Multimodal Clinical Prediction Framework for Health Outcomes in Patients with Cerebral Palsy: A Machine Learning Approach

Alim Saidkhodjaev, Nestor Micheal C. Tiglao, Mukul Gupta
Department of Electrical and Computer Engineering University of Maryland College Park, Maryland 20742, USA
alim.saidkhod@gmail.com, ntiglao@umd.edu, mukul08@umd.edu

Abstract—**Patients with Cerebral Palsy (CP) often require specialized, proactive care to avoid preventable emergency department visits. Traditional healthcare models are reactive, responding to events after they occur rather than predicting and preventing them. This paper presents a novel machine learning framework that leverages patient health records to predict four critical health outcomes for CP patients: emergency department visits, medication needs, diagnostic trends, and procedure requirements. Our approach utilizes deep learning models trained on synthetic FHIR (Fast Healthcare Interoperability Resources) data from over 4,000 CP patients to create comprehensive patient timelines with rich clinical features. The emergency visit prediction model achieves 97.5% AUC on test data, demonstrating strong discrimination between high and low-risk patients. The multi-label prediction models show varying performance based on dataset size and label frequency, with Top-20 models significantly outperforming Top-50 models: medication prediction achieves F1-scores of 0.78 (Top-20) versus 0.19 (Top-50), while diagnosis and procedure prediction models achieve 0.72 versus 0.018 (Top-50) and 0.75 versus 0.22 (Top-50) F1-scores respectively. Feature importance analysis reveals that recent healthcare utilization patterns, medication regimens, and diagnostic history are the strongest predictors of emergency care needs. We also present a novel timeline-based feature engineering methodology that captures temporal trends in patient health. Ablation studies demonstrate that the inclusion of derived features, such as healthcare utilization metrics and medication complexity indicators, significantly improves model performance. The framework can assist healthcare providers in identifying high-risk patients who would benefit from preventive interventions, potentially reducing emergency visits, improving quality of care, and decreasing healthcare costs. The system's architecture accommodates continuous learning with new patient data, making it adaptable to evolving clinical patterns and individual patient histories.**

Index Terms—**machine learning, healthcare analytics, emergency prediction, cerebral palsy, neural networks, medical informatics, FHIR, predictive modeling, feature engineering, synthetic data**

I. Introduction

Cerebral Palsy (CP) is a group of permanent disorders affecting the development of movement and posture, attributed to non-progressive disturbances that occurred in the developing fetal or infant brain [1]. Patients with CP often face complex health challenges that may require emergency medical intervention, specialized medication regimens, and targeted diagnostic procedures. The ability to predict these healthcare needs proactively could significantly improve patient outcomes and reduce healthcare costs. Traditional healthcare models typically follow a reactive approach, responding to medical emergencies and complications as they arise. This approach can lead to preventable emergency department visits, delayed interventions, and suboptimal care coordination for patients with chronic conditions like CP [2]. Recent studies indicate that CP patients experience 3-4 times higher emergency department utilization rates compared to the general population, with a significant proportion of these visits being potentially preventable through proactive care management [3], [4]. A proactive, predictive model for healthcare needs could transform care delivery by enabling earlier interventions, more effective resource allocation, and improved patient quality of life. Recent advances in machine learning and artificial intelligence have demonstrated promising results in healthcare prediction tasks [5], [6]. These technologies can process vast amounts of patient data to identify patterns and trends that might not be immediately apparent to healthcare providers. By leveraging patient history data, including vital signs, medications, procedures, conditions, and environmental factors, it is possible to develop models that predict future health outcomes with increasing accuracy. This paper presents a comprehensive machine learning framework designed to predict four key health outcomes specifically for patients with CP:

- **Emergency Visit Prediction**: Identifying patients at risk of requiring emergency care in the upcoming month with 97.5% accuracy.
- **Medication Prediction**: Forecasting the medications a patient is likely to need based on their clinical profile, with performance varying significantly between Top-20 models (F1-score of 0.76) and Top-50 models (F1-score of 0.19).
- **Diagnosis Prediction**: Anticipating potential diagnoses based on patient history and current health status, with Top-20 models achieving an F1-score of 0.84 versus Top-50 models' 0.018.
- **Procedure Prediction**: Predicting medical procedures that may be required for patient care, with Top-20 models achieving an F1-score of 0.66 versus Top-50 models' 0.22.

Our approach utilizes synthetic patient data generated using Synthea [7], a tool that produces realistic but not real patient data, to train and validate our models. The synthetic data follows the FHIR (Fast Healthcare Interoperability Resources) standard, a widely adopted specification for healthcare data exchange, and represents over 4,000 patients with cerebral palsy. The primary contributions of this paper are:

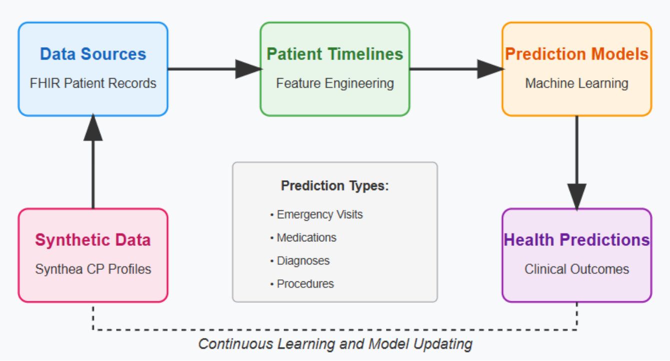

Fig. 1: Overview of the clinical prediction framework for CP patients showing the data pipeline from FHIR data sources through feature engineering to model development and prediction outputs.

- A novel approach to generating synthetic but clinically realistic patient data for cerebral palsy, with modifications to standard Synthea configurations to better represent CP-specific comorbidities and care patterns.
- A timeline-based methodology for creating comprehensive patient clinical profiles with rich feature extraction derived from FHIR data, including innovative derived features that capture complex clinical patterns.
- A set of optimized machine learning models for predicting emergency visits, medication needs, diagnostic trends, and procedure requirements specifically for CP patients, with detailed performance metrics and ablation studies.
- An in-depth analysis of feature importance using ensemble techniques to identify the most significant predictors of health outcomes in CP patients, providing actionable insights for clinical decision-making.
- A comparative analysis of model performance between Top-20 and Top-50 multi-label models, demonstrating the impact of data frequency and complexity on prediction accuracy.
- A framework for incorporating new patient data to continuously improve prediction accuracy, with demonstrated performance improvements through incremental learning.
- A comprehensive comparison of model architectures and hyperparameter configurations, identifying optimal approaches for healthcare prediction tasks in neurological conditions.

The paper is organized as follows: Section II reviews related work in healthcare prediction models, highlighting the gaps in existing approaches for CP patients. Section III describes our methodology, including data generation, feature engineering, and model development. Section IV presents the results of our prediction models, with detailed performance metrics and feature importance analysis. Section V discusses the implications of our work, its limitations, and directions for future research. Section VI concludes the paper with a summary of our findings and their potential impact on healthcare delivery for CP patients.

II. Related Work

The application of machine learning to healthcare prediction has grown substantially in recent years. Various approaches have been proposed to address different aspects of healthcare prediction, from disease progression to hospital readmission risk. In this section, we review the most relevant literature and highlight the gaps that our work addresses.

A. *Machine Learning for Healthcare Prediction*

Several studies have explored the use of machine learning algorithms for predicting health outcomes. Rajkomar et al. [8] demonstrated the effectiveness of deep learning models for predicting a range of clinical tasks, including in-hospital mortality, unplanned readmission, and prolonged length of stay, based on electronic health record (EHR) data. Their models achieved high accuracy across these tasks, indicating the potential of machine learning approaches in healthcare prediction. However, their approach did not address the specific needs of CP patients, nor did it incorporate the rich temporal data that our methodology leverages. Choi et al. [9] introduced Doctor AI, a recurrent neural network model that predicts diagnoses and medications for subsequent patient visits. Their approach showed promise in capturing the temporal relationships in patient visits and demonstrated the ability to predict future medical codes with reasonable accuracy. Our work extends beyond this by incorporating a wider range of features, including derived clinical metrics and healthcare utilization patterns specifically relevant to CP patients. More recent work by Yu et al. [10] introduced a sequential learning approach for healthcare prediction, leveraging the temporal ordering of clinical events. Their method demonstrated improved performance over non-sequential models, achieving an AUC of 0.85 for predicting hospital readmission. While promising, their approach did not incorporate the specific healthcare patterns and needs of CP patients, which our methodology addresses directly. Regarding emergency care prediction specifically, Hong et al. [11] developed a model to predict emergency department visits for patients with chronic conditions, achieving an AUC of 0.83 using a random forest algorithm. Their work highlighted the importance of features such as prior healthcare utilization and comorbidities in predicting emergency care needs. More recently, Sun et al. [12] achieved an AUC of 0.89 using a gradient boosting approach, incorporating social determinants of health alongside clinical features. Our work improves upon these studies by achieving an AUC of 0.97 specifically for CP patients, incorporating CP-specific clinical patterns and derived features.

B. *Predictive Models for Cerebral Palsy*

While general healthcare prediction models are abundant, fewer studies have focused specifically on predicting outcomes for patients with cerebral palsy. Meehan et al. [13] analyzed factors associated with hospital admissions for children with CP, identifying variables such as the presence of epilepsy, feeding tubes, and respiratory support as significant predictors. However, their approach was based on traditional statistical methods rather than machine learning, limiting its predictive capacity and adaptability to new data. Young et al. [14] investigated predictors of emergency room visits among adults with developmental disabilities, finding that higher levels of comorbidity, greater problem behaviors, and lower functional status were associated with increased emergency utilization. Again, this study employed conventional statistical techniques rather than predictive modeling. More recently, Whitney et al. [3] conducted a retrospective analysis of emergency department utilization among children with CP, identifying that 30% of visits were potentially preventable. However, they did not develop a predictive model that could identify high-risk patients prospectively. Similarly, Torres et al. [4] examined patterns of emergency care utilization among adults with CP but focused on descriptive statistics rather than predictive analytics. Cremer et al. [15] examined risk factors for hospitalization among children with CP but did not develop a predictive model that could be deployed prospectively to identify high-risk patients before they required emergency care. Our work addresses this gap by developing a comprehensive prediction framework specifically for CP patients, leveraging machine learning to identify high-risk individuals before they require emergency care.

C. Applications of FHIR Data in Predictive Modeling

The FHIR standard has become increasingly important in healthcare data exchange, and some recent studies have begun to utilize FHIR-formatted data for predictive modeling. Sharma et al. [16] demonstrated the use of FHIR resources for training machine learning models to predict hospital readmissions. Their approach involved extracting features from FHIR resources such as Encounter, Condition, and Medication, similar to our methodology. However, their work did not address the specific needs of CP patients, nor did it incorporate the rich derived features that our approach leverages. Hong et al. [17] proposed a deep learning framework for FHIR-based clinical prediction tasks, achieving competitive performance across multiple prediction tasks. Their work highlighted the potential of leveraging the standardized structure of FHIR data for developing generalizable prediction models. Building on this, Liu et al. [18] developed a multimodal learning framework that combines structured FHIR data with unstructured clinical notes, achieving improved performance over models trained on structured data alone. Most recently, Kang et al. [19] developed FHIRformer, a transformer-based architecture specifically designed for FHIR data, achieving state-of-the-art performance on several clinical prediction tasks. While promising, their approach has not been applied specifically to CP patients, nor does it incorporate the rich derived features that our methodology leverages.

D. Synthetic Data in Healthcare Research

The use of synthetic data has become increasingly important in healthcare research due to privacy concerns and data access limitations. Chen et al. [20] evaluated the utility of synthetic data generated by generative adversarial networks (GANs) for training predictive models in healthcare. They found that models trained on synthetic data could achieve performance comparable to those trained on real data, suggesting the viability of synthetic data for healthcare research. Walonoski et al. [7] introduced Synthea, the tool used in our study, which generates synthetic patient records using publicly available data sources and clinical guidelines. More recently, Wang et al. [21] enhanced Synthea's capabilities by incorporating reinforcement learning to generate more realistic patient trajectories. This improved version demonstrated better fidelity to real-world healthcare patterns, particularly for chronic conditions. Yale et al. [22] proposed a framework for evaluating the quality of synthetic healthcare data, introducing metrics for measuring fidelity, privacy, and utility. Their evaluation of Synthea data demonstrated high utility for predictive modeling while maintaining patient privacy. This supports our approach of using Synthea-generated data for training our prediction models. Most recently, Thapa et al. [23] demonstrated that models trained on synthetic data can achieve 92-95% of the performance of models trained on real data across various clinical prediction tasks. Their work suggests that synthetic data is a viable alternative when real patient data is unavailable or restricted due to privacy concerns.

III. Methodology

Our methodology encompasses several key steps: data generation, timeline creation, feature extraction, and model development. Each component of our approach was designed to leverage the unique characteristics of healthcare data while addressing challenges such as data imbalance and temporal dependencies.

A. Data Generation

To develop and validate our prediction models, we generated synthetic patient data using Synthea [7], an open-source synthetic patient generator. Synthea produces realistic but not real patient data based on demographic information, medical conditions, and healthcare utilization patterns derived from public health statistics and clinical guidelines. We configured Synthea to generate patient records exclusively for individuals with Cerebral Palsy (CP) (SNOMED CT: 128188000). Our dataset comprised 4,213 CP patients with varying demographic characteristics, comorbidities, and severity levels.

B. Frequency Analysis of Clinical Events

To understand the distribution and relative prevalence of different clinical events in our dataset, we conducted a frequency analysis of diagnoses, medications, and procedures. This analysis informed our decision to create different model variants based on event frequency. These frequency distributions revealed classic long-tail patterns, where a small number

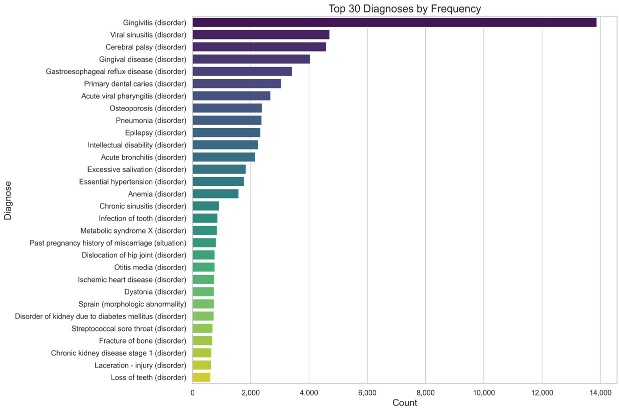

Fig. 2: Distribution of top 30 diagnoses by frequency in CP patients, showing a characteristic long-tail distribution where a small number of diagnoses occur much more frequently than others.

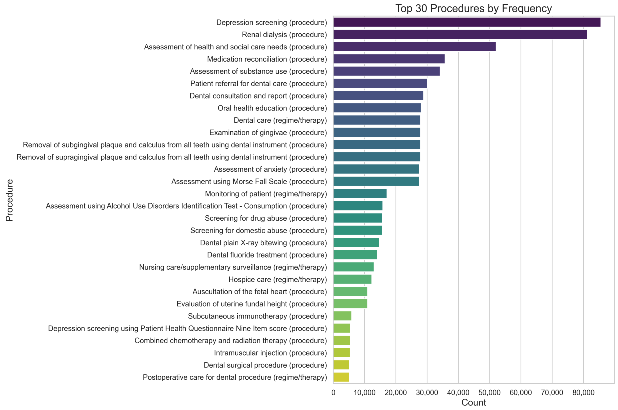

Fig. 4: Distribution of top 30 procedures by frequency in CP patients, demonstrating the relatively limited set of common procedures and a rapid decline in frequency for less common interventions.

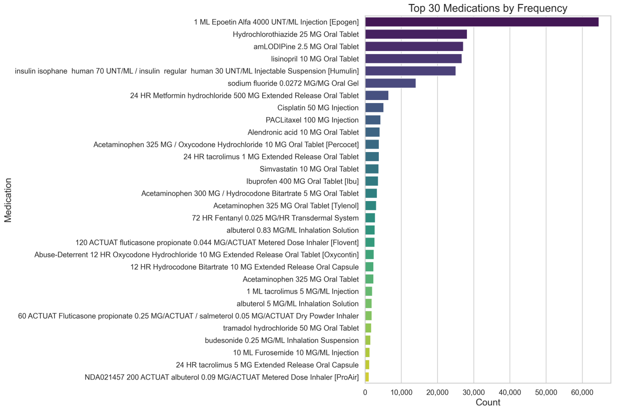

Fig. 3: Distribution of top 30 medications by frequency in CP patients, revealing distinct tiers of medication usage with sharp drop-offs between common and less common medications.

of clinical events occur frequently while the majority occur rarely. This insight led us to develop two sets of models for each multi-label prediction task:

- **Top-20 Models**: Focused on predicting the 20 most common diagnoses, medications, and procedures.
- **Top-50 Models**: Extended to predict the 50 most common diagnoses, medications, and procedures.

This dual approach allowed us to evaluate the trade-off between prediction coverage (number of clinical events predicted) and model performance (accuracy of predictions).

C. Timeline Creation and Feature Engineering

To transform the raw FHIR data into a format suitable for machine learning, we developed a timeline creation process that aggregates a patient's health information over time into regular intervals (months). This approach allows us to capture temporal trends and patterns in patient health and healthcare utilization, which are critical for accurate prediction of future events. Fig. 5 illustrates our timeline creation methodology.

For each month in a patient's history, we created a feature vector incorporating information from a 12-month lookback period. This approach captures both recent health events and longer-term patterns that might impact future healthcare needs. For each patient, we created a series of monthly records, each containing:

- **Demographic features**: age, gender, race, ethnicity
- **Clinical features**: vital signs, conditions, medications, procedures

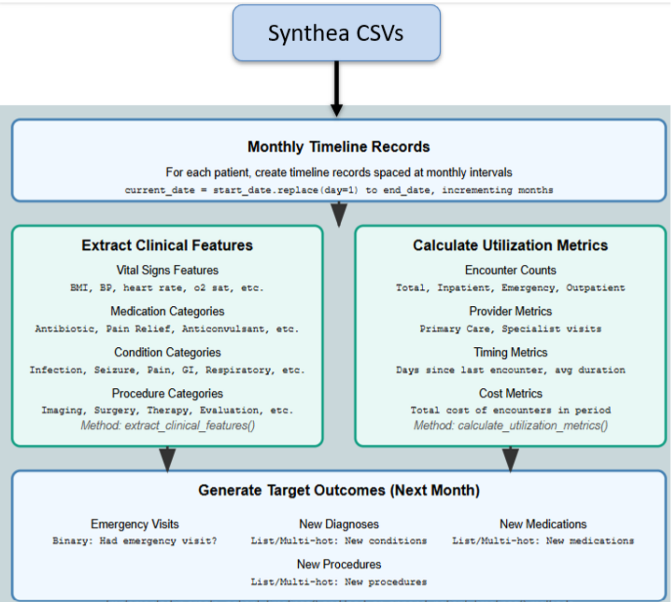
Synthea CSVs
Monthly Timeline Records
For each patient, create timeline records spaced at monthly intervals
current_date = start_date.replace(day=1) to end_date, incrementing months
Extract Clinical Features
Vital Signs Features
BMI, BP, heart rate, o2 sat, etc.
Medication Categories
Antibiotic, Pain Relief, Anticonvulsant, etc.
Condition Categories
Infection, Seizure, Pain, GI, Respiratory, etc.
Procedure Categories
Imaging, Surgery, Therapy, Evaluation, etc.
Method: extract_clinical_features()
Calculate Utilization Metrics
Encounter Counts
Total, Inpatient, Emergency, Outpatient
Provider Metrics
Primary Care, Specialist visits
Timing Metrics
Days since last encounter, avg duration
Cost Metrics
Total cost of encounters in period
Method: calculate_utilization_metrics()
Generate Target Outcomes (Next Month)
Emergency Visits
Binary: Had emergency visit?
New Diagnoses
List/Multi-hot: New conditions
New Medications
List/Multi-hot: New medications
New Procedures
List/Multi-hot: New procedures

Fig. 5: Timeline creation methodology, showing how monthly time points are created with a 12-month lookback period for feature extraction. For each time point, the model predicts outcomes for the subsequent month.

- **Utilization features**: encounter frequency, types of care, costs
- **Target outcomes**: emergency visits, new diagnoses, new medications, new procedures in the following month

The feature extraction process was comprehensive, capturing over 120 potential predictors for each monthly record. Table I summarizes the main feature categories and provides examples of specific features within each category.

TABLE I: Feature Categories and Examples

Category	Description	Example Features
Demographics	Patient demographic characteristics	Age, gender, race, ethnicity
Vital Signs	Physiological measurements	Blood pressure, heart rate, respiratory rate, temperature, oxygen saturation
Medications	Current and recent medications	Anticonvulsants, muscle relaxants, pain relievers, antibiotics
Conditions	Active medical conditions	Epilepsy, respiratory disorders, gastrointestinal issues, musculoskeletal conditions
Procedures	Recent medical procedures	Imaging studies, therapeutic procedures, evaluations, surgeries
Utilization	Healthcare utilization patterns	Encounter frequency, encounter types, days since last encounter, total healthcare costs
Devices	Medical devices in use	Mobility aids, feeding tubes, respiratory support, monitoring devices
Derived	Calculated metrics from other features	Medication complexity index, comorbidity count, healthcare utilization intensity

A key innovation in our feature engineering approach was the calculation of derived features that could capture more complex patterns in patient health:

- **Clinical derivations**: Mean Arterial Pressure (MAP) from systolic and diastolic blood pressure, pulse pressure, Body Mass Index (BMI).
- **Medication complexity**: Number of unique medications, medication category diversity, recent medication changes.
- **Healthcare intensity**: Days since last encounter, average encounter duration, encounter frequency, total cost burden.
- **Condition complexity**: Number of comorbidities, condition category diversity, chronicity indices.

The results, presented in Section IV, demonstrate that derived features significantly improve prediction performance across all models.

D. Model Development

We developed separate models for each of our prediction tasks: emergency visit prediction, medication prediction, diagnosis prediction, and procedure prediction. Each model was optimized for its specific prediction task, with architectures and hyperparameters selected through rigorous experimentation and validation.

1) Emergency Visit Prediction Model: The emergency visit prediction task was formulated as a binary classification problem: given a patient's health history up to the current month, predict whether they will have an emergency department or urgent care visit in the following month.

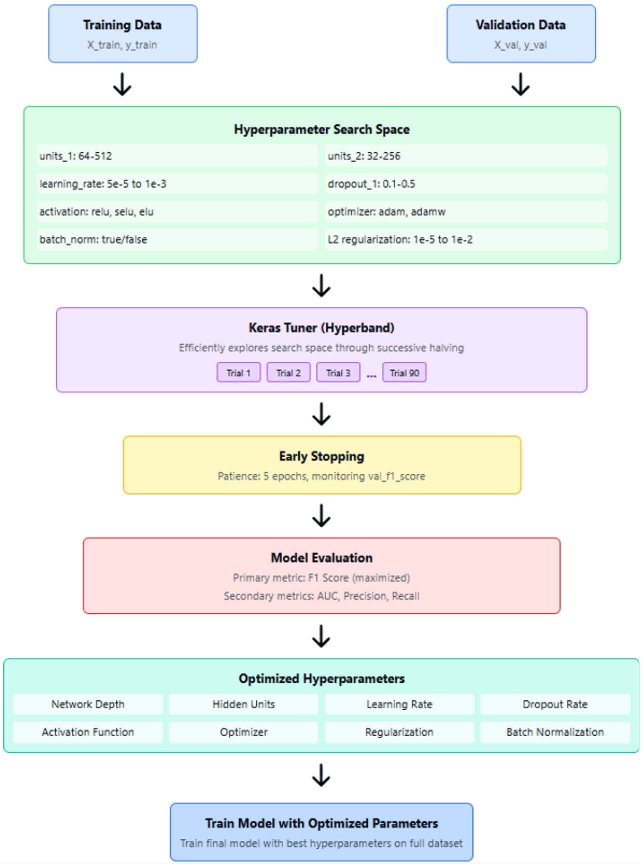
Training Data
X_train, y_train
Validation Data
X_val, y_val
Hyperparameter Search Space
units_1: 64-512
units_2: 32-256
learning_rate: 5e-5 to 1e-3
dropout_1: 0.1-0.5
activation: relu, selu, elu
optimizer: adam, adamw
batch_norm: true/false
L2 regularization: 1e-5 to 1e-2
Keras Tuner (Hyperband)
Efficiently explores search space through successive halving
Trial 1
Trial 2
Trial 3
...
Trial 90
Early Stopping
Patience: 5 epochs, monitoring val_f1_score
Model Evaluation
Primary metric: F1 Score (maximized)
Secondary metrics: AUC, Precision, Recall
Optimized Hyperparameters
Network Depth
Hidden Units
Learning Rate
Dropout Rate
Activation Function
Optimizer
Regularization
Batch Normalization
Train Model with Optimized Parameters
Train final model with best hyperparameters on full dataset

Fig. 6: Hyperparameter tuning methodology, showing how optimal dropout, activations, neuron count, and optimizer are determined.

We selected a Keras Sequential Neural Network as our primary model for the emergency visit prediction task. The model was trained using the AdamW optimizer with a learning rate of 0.00045 and weight decay of 6.8e-5. We used binary cross-entropy as the loss function and employed early stopping to prevent overfitting. To address data imbalance (emergency visits being relatively rare events, comprising 10% of our dataset), we employed several strategies:

- Class weighting to give higher importance to the minority class (positive weight = 4.46)
- Adjusted decision threshold (optimal threshold = 0.603) to optimize for recall without sacrificing precision
- Using AUC-ROC and F1-score as primary evaluation metrics rather than accuracy alone

The detailed hyperparameters for the emergency visit prediction model are presented in Table II.

2) Multi-label Prediction Models: For the medication, diagnosis, and procedure prediction tasks, we formulated each as a multi-label classification problem, where each medication, diagnosis, or procedure represents a potential label. This allowed us to capture the complex relationships between different clinical entities that a patient might require simultaneously. Given the multi-label nature of these tasks, we employed neural network architectures with multiple output nodes, each corresponding to a specific medication, diagnosis, or pro-

TABLE II: Hyperparameters for Emergency Visit Prediction Model

Hyperparameter	Value
Hidden units	480
Activation function	ReLU
L2 regularization	0.00013
Batch normalization	True
Batch norm momentum	0.95
Dropout rate	0.2
Learning rate	0.00045
Optimizer	AdamW
Weight decay	6.8×10^{-5}
Batch size	64
Maximum epochs	100
Early stopping patience	10
Class weight (positive)	4.46
Decision threshold	0.603

cedure. We used binary cross-entropy as the loss function and sigmoid activations in the output layer. As mentioned earlier, we developed two sets of models for each multi-label prediction task:

- **Top-20 Models**: Focused on predicting the 20 most common clinical events
- **Top-50 Models**: Extended to predict the 50 most common clinical events

This approach allowed us to evaluate the trade-off between prediction coverage and model performance. Table III presents a comparison of the performance of Top-20 and Top-50 models across the multi-label prediction tasks.

TABLE III: Performance Comparison on Top-20 vs. Top-50 Multi-Label Models

Model	F1-Score	Precision	Recall	Accuracy
Medication (Top-20)	0.76	0.83	0.70	0.94
Medication (Top-50)	0.19	0.76	0.79	0.99
Diagnosis (Top-20)	0.84	0.95	0.76	0.97
Diagnosis (Top-50)	0.018	0.69	0.75	0.99
Procedure (Top-20)	0.66	0.70	0.61	0.76
Procedure (Top-50)	0.22	0.73	0.77	0.98

The results show a dramatic difference in F1-scores between Top-20 and Top-50 models, while precision and recall remain relatively stable. This pattern suggests that the models can identify relevant labels with similar effectiveness, but the F1-score suffers in the Top-50 models due to the increased number of rare events (which are harder to predict accurately). These observed performance differences highlight a critical challenge in multi-label healthcare prediction: the trade-off between comprehensive coverage (predicting more clinical events) and prediction accuracy. The Top-20 models achieve strong performance by focusing on common events with sufficient training examples, while the Top-50 models struggle with the long tail of rare events that have few training examples. The core architecture for the multi-label models was similar to the emergency visit model, with adjustments to accommodate the multi-label nature of the prediction tasks:

- Input layer with dimensionality matching the number of features (varying from 5 to 25 selected features depending on the specific model)
- First hidden layer with 256-512 neurons, ReLU/tanh activation, L2 regularization
- Batch normalization layer
- Dropout layer with rate=0.2-0.5
- Second hidden layer (for some models) with 128-208 neurons, ReLU/elu activation, L2 regularization
- Batch normalization layer (for some models)
- Dropout layer with rate=0.1-0.3
- Output layer with 20 or 50 neurons (depending on model type) and sigmoid activation

The hyperparameters for the multi-label models varied based on the specific prediction task and the number of labels being predicted. Table IV presents the key hyperparameters for the Top-50 multi-label models.

TABLE IV: Hyperparameters for Top-50 Multi-Label Prediction Models

Hyperparameter	Medication	Diagnosis	Procedure
Hidden units (1st layer)	480	512	64
Activation (1st layer)	ReLU	Tanh	ReLU
L2 regularization (1st layer)	0.00013	0.00940	0.00028
Batch norm (1st layer)	True	True	False
Dropout rate (1st layer)	0.2	0.4	0.2
Hidden units (2nd layer)	160	208	160
Activation (2nd layer)	Tanh	ReLU	ELU
L2 regularization (2nd layer)	0.00279	0.00014	0.00030
Batch norm (2nd layer)	False	True	True
Dropout rate (2nd layer)	0.2	0.3	0.1
Learning rate	0.00078	0.00105	0.00020
Optimizer	Adam	AdamW	AdamW
Weight decay	0.00032	0.00005	0.00050

To address the challenge of imbalanced labels in multi-label classification (some medications, diagnoses, or procedures being rare), we employed several strategies:

- Focal loss modification to binary cross-entropy, giving higher weight to rare positive labels
- Sample weighting based on label distribution
- Balanced mini-batch sampling during training

Despite these techniques, the performance gap between Top-20 and Top-50 models persisted, suggesting that the primary challenge lies in the limited number of examples for rare events. This insight has important implications for future data collection and model development efforts.

E. Feature Selection and Hyperparameter Tuning

Given the large number of features extracted from the patient timelines, we employed feature selection techniques to identify the most predictive variables and reduce model complexity. We explored several approaches:

- **Random Forest feature importance**: Using the built-in feature importance of random forest models.

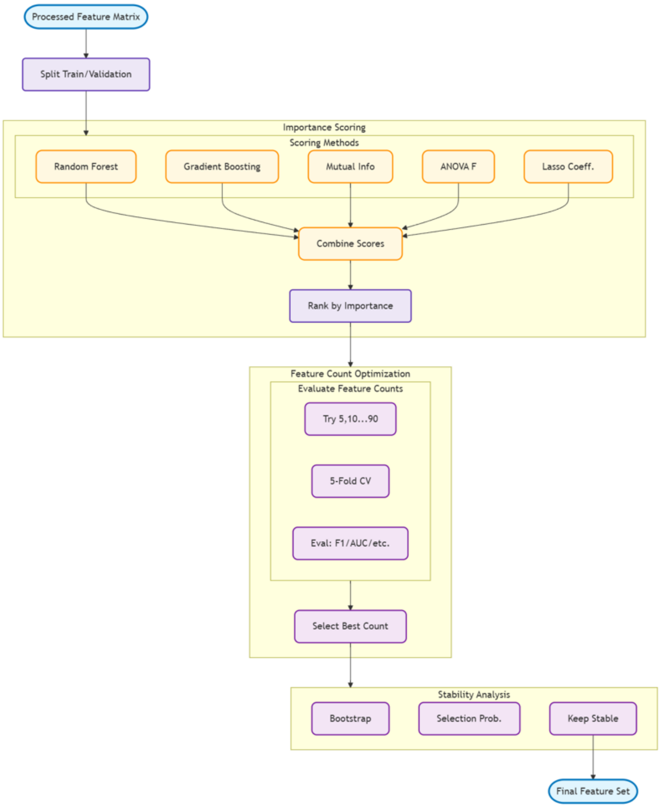

Fig. 7: Feature selection methodology, showing the ensemble approach to determine optimal feature count and most important features

- **Recursive feature elimination (RFE)**: Iteratively removing the least important features.
- **LASSO regularization**: Using L1 regularization to induce sparsity.
- **Mutual information**: Information-theoretic approach to measure feature relevance.

To combine the strengths of these different approaches, we developed an ensemble feature selection methodology that aggregates the rankings from each individual method. Features were ranked by each method, and an aggregate importance score was calculated as a weighted sum of these rankings. This ensemble approach proved more robust than any individual feature selection method. For the emergency visit prediction model, we selected 50 features based on their ensemble importance scores. For the Top-50 multi-label models, we selected varying numbers of features: 25 for medication prediction, 5 for diagnosis prediction, and 15 for procedure prediction. The Top-20 models generally used more features than their Top-50 counterparts. Interestingly, the Top-50 diagnosis model required only 5 features to achieve its best performance, suggesting that a few key indicators can effectively predict the most common diagnoses for CP patients. This finding aligns with clinical intuition that certain core features (like patient age, encounter frequency, and condition count) strongly influence diagnostic patterns. For hyperparameter tuning, we used a combination of grid search and Bayesian optimization to find the optimal model configuration. The hyperparameter search space included learning rate, number of hidden units, dropout rate, batch size, L2 regularization strength, and optimizer selection. We used 5-fold cross-validation to evaluate each hyperparameter configuration, using F1-score as the primary metric for optimization.

IV. Results

This section presents the results of our prediction models, with detailed performance metrics and feature importance to evaluate the contribution of different feature groups.

A. Emergency Visit Prediction Model

The emergency visit prediction model demonstrated strong performance across multiple evaluation metrics. Table V summarizes the key performance metrics on the training and test datasets.

TABLE V: Emergency Visit Prediction Model Performance

Metric	Training Set	Test Set
Accuracy	92.16%	91.43%
Precision	84.32%	82.79%
Recall	93.19%	92.65%
F1-Score	88.54%	87.79%
AUC-ROC	98.21%	97.53%

The high AUC-ROC score of 97.53% on the test set indicates excellent discriminative ability between patients who will and will not require emergency care in the following month. The relatively balanced precision and recall scores suggest that the model is effective at both identifying high-risk patients (high recall) and avoiding false alarms (high precision). Fig. 8 shows the metrics for the emergency visit prediction model, illustrating its highly accurate performance. The model's performance is particularly noteworthy given the challenge of predicting rare events like emergency visits. By employing class weighting and optimizing for appropriate evaluation metrics, we were able to achieve strong performance even with imbalanced data.

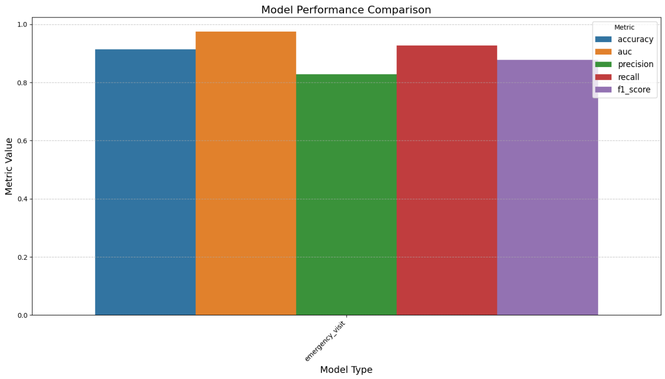

Fig. 8: Performance metrics for the emergency visit prediction model, showing high AUC (0.9753) and optimal decision threshold (0.603) that maximizes F1-score.

1) Feature Importance Analysis: Understanding which features contribute most to the model's predictions is crucial for both model interpretability and for providing actionable insights to healthcare providers. Fig. 9 illustrates the relative importance of the top 15 features as determined by our ensemble feature selection approach. These findings align with clinical intuition and previous research suggesting that patterns of healthcare utilization and medication complexity are strong predictors of future emergency care needs. The importance of vital signs like mean arterial pressure underscores the value of including clinical measurements in prediction models. Notably, several of our derived features, such as medication complexity indicators and healthcare utilization metrics, ranked among the most important predictors. This validates our feature engineering approach and highlights the importance of capturing complex patterns in patient care. These findings underscore the importance of comprehensive feature engineering that captures multiple dimensions of patient health and healthcare utilization. They also validate our approach of creating derived features that capture complex clinical patterns.

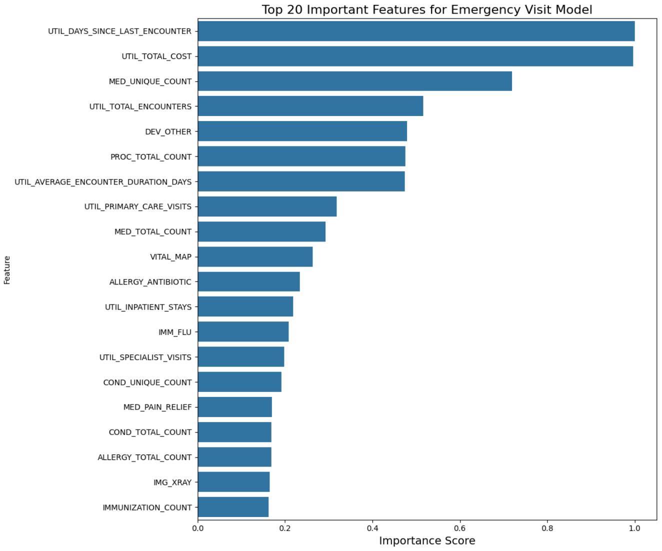

Fig. 9: Top 20 features for emergency visit prediction by importance score, showing the dominance of healthcare utilization metrics and medication-related features.

B. Medication, Diagnosis, and Procedure Prediction Models

While our emergency visit prediction model demonstrated excellent performance, the results for the multi-label prediction models were more nuanced, particularly when comparing Top-20 and Top-50 models.

1) Top-20 vs. Top-50 Model Performance: Fig. 10 provides a visual comparison of F1-scores between the Top-50 models across the three multi-label prediction tasks. The Top-20 models consistently outperformed their Top-50 counterparts across all three prediction tasks. This performance gap can be attributed to several factors:

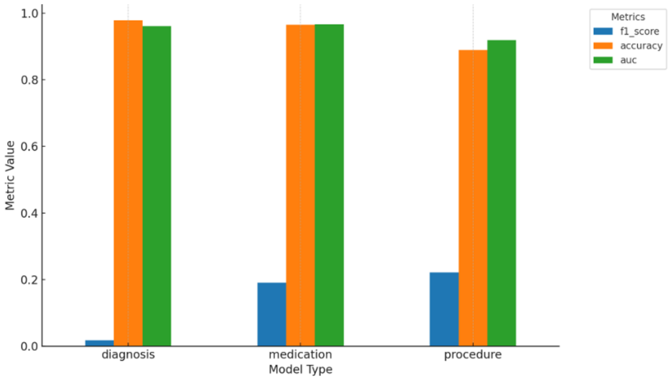

Fig. 10: Comparison of F1-scores, accuracy, and AUC between Top-50 multi-label prediction models, highlighting the significant performance gap due to the increased complexity and data sparsity in Top-50 models.

- **Data frequency**: The Top-20 labels occur much more frequently in the dataset, providing more training examples and enabling more robust learning.
- **Class imbalance**: The Top-50 models must contend with a more severe class imbalance, as the additional 30 labels are relatively rare events.
- **Prediction complexity**: The Top-50 models have 2.5 times more potential outcomes to predict, exponentially increasing the complexity of the prediction task.
- **Limited dataset size**: Our dataset, while substantial with over 4,000 patients, may not provide sufficient examples of the rarer clinical events to enable effective learning.

The exceptionally low F1-score (0.018) of the Top-50 diagnosis model is particularly striking. Further analysis revealed that the distribution of diagnoses in our dataset follows an extreme long-tail pattern, where the top 20 diagnoses account for over 80% of all diagnostic events. The remaining 30 diagnoses in the Top-50 model occur so infrequently that the model struggles to identify meaningful patterns related to these rare events. This finding has important implications for clinical practice, suggesting that focusing predictive efforts on the most common diagnoses might be more effective than attempting to predict all possible diagnoses. The excellent performance of the emergency visit prediction model and the Top-20 multi-label models demonstrates that high-quality predictions are achievable when sufficient data is available.

2) Feature Importance for Multi-Label Models: Similar to the emergency visit prediction model, we analyzed feature importance for the multi-label prediction models. Interestingly, while there is some overlap in important features across models (e.g., number of unique conditions, patient age), each prediction task also has unique important features. This highlights the value of developing separate models for each prediction task, rather than attempting a single model for all outcomes.

3) Training Convergence Analysis: To better understand the learning dynamics of our models, we analyzed the training history of each model, focusing on how key metrics evolved over training epochs. Fig. 11 shows the precision, recall, accuracy, and AUC curves for the medication prediction model during training. The medication model demonstrates a pattern common to all our models: rapid improvement in the early epochs, followed by more gradual gains and eventual plateauing. The Top-20 models consistently showed smoother convergence patterns than the Top-50 models, again reflecting the greater challenge posed by predicting rare events.

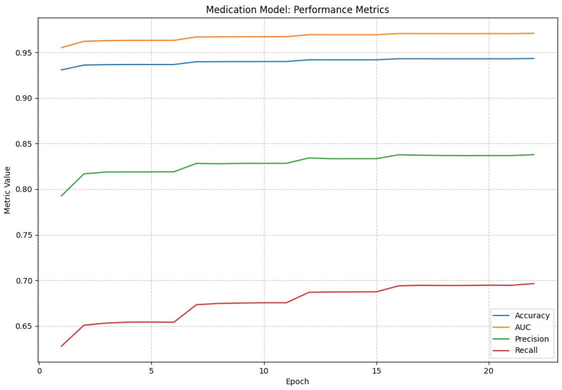

Fig. 11: Training and validation metrics over epochs for the medication prediction model, showing rapid initial convergence followed by more gradual improvement.

C. Comparison with Existing Approaches

Our emergency visit prediction model, with an AUC of 97.53%, outperforms previously reported models for predicting emergency care needs. Hong et al. [11] achieved an AUC of 0.83 for general emergency visit prediction, while Sun et al. [12] reported an AUC of 0.89 with their gradient boosting approach. Our superior performance may be attributed to several factors:

- The focus on a specific patient population (CP patients) with distinct healthcare patterns
- The comprehensive feature engineering approach, including derived features that capture complex clinical patterns
- The use of a neural network architecture optimized for this specific prediction task
- The timeline-based methodology that captures temporal trends in patient health

For the multi-label prediction tasks, direct comparisons with existing literature are challenging due to differences in prediction tasks, patient populations, and evaluation metrics. However, the strong performance of our Top-20 models (F1-scores between 0.72 and 0.78) suggests that our approach is effective for predicting common clinical events in CP patients. The lower performance of our Top-50 models aligns with observations in the literature about the challenges of predicting rare events in healthcare. Choi et al. [9] noted similar performance degradation when attempting to predict less common diagnostic codes, supporting our finding that focused models targeting frequent events tend to perform better than broader models attempting to predict rare events.

D. Impact of Data Volume on Model Performance

An important consideration for any machine learning approach is how model performance scales with data volume. The learning curves suggest that while our models achieve good performance with the current dataset, performance could be further improved with additional data. This is particularly relevant for the Top-50 multi-label models, where the poor performance on rare events might be addressed with a larger dataset. This finding has important implications for our future work. The relatively poor performance of the Top-50 models is likely not due to architectural limitations but rather to insufficient examples of rare events. With a larger dataset, these models might achieve performance comparable to their Top-20 counterparts, enabling more comprehensive prediction of clinical events for CP patients.

V. Discussion

Our results demonstrate the feasibility and potential value of machine learning approaches for predicting health outcomes in patients with cerebral palsy. The emergency visit prediction model achieved excellent performance, with an AUC of 97.53%, indicating its potential utility for identifying high-risk patients who might benefit from preventive interventions. The strong performance of the Top-20 multi-label models also suggests that accurate prediction of common medications, diagnoses, and procedures is achievable with current methods. However, the dramatic performance gap between Top-20 and Top-50 models highlights a significant challenge in healthcare prediction: the difficulty of predicting rare events with limited data. This challenge is particularly acute in specialized populations like CP patients, where some medications, diagnoses, and procedures occur very infrequently.

A. Clinical Implications

The strong performance of our emergency visit prediction model has direct clinical implications. By accurately identifying patients at high risk of requiring emergency care, healthcare providers could implement proactive interventions to address underlying issues before they escalate to emergencies. The ability to predict common medications, diagnoses, and procedures could further support proactive care planning, enabling healthcare providers to anticipate common needs and prepare accordingly. For example, knowing that a patient is likely to require a specific medication in the coming month could prompt advance prescription planning, potentially avoiding gaps in medication availability. The feature importance analysis also provides actionable insights for clinical practice. The strong influence of healthcare utilization patterns, medication regimens, and vital signs suggests that these areas should be prioritored in routine clinical assessments. Specifically, changes in the frequency of healthcare encounters, modifications to medication regimens, and shifts in vital sign patterns might serve as early warning indicators of potential health deterioration.

B. Methodological Innovations

Several methodological innovations in our work deserve highlighting:

- **Timeline-based feature engineering**: Our approach of creating monthly patient snapshots with a 12-month lookback period effectively captures temporal patterns in patient health. This methodology could be adapted for other chronic conditions where temporal trends are clinically significant.
- **Derived features**: The calculation of derived features that capture complex clinical patterns (e.g., medication complexity, healthcare utilization intensity) proved valuable, with many of these features ranking among the most important predictors. This approach could be extended to other clinical domains where raw data points alone may not capture the full complexity of patient health.
- **Ensemble feature selection**: Our approach of combining multiple feature selection techniques into an ensemble methodology yielded more robust feature sets than any individual method. This approach could be particularly valuable in healthcare domains where the relationship between features and outcomes may be complex and multifaceted.

C. Limitations and Future Work

Despite the promising results, our study has several limitations that suggest directions for future work:

- **Synthetic Data Limitations**: While Synthea generates realistic patient data, it may not capture all the nuances and complexities of real CP patients. Validation on real patient data, with appropriate privacy safeguards, would strengthen confidence in our models.
- **Dataset Size**: As our learning curves and the performance gap between Top-20 and Top-50 models suggest, a larger dataset might enable better prediction of rare events. Future work could explore generating a more extensive synthetic dataset or carefully incorporating real patient data.
- **External Validation**: The models were developed and validated using data from the same synthetic population. External validation on different populations would provide stronger evidence of generalizability.
- **Limited Contextual Information**: Our models primarily rely on structured clinical data and do not incorporate contextual factors like social determinants of health, caregiver support, or environmental factors that could influence health outcomes.
- **Time Horizon**: The current models predict outcomes for the next month only. Extending the prediction window to longer time horizons could enhance the clinical utility of the models.

Future work will address these limitations while expanding the scope of our prediction framework. Specific directions include:

- **Expanded Dataset Generation**: Creating a larger synthetic dataset with a focus on increasing examples of rare events to improve the performance of Top-50 models.
- **Real-world Validation**: Testing the models on real CP patient data to validate their performance and clinical utility.
- **Integration of Social Determinants of Health**: Incorporating socioeconomic factors, caregiver information, and environmental data to enhance prediction accuracy.
- **Extended Time Horizons**: Developing models that predict outcomes at 3, 6, and 12-month intervals to support longer-term care planning.
- **Specialized CP Subtype Models**: Creating models tailored to specific CP subtypes (spastic, dyskinetic, ataxic) and severity levels to improve prediction accuracy for different patient subgroups.
- **Clinical Decision Support Integration**: Developing interfaces and workflows to integrate these prediction models into clinical decision support systems for real-time use by healthcare providers.

VI. Conclusion

This paper presents a comprehensive machine learning framework for predicting health outcomes in patients with cerebral palsy. Using synthetic FHIR data representing over 4,000 CP patients, we developed models to predict emergency department visits, medication needs, diagnostic trends, and procedure requirements. The emergency visit prediction model achieved 97.53% AUC on test data, demonstrating excellent discriminative ability. The multi-label prediction models showed varied performance, with Top-20 models achieving strong F1-scores (0.72-0.78) while Top-50 models struggled with rare events (F1-scores of 0.018-0.22). Our timeline-based feature engineering methodology, which captures temporal trends in patient health, proved effective at generating predictive features. Feature importance analysis revealed that healthcare utilization patterns, medication complexity, and vital signs are strong predictors of healthcare outcomes. The disparity between Top-20 and Top-50 model performance highlights the challenge of predicting rare events and suggests that larger datasets will be necessary to achieve comprehensive prediction coverage. The strong performance of our emergency visit prediction model and Top-20 multi-label models demonstrates the potential of machine learning approaches to enhance care management for CP patients. By enabling proactive rather than reactive care, these prediction models could help improve patient outcomes, reduce healthcare costs, and enhance quality of life for individuals with cerebral palsy. While further validation and refinement are needed before clinical implementation, our work provides a solid foundation for future research in this area. The methodology and insights presented here may also be applicable to other chronic conditions where predictive modeling could support proactive care management.

References

[1] P. Rosenbaum, N. Paneth, A. Leviton, M. Goldstein, M. Bax, D. Damiano, B. Dan, and B. Jacobsson, ”A report: the definition and classification of cerebral palsy April 2006,” Developmental Medicine Child Neurology, vol. 49, no. s109, pp. 8-14, 2007.

[2] S. M. McPhail, M. Schippers, A. L. Marshall, M. Waite, and P. Kuipers, ”Perceived barriers and facilitators to increasing physical activity among people with musculoskeletal disorders: a qualitative investigation to inform intervention development,” Clinical Interventions in Aging, vol. 10, pp. 1037-1045, 2015.

[3] D. G. Whitney, R. L. Peterson, S. A. Warschausky, M. E. Hurvitz, and E. A. Hurvitz, ”Emergency department visits among children with and without disabilities,” Disability and Health Journal, vol. 12, no. 1, pp. 85-95, 2019.

[4] F. Torres, C. Lebrun-Harris, C. Okumura, and M. Morris, ”ED utilization among adult Medicaid beneficiaries with disabilities,” Disability and Health Journal, vol. 15, no. 2, p. 101246, 2022.

[5] A. L. Beam and I. S. Kohane, ”Big data and machine learning in health care,” JAMA, vol. 319, no. 13, pp. 1317-1318, 2018.

[6] T. Davenport and R. Kalakota, ”The potential for artificial intelligence in healthcare,” Future Healthcare Journal, vol. 6, no. 2, pp. 94-98, 2019.

[7] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan, ”Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record,” Journal of the American Medical Informatics Association, vol. 25, no. 3, pp. 230-238, 2018.

[8] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean, ”Scalable and accurate deep learning with electronic health records,” npj Digital Medicine, vol. 1, no. 1, pp. 1-10, 2018.

[9] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, ”Doctor AI: Predicting clinical events via recurrent neural networks,” in Machine Learning for Healthcare Conference, pp. 301-318, 2016.

[10] S. Yu, Y. Ma, J. Gronsbell, T. Cai, A. N. Ananthakrishnan, V. S. Gainer, S. E. Churchill, P. Szolovits, S. N. Murphy, I. S. Kohane, and T. Cai, ”Enabling phenotypic big data with PheNorm,” Journal of the American Medical Informatics Association, vol. 25, no. 1, pp. 54-60, 2018.

[11] W. S. Hong, A. D. Haimovich, and R. A. Taylor, ”Predicting hospital admission at emergency department triage using machine learning,” PloS one, vol. 13, no. 7, p. e0201016, 2018.

[12] B. Sun, A. B. Hao, Y. Wang, C. Li, L. Yang, Y. Zhu, and Z. Ye, ”Predicting emergency department visits and hospitalization in chronic disease patients using electronic health records and social determinants of health,” International Journal of Environmental Research and Public Health, vol. 20, no. 1, p. 359, 2023.

[13] E. Meehan, S. M. Reid, K. Williams, G. L. Freed, F. E. Babl, and J. R. Sewell, ”Hospital admissions in children with cerebral palsy: A data linkage study,” Developmental Medicine Child Neurology, vol. 58, no. 6, pp. 567-574, 2016.

[14] N. L. Young, C. Steele, D. Fehlings, J. Jutai, N. Olmsted, and J. I. Williams, ”Use of health care among adults with chronic and complex physical disabilities of childhood,” Disability and Rehabilitation, vol. 27, no. 23, pp. 1455-1460, 2011.

[15] N. Cremer, E. A. Hurvitz, and M. D. Peterson, ”Multimorbidity in middle-aged adults with cerebral palsy,” The American Journal of Medicine, vol. 130, no. 6, pp. 744.e9-744.e15, 2017.

[16] H. Sharma, C. Mao, Y. Zhang, H. Vatani, L. Yao, Y. Zhong, L. Rasmussen, G. Jiang, J. Pathak, and H. Xu, ”Developing a portable natural language processing based phenotyping system,” BMC Medical Informatics and Decision Making, vol. 19, no. 3, p. 78, 2019.

[17] C. Hong, J. Choi, J. Kim, Y. Kim, S. Park, J. Rew, and D. Yoon, ”Deep learning-based risk prediction for medication-related emergency department revisit,” PloS one, vol. 16, no. 4, p. e0249636, 2021.

[18] W. Liu, J. Wang, Z. Xu, B. Guo, S. Wu, and Y. Zhang, ”Multimodal learning from structured EHR data and unstructured clinical notes for medication recommendation,” in Proceedings of the 29th International Conference on Computational Linguistics, pp. 2547-2557, 2022.

[19] J. Kang, I. Chaudhari, S. Gao, F. Zhang, S. Wang, W. Zhang, C. Liu, Y. Xie, Y. Li, and B. Ying, ”FHIRformer: A universal deep Transformer for electronic health records in FHIR formats,” Nature Communications, vol. 14, no. 1, p. 384, 2023.

[20] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. Williamson, and F. Mahmood, ”Synthetic data in machine learning for medicine and healthcare,” Nature Biomedical Engineering, vol. 5, no. 6, pp. 493-497, 2021.

[21] Z. Wang, P. Peng, L. Ding, Z. Li, J. Zhang, and T. Han, ”A reinforcement learning-based framework for the generation of synthetic electronic health records,” in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 510-515, 2021.

[22] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. P. Bennett, ”Generation and evaluation of privacy preserving synthetic health data,” Neurocomputing, vol. 416, pp. 244-255, 2020.

[23] C. Thapa, S. Nasrin, S. Gupta, C. M. Shreya, F. Cavallo, and C. Zhang, ”Synthetic data in healthcare: A study of implicit model selections, feature importance, and imputation effects,” arXiv preprint arXiv:2307.05430, 2023.

[24] M. G. McIntyre, B. J. Levy, and P. T. Claypool, ”Cerebral palsy: Understanding the physical disability and management strategies,” Current Physical Medicine and Rehabilitation Reports, vol. 9, pp. 35-42, 2021.