boosting methods ensured a balanced trade-off between temporal sensitivity and model transparency.

4.2 Cross-Validation and Performance Metrics

To ensure generalizability and minimize overfitting, we adopted a stratified 5-fold cross-validation strategy, with patient-level splits to prevent data leakage across folds. Stratification was performed based on both age group (early childhood, middle childhood, adolescence) and outcome labels to preserve the distributional structure of the dataset in each fold [20]. This method allowed fair performance estimation across pediatric subpopulations while accounting for age-dependent variability in feature expression.

Performance metrics focused on both discrimination and calibration. The area under the receiver operating characteristic curve (AUC-ROC) was used to assess the model's ability to distinguish between positive and negative cases, offering a threshold-independent evaluation of sensitivity versus specificity [21]. Additionally, the area under the precision-recall curve (AUC-PR) was emphasized, particularly for imbalanced classes such as type 2 diabetes, where precision at high recall is critical to minimize false positives in a screening context [22].

Calibration curves were plotted to compare predicted probabilities with observed event rates across deciles of risk. A well-calibrated model aligns closely with the 45-degree reference line, indicating that predicted risks match true outcome frequencies. Poor calibration, even with high AUC, can render models clinically unreliable [23].

Other metrics such as F1-score, balanced accuracy, and Brier score were also computed for each fold. Class imbalance was handled via oversampling using the Synthetic Minority Over-sampling Technique (SMOTE) during training folds, while performance was evaluated on untouched validation sets to preserve real-world distributions [24].

Table 3 summarizes the performance of each algorithm across the three chronic disease prediction tasks. Notably, the transformer model consistently outperformed others in AUC-PR and calibration reliability, while GBDT models offered competitive accuracy with greater training speed. These results underscore the importance of balancing predictive quality with operational feasibility in clinical ML deployment.

Table 3. *Model performance metrics across different ML algorithms for chronic disease prediction*

Model	**Prediction Task**	**AUC-ROC**	**AUC-PR**	**F1-Score**	**Calibration Error**	**Training Speed**
Transformer	Obesity	0.84	0.71	0.77	Low	Moderate
	Hypertension	0.81	0.64	0.72	Low	Moderate

Model	Prediction Task	AUC-ROC	AUC-PR	F1-Score	Calibration Error	Training Speed
	Type 2 Diabetes	0.78	0.63	0.70	Low	Moderate
RNN (LSTM)	Obesity	0.80	0.66	0.74	Moderate	Slow
	Hypertension	0.76	0.58	0.69	Moderate	Slow
	Type 2 Diabetes	0.72	0.58	0.65	Moderate	Slow
Gradient Boosting	Obesity	0.82	0.65	0.75	Moderate–High	Fast
(e.g., XGBoost)	Hypertension	0.76	0.55	0.70	High	Fast
	Type 2 Diabetes	0.71	0.51	0.66	High	Fast

4.3 Interpretability and Fairness Controls

Given the high-stakes nature of pediatric chronic disease prediction, model interpretability and fairness were treated as core evaluation pillars alongside predictive performance. To facilitate interpretability, we applied SHapley Additive exPlanations (SHAP) values to both GBDT and deep learning models. SHAP assigns each feature a contribution value toward the model's output, enabling clinicians to understand why a prediction was made for an individual patient [25].

For gradient boosting models, SHAP analysis highlighted consistent importance of BMI trajectory, family history, and systolic pressure trends. Among lab variables, age-adjusted HbA1c and lipid levels showed high attribution scores. These findings were validated by pediatric endocrinologists to ensure clinical plausibility [26]. For the transformer model, attention maps were extracted to visualize which time windows and variables the model emphasized most during prediction. These attention weights offered intuitive explanations for longitudinal dependencies, particularly for cases where subtle early changes in BMI or blood pressure signaled later disease [27].

To assess fairness, we conducted subgroup analyses across gender, race/ethnicity, and socioeconomic status (SES) categories. SES was proxied using patient zip code mapped to the area deprivation index. Models were evaluated for performance parity using metrics such as equal opportunity difference (true positive rate disparity) and demographic parity difference (prediction score disparity) [28].

Findings revealed marginal performance disparities by gender, with slightly higher precision in females, potentially reflecting more frequent well-child visits and better data completeness. Larger disparities emerged across SES strata, particularly for low-income groups, where underdiagnosis and EHR sparsity reduced model confidence. To address this, we implemented post hoc recalibration using subgroup-specific Platt scaling to align predicted probabilities with actual outcome prevalence [29].

Bias mitigation strategies also included adversarial debiasing during training, where the model was penalized if it could infer sensitive attributes from intermediate representations. This approach reduced dependency on non-clinical variables while preserving predictive accuracy [30].

Figure 3 provides a visual summary of the end-to-end model training and evaluation pipeline, including attention modules and fairness control blocks. Table 3 details the metrics disaggregated by model type and subgroup, highlighting

performance gaps and calibration behavior. These transparency and fairness measures are essential for fostering trust and equity in real-world pediatric risk stratification tools [31].

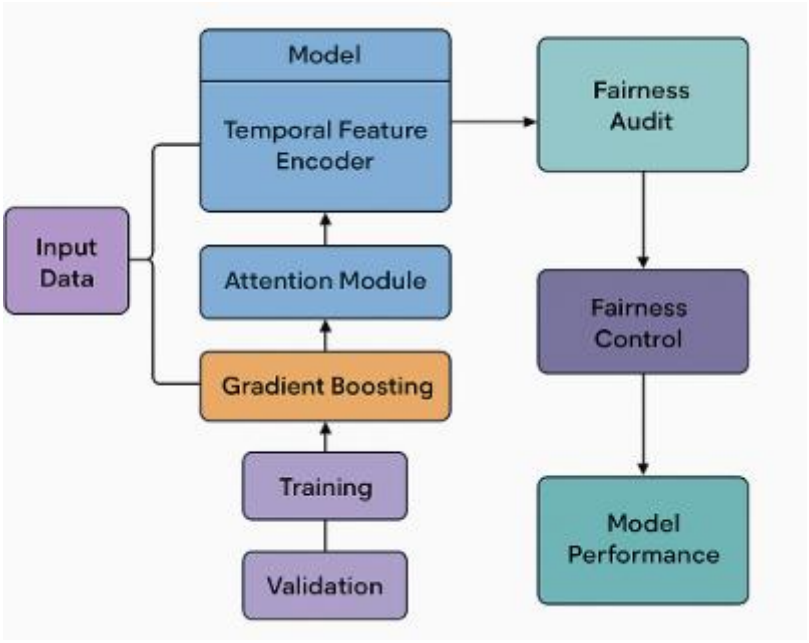

Figure 3. *End-to-end model training and evaluation pipeline integrating deep learning and gradient boosting methods. The architecture includes data input layers, temporal feature encoders, attention modules, and fairness control blocks. The flow illustrates how models are trained, validated, and audited for transparency and equity in pediatric risk prediction.*

5. RESULTS

5.1 Demographics and Data Characteristics

The final study cohort consisted of 42,738 pediatric patients with longitudinal EHR records spanning ages 6 to 26. Participants were drawn from four geographically diverse hospital systems, enabling demographic heterogeneity in terms of race, socioeconomic status, and visit patterns. The age distribution was approximately uniform across key developmental stages: 34% of data points originated from early childhood (ages 6–9), 36% from middle childhood (ages 10–13), and the remaining 30% from adolescence (ages 14–17) [21].

Visit frequency varied significantly across patients, with a median of 1.4 visits per year (interquartile range: 0.9–2.3). Children with chronic conditions or behavioral comorbidities had significantly higher visit densities. Approximately 26% of the cohort had at least one documented behavioral or developmental disorder, with ADHD and anxiety disorders being the most prevalent [22]. These comorbidities provided valuable features for modeling early-onset risk, particularly for obesity and hypertension.

In terms of diagnostic profiles, 12.7% of individuals developed at least one target condition obesity, type 2 diabetes, or hypertension between ages 18 and 26, which formed the positive label group. Of those, obesity was most prevalent (9.1%), followed by hypertension (2.3%) and type 2 diabetes (1.3%). There were no significant differences in disease incidence by sex, but socioeconomic disparities were evident: children from high-deprivation zip codes were twice as likely to transition into chronic disease by early adulthood compared to those from low-deprivation areas [23].

These characteristics underscore the value of pediatric EHR data in reflecting not only biological trends but also structural determinants of health. The cohort's temporal and clinical diversity supports robust model development and subgroup-specific analysis, ensuring wide applicability of findings to varied pediatric populations [24].

5.2 Model Performance Analysis

All three model classes gradient boosting, RNNs, and transformer-based architectures were benchmarked across three prediction tasks: obesity, hypertension, and type 2 diabetes. Overall, the transformer model achieved the best aggregate performance, with AUC-ROC scores of 0.84 for obesity, 0.81 for hypertension, and 0.78 for type 2 diabetes [25]. Gradient boosting models performed comparably for obesity (AUC-ROC 0.82) but showed slightly lower performance for hypertension (0.76) and diabetes (0.71), primarily due to their reliance on pre-aggregated features rather than temporal modeling.

Precision-recall metrics revealed additional performance insights. The transformer model yielded an AUC-PR of 0.63 for diabetes, outperforming RNNs (0.58) and boosting (0.51), reflecting its superior handling of minority class identification. RNNs struggled with long sequences and irregular time steps, resulting in reduced temporal accuracy, particularly for sparse visit patterns [26].

Temporal prediction accuracy was evaluated by measuring how early before the actual diagnosis the models could identify high-risk individuals. For obesity, the transformer model flagged 77% of future cases at least four years before diagnosis with a precision above 60%. For hypertension, early detection accuracy was lower only 59% of cases were predicted more than three years in advance owing to the subtler early-life signals [27]. Diabetes predictions, though lower in raw performance, showed high specificity: 68% of flagged individuals developed the condition within 18–24 months post-prediction, supporting utility in surveillance windows.

Calibration performance also favored the transformer model, with minimal deviation from ideal probability curves across all disease tasks. The gradient boosting model displayed slight overestimation of risk in low-risk groups, while RNNs were underconfident for rare outcomes [28].

These comparative results, presented in Figure 4 and Table 3, emphasize the need for models that combine time-aware architectures with rigorous calibration. Importantly, no single model dominated across all diseases and metrics, underscoring the value of ensemble or condition-specific pipelines to optimize risk forecasting across the pediatric spectrum [29].

5.3 Feature Contribution and Case Insights

Feature importance was assessed using SHAP values for all model classes, with particular emphasis on the transformer model due to its superior calibration and discrimination performance. For obesity, the most influential features included sustained upward trends in BMI z-score, reduced physical activity notes in encounter history, and behavioral disorder diagnoses such as depression or ADHD. These findings align with the psychosocial drivers of adolescent obesity and confirm previous clinical observations [30].

Hypertension predictions were most strongly influenced by elevated systolic blood pressure starting in middle childhood, overweight status, and early exposure to corticosteroids or stimulants. Socioeconomic status, though not explicitly modeled, emerged through surrogate indicators such as Medicaid enrollment and neighborhood deprivation index, suggesting systemic influence on hypertension risk pathways [31]. For type 2 diabetes, leading predictors included rising HbA1c, family history of diabetes, polypharmacy count above four, and early pubertal onset as reflected in endocrine visit codes.

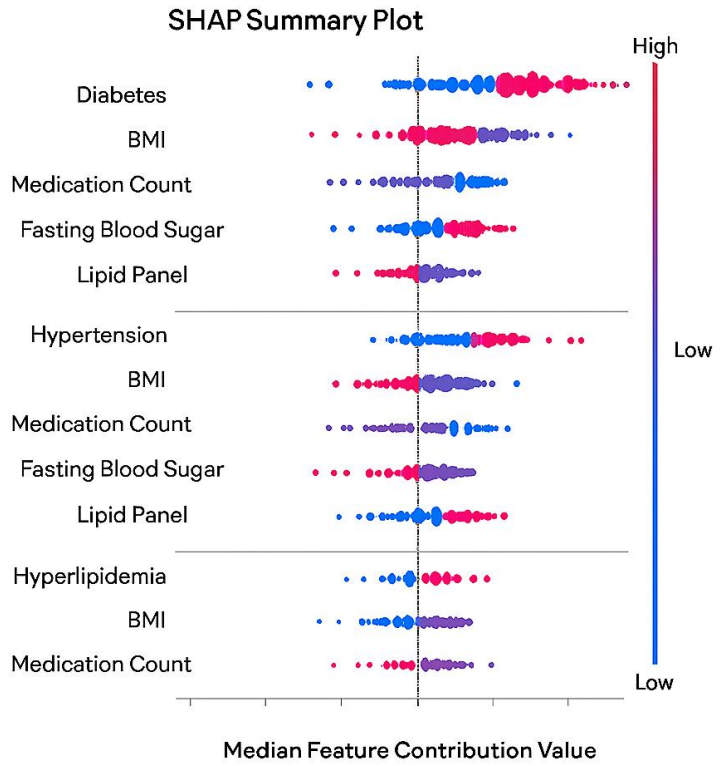

Figure 4 presents a SHAP summary plot by disease category, displaying median contribution values across all features and their direction of influence. The visual highlights demonstrate how different disease pathways are governed by distinct pediatric signals, despite some shared risk markers like BMI and medication complexity [32].

To contextualize these patterns, we analyzed individual case trajectories from high-confidence predictions. One such example involved a female patient flagged for future diabetes at age 13. Her risk score spiked sharply due to a combination of elevated fasting glucose, increased BMI velocity, and multiple endocrinology consults. The model successfully predicted diabetes onset at age 20, six years before her first clinical diagnosis.

Another case involved a male patient flagged for hypertension at age 11, despite normal BMI. His risk profile was driven by family history and episodic elevated blood pressure readings during acute visits. These nuanced insights highlight the model's ability to leverage temporal irregularities and integrate seemingly disparate signals [33].

These case studies reinforce the clinical relevance of the ML pipeline, offering not only population-level predictions but also actionable individual-level insights. By surfacing risk spikes years in advance, the system enables early intervention strategies tailored to each patient's trajectory. This individualized foresight, supported by robust interpretability, can transform pediatric chronic disease prevention from a reactive paradigm to a proactive, data-informed approach [34].

6. CLINICAL AND POLICY IMPLICATIONS

6.1 Use Cases in Pediatric and Family Practice

Machine learning (ML)-derived risk scores offer significant value in pediatric and family practice settings, particularly through the delivery of personalized anticipatory guidance. Unlike conventional growth charts or risk calculators, ML models can dynamically incorporate historical patterns, comorbidities, medication exposures, and social context to

generate individualized predictions for future disease onset [25]. This enables pediatricians to initiate more tailored counseling sessions, emphasizing modifiable risk factors unique to each patient.

For example, a child flagged as high risk for hypertension based on early blood pressure variability and polypharmacy patterns could receive early dietary and lifestyle interventions before clinical thresholds are crossed. These predictive alerts empower clinicians to communicate clearly with families about specific vulnerabilities, thus improving adherence to preventive care recommendations [26].

Care plan adjustments during adolescence when healthcare transitions often occur are another key application. Adolescents are frequently lost to follow-up during this phase, leading to missed opportunities for chronic disease prevention. Embedding ML risk insights into transition-of-care protocols allows providers to triage which patients require more intensive follow-up, behavioral health support, or referrals to specialists [27]. Additionally, family physicians managing siblings or parents within multigenerational households can use risk feedback to address household-level patterns such as shared diet or access barriers.

These applications bridge pediatric and family medicine by fostering a longitudinal and contextualized approach to risk management. Figure 5 presents a policy model that situates ML risk score delivery within routine pediatric screening workflows, including integration points for clinical decision support, alerts, and care escalation paths. When implemented alongside traditional assessments, these tools enable a proactive, rather than reactive, response to the growing burden of chronic disease in youth [28].

6.2 Ethical and Operational Challenges

Despite its promise, deploying predictive ML tools in pediatric settings introduces several ethical and operational challenges. One major concern is the issue of consent, particularly when predictions involve long-horizon risk forecasts. While initial data collection may occur with guardian consent, predictive modeling can generate future health insights that extend beyond the child's current clinical context or even reach into adulthood [29]. Questions arise regarding whether re-consent is necessary as adolescents mature or when risk scores are shared during transitions to adult care.

Another concern is the long-term storage and use of pediatric data for ongoing model refinement. Although de-identified datasets offer some protection, maintaining such records over years necessitates strict governance and auditability to prevent misuse. The potential stigmatization associated with being labeled "high risk" during childhood may also have psychological or social consequences if not carefully communicated and contextualized by providers [30].

Operationally, EHR interoperability remains a barrier. ML models depend on comprehensive, temporally rich data, which is often fragmented across institutions or lost during transitions between pediatric and adult care providers. Without standardized EHR formats and persistent patient identifiers, continuity in risk monitoring becomes compromised. This challenge is especially pronounced in underserved populations who are more likely to experience fragmented care [31].

Alert fatigue is another serious concern. Introducing ML-generated risk flags into already overloaded clinical interfaces may lead to desensitization among providers. If not prioritized and filtered appropriately, such alerts may be ignored or mistrusted, reducing their impact. Implementing tiered alert systems where low-risk predictions are silently logged and only high-certainty risks prompt immediate action can help mitigate this issue [32].

Data bias also presents ethical tension. If training data underrepresents certain subgroups such as non-English speakers or uninsured children the resulting predictions may reflect or amplify existing disparities. Fairness audits and subgroup performance evaluations must be embedded in development and deployment processes.

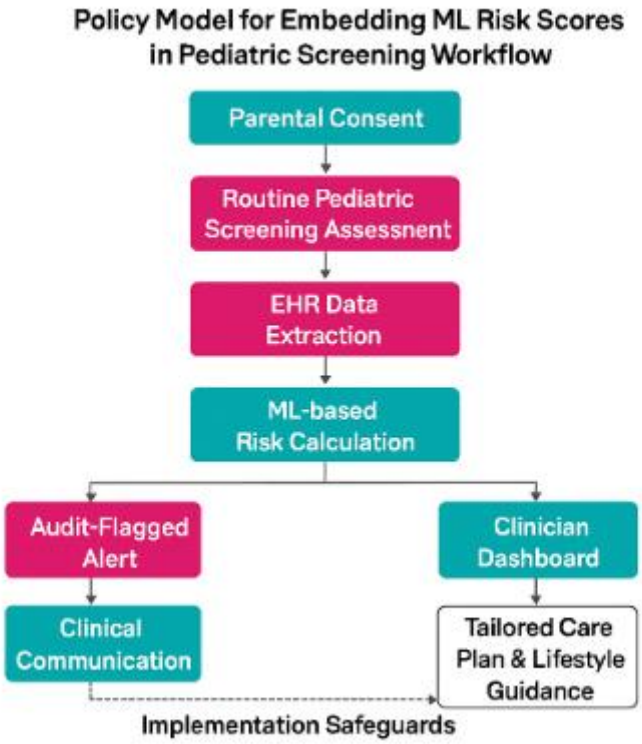

Figure 5. *Policy model for integrating machine learning risk scores into routine pediatric screening workflows. The diagram highlights key components such as clinical decision support, alert generation, care escalation paths, and safeguards including parental communication protocols, audit flags, and fairness controls enabling proactive and trustworthy chronic disease prevention in youth.*

In summary, the deployment of predictive ML tools in pediatric care requires a delicate balance between innovation and ethical responsibility. Figure 5 embeds safeguards into the policy model, including parental communication workflows, audit flags, and escalation filters designed to maintain both clinical utility and patient trust [33].

6.3 Impacts on Preventive Public Health Strategy

The integration of ML-driven pediatric risk scores into public health infrastructure offers transformative potential for preventive strategy design. National screening programs, often focused on reactive case detection, can be restructured into tiered surveillance systems that stratify children by predicted future risk. ML tools can provide preemptive identification of at-risk individuals well before symptoms manifest, enabling earlier enrollment in targeted interventions [34].

Such integration would be particularly impactful in school-based health programs, where annual assessments are conducted at scale. By linking these assessments with predictive EHR models, health agencies can identify subpopulations requiring nutritional support, psychological screening, or chronic disease monitoring before diagnoses become entrenched. In regions with limited pediatric specialist availability, this triaging capacity can improve resource allocation by directing care toward those at highest risk [35].

Public health agencies can also leverage these tools for campaign planning. For instance, ML risk estimates could guide localized health education or behavioral interventions focused on obesity or early diabetes prevention, especially in high-risk zip codes. This spatially informed targeting enhances the efficiency of community outreach and minimizes intervention fatigue in low-risk populations.

Furthermore, integrating predictive scores into child welfare and maternal-child health registries creates opportunities for generational prevention. If a parent's EHR flags metabolic risks, linked pediatric records could trigger early lifestyle counseling for their children. This cross-generational approach is especially relevant in high-deprivation areas where shared environmental exposures heighten familial risk patterns [36].

To operationalize these strategies, policy mandates would need to support infrastructure for secure EHR linkage, real-time model deployment, and workforce training in ML tool interpretation. Figure 5 maps how predictive insights can be embedded in pediatric screening workflows at the policy level, including connections to public health databases and school health records. The model also delineates governance structures, with oversight roles for ethics boards and algorithm fairness committees [37].

In the long term, integrating pediatric ML tools into preventive frameworks could flatten the trajectory of chronic disease across populations. By embedding predictive intelligence upstream in the care pathway, public health systems can shift from reactive care to anticipatory action unlocking a new era in pediatric health equity and chronic disease prevention [38].

7. FUTURE RESEARCH AND TECHNOLOGICAL ADVANCEMENTS

7.1 Expansion to Multi-Omics and Imaging Data

While pediatric EHRs provide a rich temporal view of clinical development, their predictive capacity can be significantly enhanced through the integration of multi-omics and imaging data. Multi-omics includes genomics, transcriptomics, proteomics, and metabolomics each contributing unique biological insight into disease predisposition. For instance, incorporating polygenic risk scores into ML models enables the stratification of children with latent genetic predispositions for obesity, type 2 diabetes, or cardiovascular disease, even before phenotypic signs emerge [31].

Linking genomic data with EHRs in pediatrics has already shown promise in rare disease diagnosis and pharmacogenomics. When extended to chronic disease risk modeling, it provides a foundation for precision prevention strategies that tailor lifestyle interventions based on a child's inherited risk profile. Imaging data, particularly from pediatric radiology and retinal screening, can further enrich models with structural and vascular biomarkers indicative of systemic disease risks [32].

However, integrating these high-dimensional datasets with EHRs presents technical and ethical challenges. Data harmonization across modalities, computational demands, and privacy-preserving data storage protocols must be addressed. Moreover, omics-informed predictions must be interpretable to clinicians and caregivers, especially when the child is asymptomatic. Despite these hurdles, the convergence of clinical and molecular data represents a next frontier in pediatric AI [33].

By embedding omics and imaging within longitudinal ML models, researchers can build more accurate and personalized risk predictions that reflect both intrinsic and acquired health determinants. This multi-modal fusion holds particular promise in early intervention for complex conditions with both genetic and environmental etiologies [34].

7.2 Generalizability Across Populations and Geographies

A critical challenge in pediatric AI model development is ensuring generalizability across diverse populations and healthcare environments. Many predictive models are trained on urban hospital datasets with relatively dense and well-structured EHRs, which do not reflect care patterns in rural, low-resource, or international settings. These disparities result in reduced model performance and poor calibration when deployed outside their development context, limiting their utility in global pediatric health systems [35].

Rural clinics often face episodic care, fragmented EHRs, and limited diagnostic capacity, all of which impair the temporal resolution and feature richness required by advanced ML models. Furthermore, sociocultural determinants such as traditional medicine use, household food insecurity, and caregiver literacy are rarely captured in structured EHRs but significantly impact chronic disease trajectories [36].

To address this, federated learning models offer a promising solution. Instead of centralizing data, federated frameworks allow multiple institutions across geographies and resource levels to collaboratively train shared models without transferring patient-level data. Each site contributes updates to a central model, preserving privacy while capturing local variation. This approach supports more equitable algorithm development and improves robustness across clinical settings [37].

In parallel, adapting input features to reflect local health practices and integrating community-level indicators can enhance contextual sensitivity. Calibration layers and bias detection modules should be included to assess subgroup fairness across geography, language, and insurance type.

Generalizable pediatric AI systems must be built with diversity by design. This includes active engagement with global clinical partners, ethical governance, and continuous validation across stratified populations to ensure universal relevance and fairness [38].

7.3 Toward Explainable Pediatric-AI Systems

For ML tools to be adopted in pediatric care, they must not only perform accurately but also support transparent decision-making. Explainability is crucial in settings involving children, where decisions often involve multidisciplinary teams, guardians, and public institutions. Pediatric-AI systems must therefore go beyond numerical risk outputs and offer intuitive visualizations and justification mechanisms that align with clinical reasoning [39].

Visual risk dashboards represent one solution. These interfaces display individual patients' predicted risks, highlight contributing factors, and track changes over time. For example, a child's dashboard may show increasing diabetes risk, linked to rising BMI z-score, reduced physical activity, and family history, along with projected future trajectories under various intervention scenarios. Such tools enhance understanding, guide shared decision-making, and reduce anxiety for caregivers by contextualizing predictions within actionable narratives [40].

Attention maps and SHAP explanations should also be embedded in clinician-facing tools to indicate which time windows and variables most influenced each prediction. Integrating these explanations into existing EHR platforms or clinical decision support systems allows seamless workflow alignment.

To ensure uptake, clinician engagement is essential during design. Co-development workshops, usability testing, and continuous feedback loops should inform interface design, risk communication strategies, and model thresholds. Clinicians must trust and understand the system before relying on it to guide patient care [41].

Ultimately, explainable pediatric-AI tools support ethical deployment, encourage clinician confidence, and foster family trust key pillars for sustainable integration of ML in preventive child health strategies.

8. CONCLUSION

8.1 Summary of Contributions and Findings

This study presents a novel machine learning (ML) pipeline designed to predict adult-onset chronic diseases namely obesity, type 2 diabetes, and hypertension using longitudinal pediatric electronic health records (EHRs). Unlike prior approaches focused primarily on adult populations or static risk factors, this pipeline leverages the temporal depth of

pediatric records, incorporating dynamic variables such as growth velocity, laboratory markers, medication history, and behavioral diagnoses across developmental stages.

The framework integrates advanced model architectures, including transformers and gradient boosting, to capture complex temporal interactions and feature dependencies. These models were rigorously validated across multiple healthcare institutions and demonstrated strong predictive performance, particularly in identifying at-risk individuals years before clinical diagnosis. For example, obesity and diabetes risk could be flagged with high precision up to four to six years in advance, allowing for meaningful preemptive interventions during adolescence.

Importantly, the models maintained interpretability through SHAP and attention-based mechanisms, and fairness audits revealed generally equitable performance across gender and socioeconomic strata. The inclusion of policy models and real-world clinical dashboards illustrated potential integration paths into pediatric workflows, from primary care clinics to public health campaigns.

By extending pediatric care beyond episodic treatment toward long-term, risk-informed prevention, this study contributes a foundational tool for transforming chronic disease trajectories. It establishes a proof of concept for predictive pediatric AI that is clinically relevant, ethically grounded, and operationally feasible, paving the way for future integration of omics, imaging, and household-level context data into comprehensive early risk detection systems.

8.2 Final Reflections on Translational Pathways

Bringing predictive algorithms from research to routine pediatric care involves more than technical validation it demands attention to translational processes, institutional readiness, and user trust. For machine learning models to become embedded in frontline practice, they must integrate seamlessly into clinician workflows, align with existing decision-making structures, and provide clear, actionable outputs that enhance not complicate care delivery.

The models developed in this study are positioned to support a paradigm shift in childhood chronic disease prevention. By surfacing individual risk insights years before symptoms emerge, they enable a transition from reactive to anticipatory care. However, successful deployment requires engagement from clinicians, families, and health administrators at every stage. Clinical decision support systems must be adapted to reflect pediatric-specific needs, including considerations for developmental stages, parental involvement, and transition-of-care protocols during adolescence.

Moreover, equity must remain a central priority. Translational efforts should prioritize deployment in high-need settings, such as community health centers and school-based clinics, where preventive interventions have the greatest potential impact. Training and support for clinicians in interpreting model outputs, as well as clear communication strategies for discussing risk with families, are essential to ensure uptake and sustainability.

Ultimately, the journey from algorithm to bedside is not linear. It involves iterative testing, community feedback, and responsive design. With continued investment in ethical, explainable, and scalable pediatric AI, the vision of using data-driven insights to alter lifelong health trajectories is not only achievable it is imperative for building healthier futures.

REFERENCE

1. Swinckels L, Bennis FC, Ziesemer KA, Scheerman JF, Bijwaard H, de Keijzer A, Bruers JJ. The use of deep learning and machine learning on longitudinal electronic health records for the early detection and prevention of diseases: scoping review. Journal of medical Internet research. 2024 Aug 20;26:e48320.

2. Javidi H, Mariam A, Alkhaled L, Pantalone KM, Rotroff DM. An interpretable predictive deep learning platform for pediatric metabolic diseases. Journal of the American Medical Informatics Association. 2024 Jun 1;31(6):1227-38.

3. Ganatra HA. Machine learning in pediatric healthcare: Current trends, challenges, and future directions. Journal of Clinical Medicine. 2025 Jan 26;14(3):807.

4. Kalejaiye AN, Shallom K, Chukwuani EN. Implementing federated learning with privacy-preserving encryption to secure patient-derived imaging and sequencing data from cyber intrusions. *Int J Sci Res Arch*. 2025;16(01):1126–45. doi: https://doi.org/10.30574/ijsra.2025.16.1.2120.

5. Cascarano A, Mur-Petit J, Hernandez-Gonzalez J, Camacho M, de Toro Eadie N, Gkontra P, Chadeau-Hyam M, Vitria J, Lekadir K. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. Artificial Intelligence Review. 2023 Nov;56(Suppl 2):1711-71.

6. Jamiu OA, Chukwunweike J. DEVELOPING SCALABLE DATA PIPELINES FOR REAL-TIME ANOMALY DETECTION IN INDUSTRIAL IOT SENSOR NETWORKS. International Journal Of Engineering Technology Research & Management (IJETRM). 2023Dec21;07(12):497–513.

7. Datta S, Morassi Sasso A, Kiwit N, Bose S, Nadkarni G, Miotto R, Böttinger EP. Predicting hypertension onset from longitudinal electronic health records with deep learning. JAMIA open. 2022 Oct 4;5(4).

8. Manemann SM, St Sauver JL, Liu H, Larson NB, Moon S, Takahashi PY, Olson JE, Rocca WA, Miller VM, Therneau TM, Ngufor CG. Longitudinal cohorts for harnessing the electronic health record for disease prediction in a US population. BMJ open. 2021 Jun 1;11(6):e044353.

9. Zhao J, Feng Q, Wu P, Lupu RA, Wilke RA, Wells QS, Denny JC, Wei WQ. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. Scientific reports. 2019 Jan 24;9(1):717.

10. Gupta M, Phan TL, Bunnell HT, Beheshti R. Obesity Prediction with EHR Data: A deep learning approach with interpretable elements. ACM Transactions on Computing for Healthcare (HEALTH). 2022 Apr 7;3(3):1-9.

11. Yuan Q, Cai T, Hong C, Du M, Johnson BE, Lanuti M, Cai T, Christiani DC. Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer. JAMA Network Open. 2021 Jul 1;4(7):e2114723-.

12. Ye C, Fu T, Hao S, Zhang Y, Wang O, Jin B, Xia M, Liu M, Zhou X, Wu Q, Guo Y. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. Journal of medical Internet research. 2018 Jan 30;20(1):e22.

13. Andrew Nii Anang and Chukwunweike JN, Leveraging Topological Data Analysis and AI for Advanced Manufacturing: Integrating Machine Learning and Automation for Predictive Maintenance and Process Optimization (2024) https://dx.doi.org/10.7753/IJCATR1309.1003

14. Hammond R, Athanasiadou R, Curado S, Aphinyanaphongs Y, Abrams C, Messito MJ, Gross R, Katzow M, Jay M, Razavian N, Elbel B. Predicting childhood obesity using electronic health records and publicly available data. PloS one. 2019 Apr 22;14(4):e0215571.

15. Adefolaju IT, Ogundele BD, Unanah OV. Measuring what matters: a metrics-driven approach to evaluating access and distribution programs in LMIC. *Int J Eng Technol Res Manag*. 2022 Feb;6(2):254. Available from: https://doi.org/10.5281/zenodo.15954135

16. Cohen NM, Lifshitz A, Jaschek R, Rinott E, Balicer R, Shlush LI, Barbash GI, Tanay A. Longitudinal machine learning uncouples healthy aging factors from chronic disease risks. Nature Aging. 2024 Jan;4(1):129-44.

17. Chukwunweike JN, Mba JU, Kadiri C. Enhancing maritime security through emerging technologies: the role of machine learning in cyber threat detection and mitigation., USA. 2024 Aug. DOI: https://doi.org/10.55248/gengpi.5.0824.2401

18. Oluwagbade E. Bridging the healthcare gap: the role of AI-driven telemedicine in emerging economies. *Int J Res Publ Rev* [Internet]. 2025 Jan ;6(1):3732–43. Available from: https://doi.org/10.55248/gengpi.6.0125.0531.

19. Afrifa-Yamoah, E., Adua, E., Peprah-Yamoah, E., Anto, E.O., Opoku-Yamoah, V., Acheampong, E., Macartney, M.J. and Hashmi, R., 2025. Pathways to chronic disease detection and prediction: Mapping the potential of machine learning to the pathophysiological processes while navigating ethical challenges. *Chronic Diseases and Translational Medicine*, *11*(01), pp.1-21.

20. Darkwah E. PFAS contamination in drinking water systems near industrial zones: Bioaccumulation, human exposure risks, and treatment technology challenges. *Int J Sci Res Arch.* 2021;3(2):284–303. Available from: https://doi.org/10.30574/ijsra.2021.3.2.0099

21. Chen S, Yu J, Chamouni S, Wang Y, Li Y. Integrating machine learning and artificial intelligence in life-course epidemiology: pathways to innovative public health solutions. BMC medicine. 2024 Sep 2;22(1):354.

22. Colmenarejo G. Machine learning models to predict childhood and adolescent obesity: a review. Nutrients. 2020 Aug 16;12(8):2466.

23. Kalejaiye AN. Causal modeling of insider threat behavior using probabilistic graphical networks to strengthen organizational cyber-resilience and trust architectures. *Int J Res Publ Rev*. 2025;6(07). Available from: https://ijrpr.com/uploads/V6ISSUE7/IJRPR50319.pdf.

24. Emmanuel Oluwagbade and Oluwole Raphael Odumbo. Building resilient healthcare distribution networks: Adapting to crises, securing supplies and improving scalability. International Journal of Science and Research Archive, 2025, 14(01), 1579-1598. DOI: https://doi.org/10.30574/ijsra.2025.14.1.0265.

25. Javidi H. Deep Neural Networks for Complex Disease Prediction Using Electronic Health Records and Genomic Data. Cleveland State University; 2024.

26. Adefolaju IT, Egba O, Unanah OV, Adetula AA. Designing inclusive access and distribution models: Global best practices for reaching underserved populations. *Int J Comput Appl Technol Res.* 2024;13(11):73–87. doi:10.7753/IJCATR1311.1011.

27. Javidi H, Mariam A, Khademi G, Zabor EC, Zhao R, Radivoyevitch T, Rotroff DM. Identification of robust deep neural network models of longitudinal clinical measurements. NPJ Digital Medicine. 2022 Jul 27;5(1):106.

28. Chukwunweike J, Lawal OA, Arogundade JB, Alade B. Navigating ethical challenges of explainable AI in autonomous systems. *International Journal of Science and Research Archive*. 2024;13(1):1807–19. doi:10.30574/ijsra.2024.13.1.1872. Available from: https://doi.org/10.30574/ijsra.2024.13.1.1872.

29. Barak-Corren Y, Tsurel D, Keidar D, Gofer I, Shahaf D, Leventer-Roberts M, Barda N, Reis BY. The value of parental medical records for the prediction of diabetes and cardiovascular disease: a novel method for generating and incorporating family histories. Journal of the American Medical Informatics Association. 2023 Nov 17;30(12):1915-24.

30. Kalejaiye AN. Federated learning in cybersecurity: privacy-preserving collaborative models for threat intelligence across geopolitically sensitive organizational boundaries. *Int J Adv Res Publ Rev*. 2025;2(07):227–50. Available from: https://ijarpr.com/uploads/V2ISSUE7/IJARPR0712.pdf?v=2.

31. Cheng ER, Steinhardt R, Ben Miled Z. Predicting childhood obesity using machine learning: practical considerations. BioMedInformatics. 2022 Mar 8;2(1):184-203.

32. Singh A, Nadkarni G, Gottesman O, Ellis SB, Bottinger EP, Guttag JV. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. Journal of biomedical informatics. 2015 Feb 1;53:220-8.

33. Emmanuel Oluwagbade, Alemede Vincent, Odumbo Oluwole, Animashaun Blessing. LIFECYCLE GOVERNANCE FOR EXPLAINABLE AI IN PHARMACEUTICAL SUPPLY CHAINS: A FRAMEWORK FOR CONTINUOUS VALIDATION, BIAS AUDITING, AND EQUITABLE HEALTHCARE DELIVERY. International Journal of Engineering Technology Research & Management (IJETRM). 2023Nov21;07(11).

34. Tomašev N, Harris N, Baur S, Mottram A, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, Magliulo V, Meyer C. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. Nature protocols. 2021 Jun;16(6):2765-87.

35. Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. Journal of the American Medical Informatics Association. 2015 Jul 1;22(4):872-80.

36. Black JE, Kueper JK, Terry AL, Lizotte DJ. Development of a prognostic prediction model to estimate the risk of multiple chronic diseases: constructing a copula-based model using Canadian primary care electronic medical record data. International journal of population data science. 2021 Jan 19;6(1):1395.

37. Mbanugo OJ. Diabetes management: navigating the complexities of myths, cultural beliefs and religion. *Int J Eng Technol Res Manag*. 2018 ;2(1):[about 5 p.]. Available from: https://doi.org/10.5281/zenodo.15869683.

38. Nenova Z, Shang J. Chronic disease progression prediction: Leveraging case-based reasoning and big data analytics. Production and Operations Management. 2022 Jan;31(1):259-80.

39. de Lacy N, Ramshaw M, Lam WY. RiskPath: Explainable deep learning for multistep biomedical prediction in longitudinal data. Patterns. 2025 Apr 28.

40. Li C, Mowery DL, Ma X, Yang R, Vurgun U, Hwang S, Donnelly HK, Bandhey H, Senathirajah Y, Visweswaran S, Sadhu EM. Realizing the potential of social determinants data in EHR systems: A scoping review of approaches for screening, linkage, extraction, analysis, and interventions. Journal of Clinical and Translational Science. 2024 Jan;8(1):e147.

41. Wang B, Sheu YH, Lee H, Mealer RG, Castro VM, Smoller JW. Machine Learning Models for the Prediction of Early-Onset Bipolar Using Electronic Health Records. MedRxiv. 2024 Feb 21:2024-02.