

Software Engineering Department

Braude College

**Predicting Life Expectancy Based on Individual Lifestyle Indicators Using Machine Learning and Neural Network Models- [25-1-R-12]**

By:

Maher Salman

Adan Butto

Advisors :

Dr. Naomi Unkelos-Shpigel

Dr. Samah Idrees Ghazawi

GitHub :<https://github.com/mahersalman/life-expectancy>

**Table of Contents**

[**Abstract 3**](#_1fob9te)

[**1. Introduction 3**](#_30j0zll)

[**2. Related Work - Literature Review 4**](#_gbuhcfdtd86s)

[2.1. Life Expectancy Research: Key Studies and Predictive Models 4](#_e9iumx3j34ox)

[**2.2. Indicators for Life Expectancy 5**](#_3znysh7)

[2.2.1. Key Indicators and Their impact on Life Expectancy 5](#_x75qybp41ddv)

[2.2.2. Enhancing Predictive Accuracy with Additional Indicators 6](#_d7675fav314g)

[2.3. Machine learning in life expectancy 7](#_x4mqd13nrtv9)

[**3. Methodology and Research Process 7**](#_efsmo7b90bg7)

[3.1. Data Source 7](#_jkegvyeqaqsd)

[3.2. Preprocessing 8](#_nhu0kvksjjyk)

[3.2.1. Synthetic Individual Data Based on WHO 9](#_wy7ylz1quf2)

[3.3 Predictive Modeling Approaches for LE 11](#_qhdx9u276gla)

[3.3.1 XGBoost 11](#_wu1t0xr0qoe8)

[3.3.1.1 Flow of the XGBoost Model 13](#_wh4oqw23oehx)

[3.3.1.2 Example : Step-by-Step Flow of the XGBoost Model 14](#_jv4kd3a07sdw)

[3.3.2 SAINT: Self-Attention and Intersample Attention Transformer 16](#_xizooou2sq2p)

[3.3.2.1 Example Application of SAINT 17](#_m4l1kqtldvn8)

[**4. Evaluation 19**](#_fomguh5gn5ub)

[4.1. Comparative Evaluation of XGBoost and SAINT 19](#_iwzp8w3s1kb4)

[4.2. Metrics in Detail 20](#_eq2wuqv7g6kp)

[**5. Project Architecture and Deployment 21**](#_4iw9ifh7az0f)

[**6. Expected Result 22**](#_ir8d0754z7cm)

[**References: 23**](#_3dy6vkm)

# Abstract

Life expectancy (LE) is a critical measure of public health and longevity potential, offering valuable insights into societal well-being. This project aims to predict life expectancy by leveraging advanced machine learning (ML) and deep learning (DL) techniques. The study examines a comprehensive set of lifestyle, behavioral, and socioeconomic indicators using real-world data from reputable sources, including the World Health Organization (WHO) and the National Health and Nutrition Examination Survey (NHANES). Key factors such as Body Mass Index (BMI), physical activity, smoking habits, alcohol consumption, and access to healthcare are analyzed to understand their contributions to LE prediction.

The project employs a comparative analysis of various predictive models, including ensemble methods XGBoost, and advanced deep learning models such as SAINT. These models are rigorously optimized and evaluated using metrics RMSE, MAE, and R² to ensure accuracy and generalizability. By integrating traditional ML techniques with state-of-the-art DL approaches, this research seeks to enhance prediction accuracy and provide actionable insights for public health planning. The findings aim to demonstrate the transformative potential of machine learning in life expectancy forecasting, contributing to advancements in predictive modeling and personalized healthcare strategies.

# 1. Introduction

Life expectancy (LE) is a vital measure of public health and societal well-being. It offers insights into the average potential longevity of individuals within a population. For individuals at a given age, it estimates the additional years they are likely to live based on prevailing death rates. By summarizing mortality patterns, life expectancy serves as a key indicator of health outcomes, societal progress, and the effectiveness of healthcare systems [1].

Predicting life expectancy is a multifaceted challenge influenced by a wide range of factors. Personal attributes such as body mass index (BMI), physical activity levels, and smoking habits interact with systemic determinants like access to healthcare, socioeconomic status, and environmental conditions[11]. These variables often have complex, non-linear relationships, making life expectancy prediction both complicated and rich in potential insights. Accurate predictions of LE are essential for public health planning, enabling policymakers to identify key determinants and design targeted interventions to improve health and longevity.

While previous research has explored individual indicators of life expectancy, integrating machine learning presents a transformative approach to tackling this problem. Machine learning algorithms allow for the simultaneous analysis of multiple factors, uncovering intricate patterns and relationships with exceptional precision and scalability[1]. By leveraging these capabilities, predictive models can be constructed to estimate life expectancy with greater accuracy than traditional statistical methods.

This project builds on these advancements by utilizing real-world data from diverse sources to ensure a comprehensive representation of indicators. These indicators include behavioral factors like smoking and physical activity, physiological measures such as BMI, and systemic factors like healthcare access and income levels. By testing and comparing machine learning and deep learning models—including neural networks, and ensemble methods—this project aims to identify the most effective approach for predicting life expectancy. Through rigorous model optimization and validation, the results are designed to be both accurate and generalizable, providing valuable tools for improving public health outcomes across diverse populations.

# 2. Related Work - Literature Review

Understanding the factors influencing life expectancy is a critical area of research with profound implications for public health and policy-making. Numerous studies have explored the relationship between health, economic, and social indicators and life expectancy, leveraging both traditional statistical approaches and modern machine learning techniques. These works provide valuable insights into the predictors of longevity and form the foundation for more advanced modeling and analysis. Below, we review key studies that have significantly contributed to this field.

## 2.1. Life Expectancy Research: Key Studies and Predictive Models

1) The project “What Really Drives Higher Life Expectancy?” analyzed global data from the World Health Organization, focusing on identifying key factors that influence life expectancy across countries from 1960 to 2015. Using a regression model, the study aimed to predict life expectancy based on 21 development indicators, achieving a high degree of accuracy. Significant predictors highlighted in the study included access to sanitation facilities, which showed a strong positive correlation (r = 0.88), and gross domestic product (GDP) per capita, which also had a positive correlation (r = 0.698). Other key factors were health expenditure per capita (r = 0.63), percentage of the population in rural areas, which exhibited a strong negative correlation (r = -0.80), and adolescent fertility rate, which had a negative correlation (r = -0.77).

The regression model achieved an accuracy of 89.7% using all 21 indicators. A simplified model that used only access to sanitation and rural population percentage maintained an impressive accuracy of 80.4%. Notably, the study found that a 0.18% increase in access to sanitation corresponded to a one-year increase in life expectancy, while a 0.11% decrease in the rural population percentage resulted in a one-year increase in life expectancy.[21]

2) The article **“An Application of a Supervised Machine Learning Model for Predicting Life Expectancy”** investigates the use of machine learning techniques, specifically the Extreme Gradient Boosting (XGBoost) algorithm, to forecast life expectancy across various countries. The study focuses on identifying key determinants influencing life expectancy by analyzing data from the World Health Organization (WHO) and the United Nations (UN).

Key factors examined include:

**Health Indicators**: Adult mortality rate, infant deaths, immunization coverage (e.g., Polio, Hepatitis B), and disease prevalence (e.g., HIV/AIDS).

**Economic Factors**: GDP per capita, total health expenditure, and alcohol consumption.

**Social Factors**: Average years of schooling, population demographics, and country development status.

The XGBoost algorithm demonstrated exceptional performance, surpassing earlier models in accuracy and handling dataset complexity. The results highlighted significant predictors of life expectancy, such as adult mortality rate, GDP per capita, and immunization coverage. These findings emphasize the importance of enhancing healthcare access, economic growth, and disease prevention to positively impact life expectancy [1].

## 2.2. Indicators for Life Expectancy

### 2.2.1. Key Indicators and Their impact on Life Expectancy

Using insights from previous studies, we selected key indicators with a high correlation to life expectancy to serve as the foundation of our analysis. We will be utilizing machine learning (ML) techniques to evaluate the relevance and predictive power of these indicators and explore how adding additional factors could refine our models. Below, we define each indicator, summarize the observed results, and explain its significance in life expectancy research.

**Table 1. Key Determinants and Indicators of Life Expectancy: Explanations and Insights**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Indicator** | **Explanation** | **REF** |
| 1 | **Country** | Life expectancy varies significantly across countries due to differences in healthcare systems, economic conditions, cultural practices, and environmental factors. Even when other indicators remain constant, the disparities in life expectancy highlight the impact of geographical and systemic variations. Understanding these differences provides critical insights into the unique challenges and opportunities present in various regions, aiding in the development of targeted health interventions and policies. |  |
| 2 | **Body Mass Index (BMI)** | BMI is a widely recognized measure of body fat based on weight and height, categorizing individuals as underweight, normal weight, overweight, or obese. Studies reveal that individuals with a BMI in the range of 20.0–25.0 exhibit the lowest mortality risk (Hazard Ratio = 1.00). Conversely, being underweight (BMI < 20.0) or overweight/obese (BMI > 25.0) significantly increases mortality risk, with the hazard ratio reaching as high as 2.76 for severe obesity (BMI 40.0–60.0). These findings underscore the importance of maintaining a healthy BMI to mitigate health risks and enhance longevity . | [4] |
| 3 | **Physical Activity** | The frequency and intensity of physical activity are critical factors influencing life expectancy. Participants categorized as ‘Weekend Warriors’ (exercising 1–2 times per week) or ‘Regularly Active’ (frequent exercise) demonstrated reduced mortality risks compared to inactive individuals, with hazard ratios of 0.86 and 0.85, respectively. These findings highlight the significant health benefits of physical activity, even when performed sporadically, and support the promotion of exercise as a viable strategy for increasing life expectancy . | [5] |
| 4 | **Smoking** | Smoking remains one of the most significant contributors to preventable mortality globally, accounting for approximately 20% of adult deaths in studied regions. Eliminating smoking-related deaths could increase life expectancy by an average of 2.4 years for men and 1 year for women. These statistics emphasize the urgent need for effective tobacco control policies to reduce smoking prevalence and its associated health burden, thereby improving overall life expectancy . | [6] |
| 5 | **Alcohol Consumption** | Excessive alcohol consumption is closely linked to higher mortality rates and a range of chronic health conditions. Research indicates that consuming more than five drinks per week leads to a measurable reduction in life expectancy, with individuals drinking 10 or more drinks per week losing 1–2 years of life expectancy and those drinking 18 or more losing 4–5 years. These findings highlight the importance of promoting moderation in alcohol consumption to mitigate its adverse effects on health and longevity . | [7] |
| 6 | **Dietary Habits** | Healthy dietary habits, such as consuming adequate portions of fruits, vegetables, whole grains, and lean proteins, are strongly associated with better health outcomes and increased life expectancy. Balanced diets help reduce the risk of chronic diseases such as cardiovascular issues, diabetes, and obesity, contributing to improved longevity. This indicator underscores the role of nutrition in shaping public health and guiding dietary recommendations . | [20] |
| 7 | **Sleep Patterns** | Sleep quality and duration play a critical role in overall health and longevity. Poor sleep patterns, whether in terms of insufficient duration or inadequate quality, are linked to increased risks of chronic diseases and reduced life span. This highlights the necessity of promoting healthy sleep behaviors as part of a holistic approach to improving life expectancy and preventing mortality . | [20] |

### 2.2.2. Enhancing Predictive Accuracy with Additional Indicators

To further refine the analysis, additional indicators were considered to enhance predictive accuracy and provide a comprehensive understanding of life expectancy determinants.

**Table 2. Additional Indicators Enhancing the Analysis of Life Expectancy Determinants**

|  | **Indicator** | **Explanation** | **REF** |
| --- | --- | --- | --- |
| 1 | **Schooling** | Education levels are closely associated with better health outcomes and reduced mortality risks. Each additional year of schooling reduces the risk of mortality by 1.9%, with individuals having 12 years of schooling enjoying a 24.5% lower mortality risk compared to those with no formal education. This reflects the profound influence of education on health awareness, healthcare access, and socioeconomic conditions. | [9] |
| 2 | **Socioeconomic Status** | Factors such as income levels, employment, and social standing significantly impact health outcomes. Higher socioeconomic status correlates with improved access to healthcare services, healthier lifestyles, and longer life expectancy, emphasizing the need for policies aimed at reducing inequalities to promote health equity . | [10] |
| 3 | **Access to Clean Water and Sanitation** | Access to clean water and proper sanitation facilities is essential for reducing mortality from communicable diseases. Regions with improved infrastructure in this regard show lower mortality rates, reflecting the fundamental role of clean water and sanitation in enhancing life expectancy. | [11] |
| 4 | **Healthcare Access** | The availability and utilization of healthcare services, including advancements in telemedicine, are pivotal in improving health outcomes and life expectancy. Regions with better healthcare access demonstrate higher life expectancy due to early diagnosis, timely treatment, and effective disease management. | [12] |

## 2.3. Machine learning in life expectancy

In related studies, commonly used machine learning techniques for life expectancy prediction include linear regression and multiple regression models, which are effective for analyzing relationships between indicators and life expectancy. Ensemble methods such as random forests and decision trees are widely employed for feature importance analysis and to enhance predictive performance [11]. Advanced techniques like support vector machines (SVM) are utilized to model complex non-linear relationships, while gradient boosting algorithms, such as XGBoost, are favored for handling large datasets and optimizing predictions[1].

Despite these advancements, existing approaches still face limitations. A key challenge lies in the generalizability of these models, as they often struggle to adapt to diverse datasets from different countries or regions [4]. Additionally, publicly available datasets tend to represent averaged data [2] , which weakens the ability to identify precise relationships between life expectancy and specific indicators. Addressing these gaps is crucial for developing more accurate, adaptable, and practical life expectancy prediction models that can offer actionable insights to inform global health policies and strategies.

# 3. Methodology and Research Process

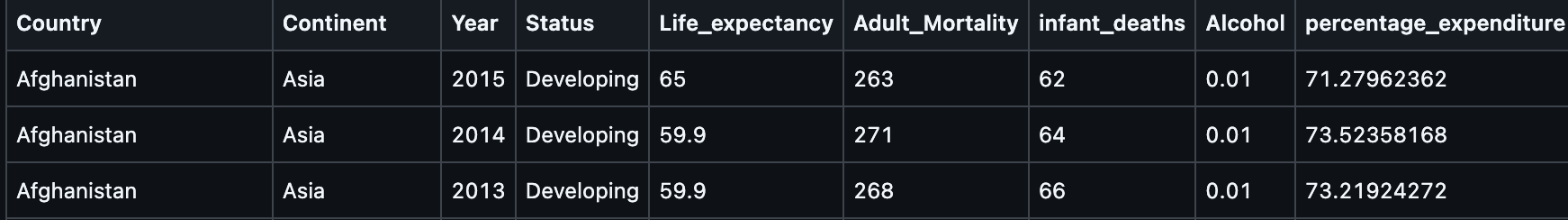
We began our research by exploring the concept of life expectancy (LE) and identifying related indicators with a strong correlation to LE. Initially, we searched for a dataset that provided individual-level data, but unfortunately, such a dataset did not exist. Instead, we found datasets like the one from WHO [23], which contains averaged data based on various indicators. A larger version of this dataset was also available [22]. Due to privacy concerns, we could not access individual-level data, so we decided to create a synthetic individual dataset. To ensure the realism and accuracy of this synthetic dataset, we based it on the average data we found and used a model called GAN, which we will discuss further in Section 3.2.1 .

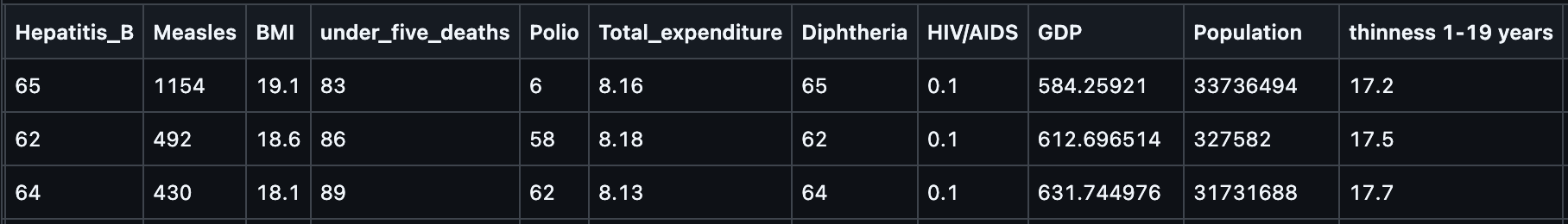
Given that our dataset is tabular, we sought machine learning and neural network models with high accuracy on medical and tabular data. We identified XGBoost, a well-known model that has excelled in Kaggle challenges and medical data experiments[3], achieving top accuracy on tabular data [2]. While exploring neural networks, we noted their limitations with tabular data [24] until discovering a model called SAINT, introduced in 2021. SAINT claimed superior accuracy on tabular data compared to gradient boosting models like XGBoost, outperforming them in several datasets[13]. Therefore, we decided to compare the performance of XGBoost and SAINT on our synthetic dataset to determine which algorithm achieves better accuracy.

## 3.1. Data Source

The Life Expectancy dataset, obtained from Kaggle [23], provides valuable insights for predicting life expectancy using various factors over a 15-year span. This dataset is based on information from the Global Health Observatory (GHO) under the World Health Organization (WHO), which tracks health status and related factors for all countries. It includes data on health factors for 193 countries collected from the WHO data repository, along with corresponding economic data obtained from the United Nations [22]. This comprehensive dataset has been made publicly available for health data analysis and is suitable for machine learning and data visualization purposes. Although previous studies have explored factors influencing life expectancy—such as demographic variables, income composition, and mortality rates—the effects of immunization and the human development index had not been fully accounted for in earlier analyses [23].

Each row in the dataset contains information about various parameters, including the country, continent, year, and status, as well as key indicators such as life expectancy, adult mortality, infant deaths, alcohol consumption, and GDP. These variables, along with other health and economic indicators, offer a rich foundation for understanding global patterns in life expectancy and the socio-economic factors that influence it. See Fig. 1 for an example of the dataset showcasing these indicators [22].







**Fig. 1 [22] - Example of WHO data showcasing various indicators**

## 3.2. Preprocessing

Preprocessing is a vital step in preparing raw data for machine learning analysis and it ensures consistency and reliability. In this study, a structured pipeline was implemented to address missing values, transform categorical variables, normalize numerical features, and split the dataset into training, validation, and test subsets for accurate and generalizable life expectancy predictions.

The process begins with data cleaning, where missing values are handled using imputation techniques. For instance, missing Body Mass Index (BMI) values are replaced with averages from participants with similar demographics or lifestyles. Outliers in critical variables, such as extreme BMI values, were identified and mitigated to minimize their impact on the model.

Categorical variables, such as smoking status and physical activity levels, were transformed into numerical representations using one-hot encoding, enabling machine learning algorithms to process them effectively. Numerical variables, including GDP and BMI, were normalized and standardized to ensure uniformity, preventing features with larger ranges from dominating the model and enhancing training efficiency.

The dataset was then split into training, validation, and test sets in a stratified manner, preserving the diversity and distribution of the original data. This approach ensures the model is trained on representative samples and tested on unseen data, providing an unbiased evaluation of its performance.

By addressing missing values, transforming features, and scaling data appropriately, the preprocessing pipeline refines the raw dataset into a robust format, ready for machine learning analysis. These steps are critical for building reliable and generalizable predictive models in this study.

Obtaining accurate and real individual-level data is challenging due to its private nature and limited availability. To address this, we utilize aggregated data provided by WHO, which represents averages across various indicators. The following approach is employed to preprocess and adapt this data for our analysis:

### 3.2.1. Synthetic Individual Data Based on WHO

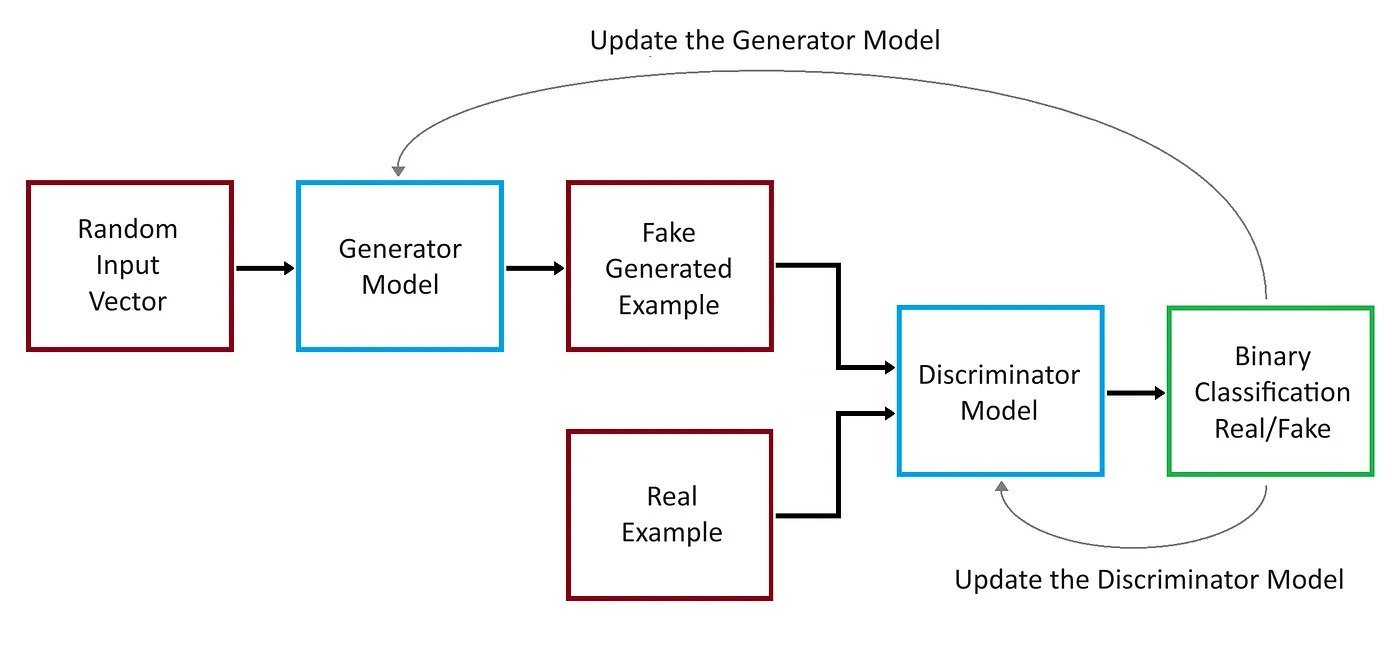
In this research, due to the difficulty of reaching individual-level healthcare data, we propose leveraging Generative Adversarial Networks (GANs) to generate individual-level data based on aggregated averages provided by WHO datasets. We will be synthesizing the life expectancy (LE) and other related indicators (e.g., BMI, alcohol consumption, access to water).

GANs are a class of generative models that employ a two-network architecture: a generator and a discriminator, working in a zero-sum game framework. The generator synthesizes data points that mimic the statistical properties of real data, while the discriminator evaluates the authenticity of the generated data.

Through adversarial training, these two networks iteratively improve each other, enabling the generation of realistic synthetic data. This approach allows us to create high-resolution individual-level data that aligns closely with the aggregated averages, preserving essential patterns and relationships.

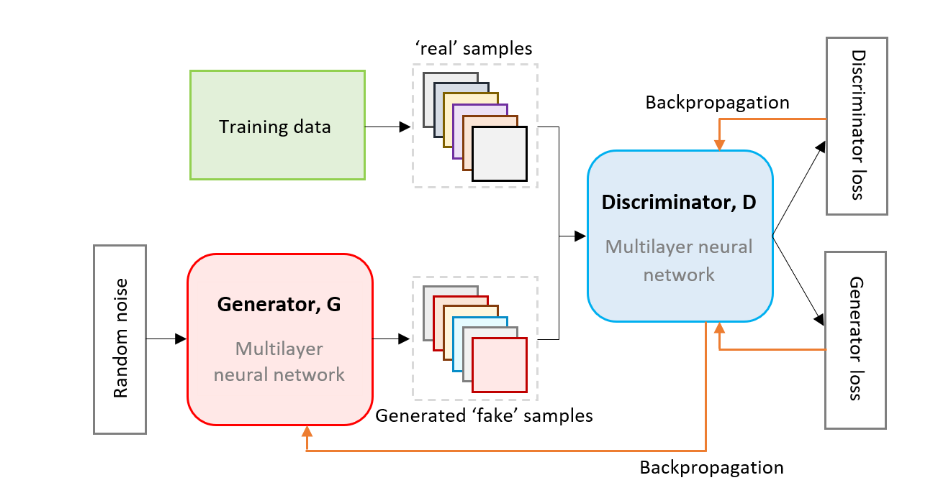
Generative Adversarial Networks (GANs) are powerful tools for generating synthetic data, as they can effectively capture the correlations between features and are scalable for large datasets. [8] GANs can generate high-quality synthetic data to benefit data sciences[8]. At the core of a GAN is the interaction between two neural networks: the generator and the discriminator. The generator creates synthetic data by starting with a noise vector, sampled from a random distribution such as Gaussian or uniform, and passing it through a series of non-linear transformations governed by its weights and biases. This transformation results in synthetic data points that aim to mimic real-world data.

The discriminator acts as a binary classifier, evaluating whether the input data is real (from the original WHO dataset) or synthetic (produced by the generator). It outputs a probability score indicating the likelihood that the input is real. The generator and discriminator are trained in an adversarial setup, where the discriminator strives to improve its accuracy in distinguishing real from fake data, while the generator works to minimize the discriminator’s ability to differentiate between them. **[Fig. 2a]**



**Fig. 2a: Workflow of a Generative Adversarial Network (GAN). [16]**

Training a GAN involves iterative optimization of both networks. Initially, the discriminator is trained by feeding it batches of real data from the WHO dataset and batches of synthetic data from the generator. Its loss is computed using a binary cross-entropy function, and its weights are updated via backpropagation to improve its ability to correctly classify data as real or fake. Subsequently, the generator is trained to produce synthetic data that the discriminator cannot distinguish from real data. The generator’s loss is computed based on the discriminator’s feedback, and its weights are adjusted to generate data that appears more realistic **[Fig. 2b]**. This process is repeated iteratively, with both networks improving over time—the generator creating increasingly plausible data and the discriminator becoming more adept at identifying fakes.



**Fig. 2b: Demonstration of GAN architecture. [15]**

For structured tabular data, such as the WHO dataset, additional enhancements are applied to adapt GANs effectively. The generator is conditioned on features such as region and year to ensure that the generated data aligns with the contextual distributions in the real dataset. Moreover, advanced techniques such as Wasserstein GAN with Gradient Penalty (WGAN-GP) are utilized to improve training stability and prevent issues like mode collapse, where the generator produces a limited variety of outputs.

By employing these methods, the GAN learns to generate individual-level data points that are statistically consistent with aggregated averages while preserving complex, non-linear relationships between variables. This approach provides a robust framework for creating realistic synthetic data tailored to specific contexts, enabling applications such as policy simulations, predictive modeling, and granular trend analysis in health and demographic studies.

## 3.3 Predictive Modeling Approaches for LE

To predict life expectancy (LE), we employ a machine learning model and a deep learning approach. Each model is chosen based on its unique strengths and suitability for addressing the complexities of LE prediction.

In this study, we utilize **XGBoost**, an ensemble-based gradient boosting algorithm known for its efficiency and high accuracy, particularly in handling tabular data. XGBoost sequentially improves weak models, making it a robust choice for LE prediction. Additionally, XGBoost has demonstrated significant effectiveness in medical data applications, handling complex and diverse datasets while meeting accuracy requirements for auxiliary diagnosis [3]. It also serves as a baseline for comparing its performance against deep learning approaches, which are explored later in this research.

For this research, we could not find previous studies or projects that employed deep learning models for predicting life expectancy. This gap may stem from the challenges deep learning faces when applied to tabular datasets, particularly small ones, where traditional machine learning approaches like gradient boosting often outperform.[2]

Despite these limitations, we chose to explore a deep learning approach to assess its potential and examine the accuracy achievable in this domain. To this end, we selected the SAINT algorithm [13], which has demonstrated superiority over gradient boosting methods, including XGBoost, even on smaller datasets. This exploration seeks to evaluate the feasibility of deep learning for tabular data and its potential for life expectancy prediction.

### 3.3.1 XGBoost

XGBoost, or eXtreme Gradient Boosting, is a powerful and efficient machine learning algorithm designed to enhance the performance of gradient boosting models. By leveraging sequential decision trees and optimizing their outputs, XGBoost iteratively refines predictions, making it an essential tool for both research and real-world applications. Its robust features, including scalability, flexibility, and efficiency, have made it a popular choice in machine learning competitions and industry use cases.

As illustrated in Fig. 3, XGBoost represents the evolution of boosting techniques, building on the strengths of decision trees, bagging, and gradient boosting to deliver optimized gradient boosting. This evolutionary process demonstrates how XGBoost refines its approach to achieve greater efficiency and predictive power.

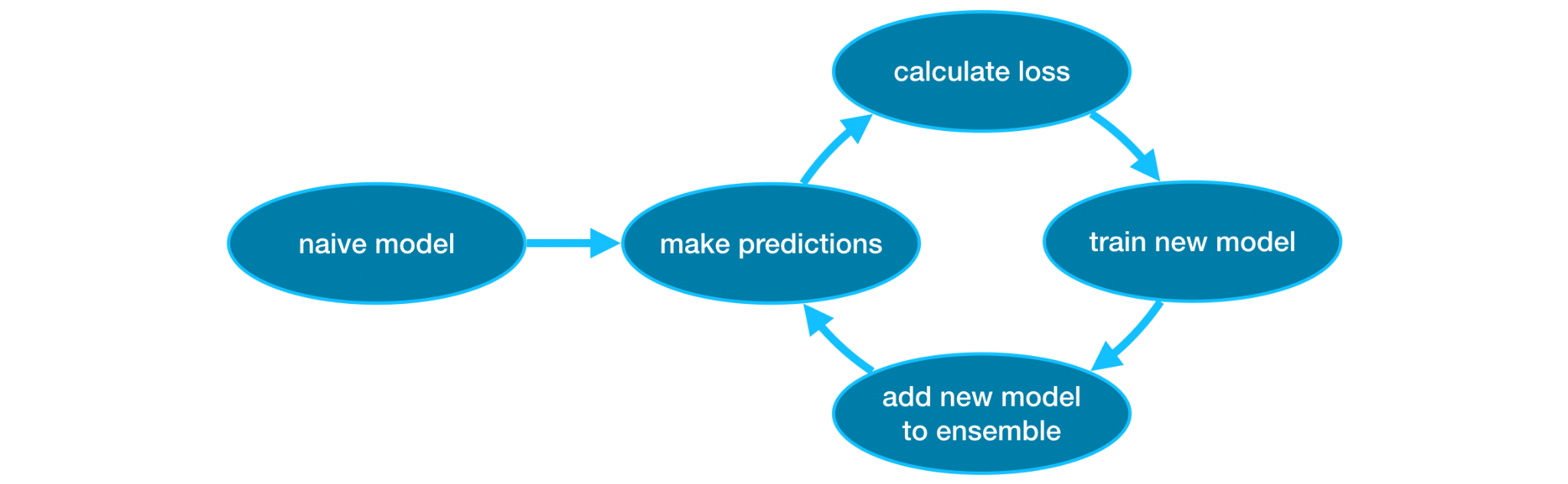
XGBoost’s primary strength lies in its ability to convert weak learners, in our case a decision tree, into a strong predictive model through an iterative boosting process. Each subsequent tree focuses on correcting the residual errors from the previous iterations, allowing the model to capture intricate patterns in the data. This step-by-step refinement ensures that the final ensemble model is both accurate and robust.

In addition to its boosting mechanism, XGBoost is known for its computational efficiency and scalability. It employs parallel processing and memory-efficient techniques, enabling it to handle large datasets with high-dimensional features effectively. Another crucial feature is its ability to automatically handle missing data, allowing users to train models without extensive preprocessing. To prevent overfitting, XGBoost support L1 (lasso) and L2 (ridge) regularization, which penalize overly complex models.

# 

**Fig. 3: Evolution of XGBoost: From Decision Trees to Optimized Gradient Boosting [17]**

#### 3.3.1.1 Flow of the XGBoost Model



**Fig. 4: Workflow of the XGBoost Model Showing Iterative Training and Loss Minimization [26]**

The flow of the XGBoost model involves several key steps, starting from defining the objective to iteratively building trees that enhance prediction accuracy. **Fig. 4** illustrates the workflow of the XGBoost model, highlighting its iterative process from making initial predictions to adding new models to the ensemble and minimizing loss through repeated refinements. Below is a detailed breakdown of this process:

**Define the Objective Function:** XGBoost minimizes a loss function by adding trees sequentially. For regression tasks, the loss function is typically the **Mean Squared Error (MSE)**, which quantifies the difference between the actual values and the predicted values.

The MSE is expressed as:

Here, is the Actual value and Predicted value, Land is the loss function, specifically the Mean Squared Error in this case.

This formulation ensures that XGBoost iteratively adjusts its predictions to minimize the squared errors between and, improving the accuracy of the model with each additional tree.

**Initialize the Model:** Start with an initial prediction, typically the mean value of the target variable for regression tasks: , this serves as the baseline for subsequent iterations.

**Calculate Residuals:** To improve predictions, XGBoost computes the residuals between the actual values and the current predictions. The residuals represent the direction and magnitude of adjustment needed to minimize the loss. These are calculated as the negative gradient of the loss function with respect to the predicted values:

For regression tasks, such as those using the Mean Squared Error (MSE) as the loss function, the residuals simplify to:

Here, is the loss function, specifically the Mean Squared Error, is the actual target value, is the predicted value at iteration t, and is the residual at iteration t.

The residuals serve as the basis for the next tree in the sequence, allowing the model to iteratively reduce errors and improve its predictions. By incorporating this gradient-based approach, XGBoost systematically refines the model’s accuracy.

**Fit a Tree to the Residuals:** To predict the residuals, a decision tree is constructed. This tree identifies patterns in the data that previous predictions missed, enabling the model to refine its performance. Each leaf node in the tree corresponds to a predicted residual value.

**Compute Leaf Values:** Once the tree is built, the next step is to compute the optimal values for each leaf node that minimize the loss function. This involves calculating the weighted sum of gradients and applying regularization penalties to avoid overfitting and overly complex trees.

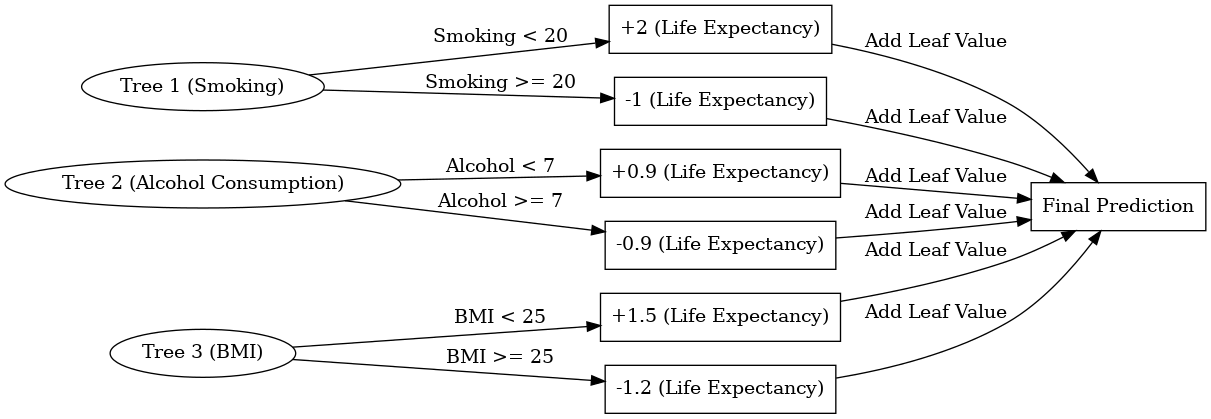
**Update Predictions:** Update the model’s predictions by adding the contribution of the new tree:

Here , is the learning rate (controls the contribution of each tree) and is the prediction of the new tree for input .

**Iterate Through Multiple Trees:** The process of calculating residuals, fitting trees, and updating predictions is repeated for a predefined number of iterations (K) or until a stopping criterion is met. This iterative approach enables the model to continuously refine its predictions.

**Generate Final Predictions:** After completing all iterations, the final prediction is obtained by summing the outputs of all trees, weighted by the learning rate:

#### 3.3.1.2 Example : Step-by-Step Flow of the XGBoost Model



**Fig. 5: Simulation of XGBoost Using Smoking, Alcohol Consumption, and BMI to Predict Life Expectancy**

Here’s a simulation of how XGBoost works using the indicators of **Smoking**, **Alcohol Consumption**, and **BMI** to predict **Life Expectancy**. Each step includes the **formulas** for better understanding.

**Fig. 5** illustrates how XGBoost predicts life expectancy by iteratively refining predictions through a series of decision trees. The process begins with an initial prediction, such as the mean life expectancy in the training data, set at 70 years . Each tree contributes adjustments to this initial prediction based on specific decision rules derived from features like Smoking, Alcohol Consumption, and BMI.

To calculate this step-by-step, consider the following example data:

• **Smoking**: 15 (Smoking < 20, so the tree adds +2 to the prediction)

• **Alcohol Consumption**: 6 (Alcohol < 7, so the tree adds +0.9 to the prediction)

• **BMI**: 24 (BMI < 25, so the tree adds +1.5 to the prediction)

The first step is to calculate the residuals for each observation. Residuals are defined as the difference between the actual observed values and the predicted values. They play a critical role in guiding the corrections made by subsequent trees. For instance, if the actual life expectancy for a data point is 73 years and the initial prediction is 70, the residual is calculated as:

This residual informs the first tree about the error it needs to address.

**In the first iteration**, the model trains a decision tree to minimize the residuals. The first tree uses the feature Smoking and applies the decision rule: if Smoking is less than 20, the tree adds +2 to the prediction; otherwise, it subtracts 1. For the given example, Smoking is 15, so . The updated prediction after this tree is expressed as:

The new residual is then recalculated as:

**In the second iteration**, the model focuses on the residuals left by the first tree. The second tree uses the feature Alcohol Consumption and applies the decision rule: if Alcohol Consumption is less than 7, the tree adds +0.9 to the prediction; otherwise, it subtracts 0.9. For the given example, Alcohol Consumption is 6, so . The updated prediction becomes:

The residual is updated again as:

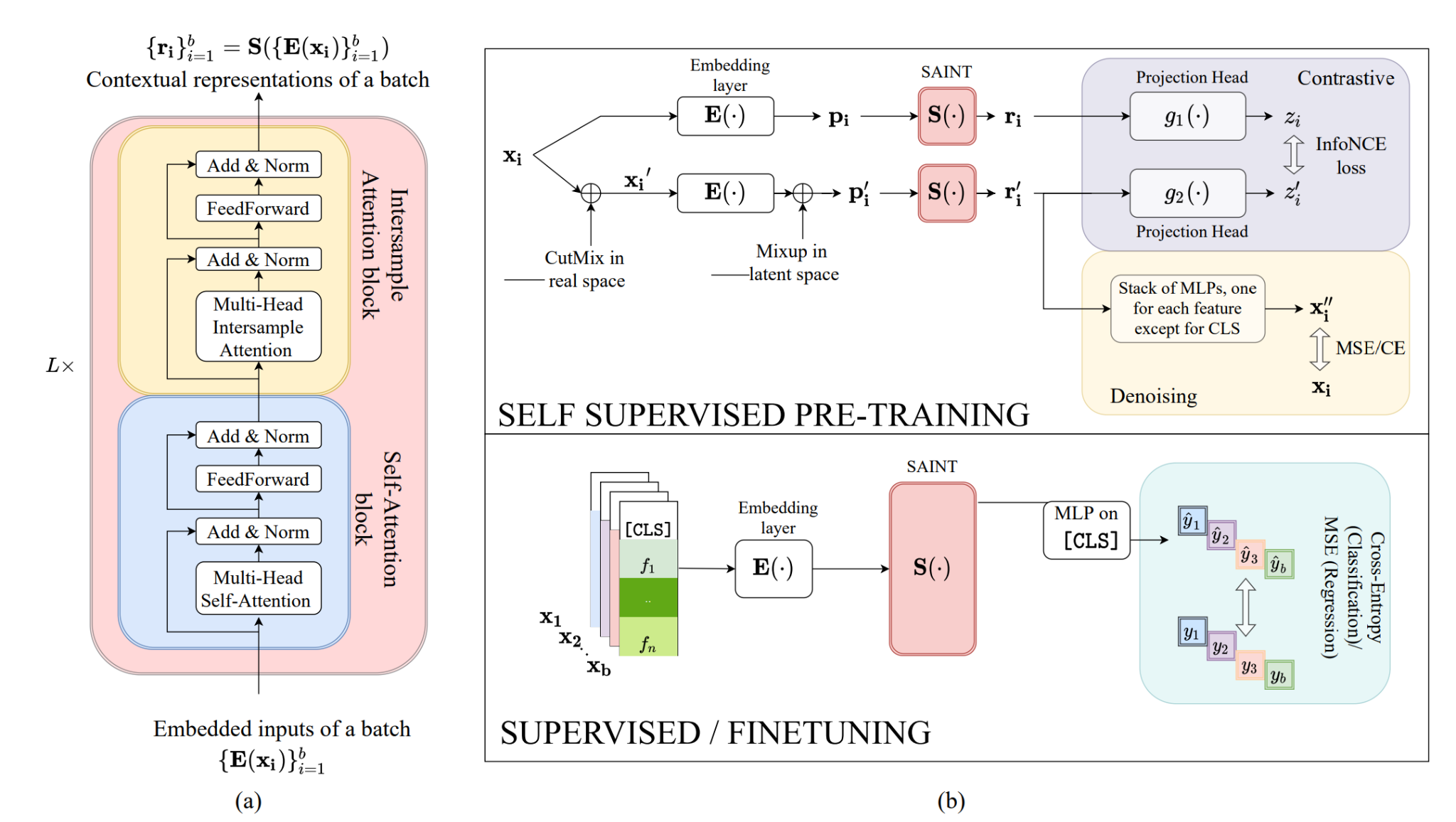
**Finally**, the third tree addresses the remaining residuals using the feature BMI. The decision rule for this tree is: if BMI is less than 25, the tree adds +1.5 to the prediction; otherwise, it subtracts 1.2. For the given example, BMI is 24, so . The updated prediction becomes:

Equivalently**,** The final prediction is achieved by summing the contributions from all trees, as shown in Fig. 5 :

This iterative process ensures that each tree incrementally reduces the error by focusing on the residuals from the previous step. By combining the initial prediction with the adjustments from all three trees, the model produces a final prediction of , which is closer to the actual value of 73. This example demonstrates the effectiveness of XGBoost in refining predictions through its boosting mechanism.

### 3.3.2 SAINT: Self-Attention and Intersample Attention Transformer

SAINT (Self-Attention and Intersample Attention Transformer) [13] represents a breakthrough in deep learning for tabular data. Its architecture integrates advanced attention mechanisms and data augmentation techniques to model feature-level and intersample relationships, overcoming traditional challenges with tabular datasets. Unlike standard models, SAINT handles mixed data types, including numerical and categorical features, without requiring extensive preprocessing. It offers state-of-the-art performance on classification and regression tasks, even outperforming traditional tree-based models like XGBoost, LightGBM, and CatBoost.

****

**Fig. 6: Architecture of SAINT for Self-Supervised Pre-Training and Supervised Fine-Tuning[13]**

The architecture of **SAINT** is crafted to address the unique challenges of tabular data, incorporating transformer-based mechanisms for embedding and modeling relationships within and across data samples. The architecture’s components, as depicted in **Fig 6(a)** and **Fig 6(b)**, play distinct roles in processing the data to ensure optimal performance.

The **embedding layer** processes both numerical and categorical features by converting them into dense vector representations suitable for the transformer model. Numerical features, such as BMI or alcohol consumption, are passed through fully connected layers with ReLU activation to transform the raw values into high-dimensional embeddings. Categorical features, like occupation or gender, are embedded using learnable embedding layers, akin to word embeddings in NLP. This embedding step ensures that all features, regardless of type, are represented in a uniform format and concatenated to form a single row embedding, encapsulating all feature information for that sample.

The **self-attention blocks**, highlighted in **Fig 6(a)**, model feature-level interactions within a single row. By using a multi-head attention mechanism, these blocks capture how features within a sample relate to one another. For example, in a dataset with BMI, smoking, and alcohol consumption, self-attention determines how these variables jointly influence the target variable, which in our case is life expectancy. This refined representation enriches the model’s understanding of the internal structure of each sample.

In contrast, the **intersample attention blocks** extend the focus beyond individual rows to the relationships across samples, as shown in **Fig 6(a)**. These blocks employ multi-head attention to compare rows, identifying patterns between samples with similar feature distributions. For instance, individuals with comparable BMI and smoking habits may exhibit analogous life expectancy trends. By attending to other rows in the dataset, the model enhances its prediction accuracy and robustness by incorporating insights from similar data points.

The training pipeline, detailed in **Fig 6(b)**, incorporates a two-stage process involving pre-training and fine-tuning. Pre-training employs self-supervised learning techniques, such as contrastive learning and denoising objectives, to initialize the model with a robust understanding of general patterns in the data. Techniques like **CutMix** and **Mixup** are applied during this phase to augment the data, combining parts of samples or interpolating embeddings to create new data points. These augmentation strategies, visualized in **Fig 6(b)**, improve the model’s generalization capabilities.

Following pre-training, supervised fine-tuning optimizes the model for specific tasks like regression or classification. During this phase, loss functions such as Mean Squared Error (MSE) or Cross-Entropy Loss (CE) are employed to refine the model’s predictions. By the end of this pipeline, SAINT can effectively process tabular datasets, leveraging both intra-row feature relationships and inter-row sample relationships to achieve superior predictive performance.

Diagrams such as **Fig 6(a)** and **Fig 6(b)** play a pivotal role in understanding the flow and interactions within SAINT’s architecture, from embedding features to leveraging attention blocks and data augmentation for robust training.

#### 3.3.2.1 Example Application of SAINT

To illustrate how **SAINT** operates on tabular data, consider a dataset predicting life expectancy based on features like BMI, smoking habits, and alcohol consumption. Each row represents a sample, with columns capturing numerical and categorical features. For this example, we analyze the steps as visualized in **Fig 6(a)** and **Fig 6(b)** to understand the transformation and learning process.

**Input Data:** The dataset is structured as follows:

Each row represents a sample, with BMI, smoking, and alcohol consumption as the features, and life expectancy as the target variable.

**Embedding Features:** Each feature is transformed into a dense vector representation using the **embedding layer**, depicted in **Fig 6(b)**. Numerical features such as BMI and alcohol consumption are processed through fully connected layers, while categorical features are embedded into vectors. For instance:

BMI of 23.5 might be represented as [0.2, 0.8, -0.1, 0.5].

Smoking of 15 might translate to [0.3, 0.1, 0.7, -0.4].

Alcohol of 5 might result in [-0.2, 0.6, 0.9, 0.1].

These feature embeddings are combined for each row, creating a comprehensive representation of the sample.

**Attention Mechanisms:** The row embeddings are processed through **self-attention blocks** and **intersample attention blocks**, as shown in **Fig 6(a)**.

1. **Self-Attention**: Captures relationships between features within the same row. For example, SAINT learns how BMI and smoking habits interact to influence life expectancy. The multi-head attention mechanism allows the model to explore various combinations of feature relationships simultaneously.
2. **Intersample Attention**: Identifies patterns across different rows. For example, if two individuals have similar BMI and smoking profiles, their life expectancy trends may also align. This mechanism allows the model to integrate information from similar samples, enriching its predictions.

**Data Augmentation:** To improve robustness and generalization, **CutMix** and **Mixup** are applied during pre-training, as seen in **Fig 6(b)**. These techniques augment the data by:

1. **CutMix**: Combining parts of one sample with another, such as taking BMI from one row and smoking habits from another.
2. **Mixup**: Blending embeddings of two rows to create new samples in the latent space.

**Training and Prediction:** Training proceeds in two stages:

1. **Pre-Training**: Using self-supervised techniques like contrastive loss and denoising objectives, SAINT learns general patterns in the dataset.
2. **Fine-Tuning**: The model is optimized for the target task, predicting life expectancy using labeled data. Loss functions like MSE refine the predictions.

Once trained, SAINT predicts the life expectancy for new samples by embedding the input features, passing them through the attention blocks, and generating predictions from the final layer.

**In conclusion**, by employing both machine learning and deep learning approaches, this research seeks to provide a comprehensive evaluation of the methodologies best suited for predicting life expectancy.

# 4. Evaluation

Evaluation is a critical component of this study, ensuring that the models’ predictions are both accurate and reliable. The XGBoost and SAINT models are evaluated using metrics commonly employed in regression tasks, with detailed analysis tailored to each model’s unique characteristics. The metrics used are **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R-squared (R²)**, as both models demonstrate strong performance using these measures. Below, we detail the evaluation approach for each model and provide an in-depth explanation of the metrics.

## 4.1. Comparative Evaluation of XGBoost and SAINT

To comprehensively assess the performance of XGBoost and SAINT, we compare their evaluation strategies and unique strengths in handling regression tasks. The table below highlights the differences and similarities in their approaches, emphasizing how each model leverages specific techniques to optimize predictions and interpret results.

**Table 3: Comparative Analysis of XGBoost and SAINT Evaluation Approaches**

|  |  |  |
| --- | --- | --- |
| Aspect | XGBoost | SAINT |
| Cross-Validation | Evaluated using cross-validation, dividing the dataset into multiple folds to iteratively train and validate the model. This reduces overfitting and ensures robust generalization. | Employs cross-validation with similar regression metrics but focuses on monitoring training dynamics, such as loss over epochs, to ensure convergence. |
| Feature Importance | Utilizes feature importance analysis to interpret the contribution of each independent variable to the predictions. | Leverages attention maps to identify influential features and interactions between samples, enhancing interpretability. |
| Hyperparameter Tuning | Hyperparameters such as learning rate, maximum depth, and boosting rounds are optimized using grid search or Bayesian optimization | Implements early stopping to prevent overfitting when validation loss stops improving, ensuring efficient training. |
| Strengths | Demonstrates high efficiency in handling tabular data and strong capability in modeling non-linear relationships. | Excels at capturing complex feature interactions through self-attention mechanisms, offering deeper insights into the relationships within the data. |

This comparison underscores the complementary strengths of XGBoost and SAINT. While XGBoost focuses on computational efficiency and robust modeling of non-linear relationships, SAINT’s advanced attention mechanisms provide an added layer of interpretability and adaptability to complex datasets. Together, they offer a diverse toolkit for solving regression problems effectively.

## 4.2. Metrics in Detail

**Mean Absolute Error (MAE) -** MAE measures the average magnitude of errors between the predicted and actual values, regardless of their direction. It is a straightforward metric that provides a clear understanding of how far predictions deviate from the true values on average. MAE treats all errors equally, making it robust against the influence of large outliers. This simplicity makes it a reliable choice for comparing models, especially when a balanced view of prediction accuracy is desired. The formula used **,** where is the actual value, is the predicted value, and is the number of observations. The absolute value ensures that the errors are not canceled out by their direction.

**Root Mean Squared Error (RMSE) -** RMSE measures the square root of the average squared differences between the predicted and actual values. Unlike MAE, RMSE gives more weight to larger errors, penalizing them more heavily. This makes RMSE particularly useful when large prediction errors are more concerning than small ones. RMSE provides an intuitive sense of error magnitude, as it is expressed in the same unit as the target variable, making it easy to interpret. The formula used is  , where and represents the actual and predicted values, respectively, and is the total number of observations. The squaring amplifies larger errors, while the square root brings the result back to the same scale as the original data.

**R-squared (R²) -** R² quantifies the proportion of variance in the dependent variable that is explained by the independent variables in the model. It provides a measure of how well the model fits the data, with values ranging from 0 to 1. An R² value close to 1 indicates that the model explains most of the variability in the target variable, while a value near 0 suggests limited explanatory power. R² is particularly useful for comparing models to see which one better captures the underlying patterns in the data. The formula used is , where represents the actual values, the predicted values, and the mean of the actual values. The numerator calculates the residual sum of squares (unexplained variance), while the denominator calculates the total sum of squares (total variance). The resulting value represents the proportion of variance explained by the model, with higher values indicating better performance.

These metrics together provide a comprehensive evaluation framework. While MAE offers a straightforward view of average errors, RMSE highlights significant deviations, and R² indicates the overall explanatory power of the model. Both XGBoost and SAINT are rigorously evaluated using these metrics, providing comprehensive insights into their predictive capabilities. XGBoost excels in efficiency and feature importance analysis, while SAINT demonstrates its strength in capturing complex interactions through its attention mechanisms. The use of these metrics ensures that the models’ predictions are not only accurate but also interpretable and actionable.

# 5. Project Architecture and Deployment

# 

**Fig. 7: Project Architecture and Deployment Workflow**

This project is designed to predict life expectancy based on user-provided information through an interactive web application. Below is a detailed overview of the system architecture, components, and their roles.

**Frontend:**

**Next.js [27]** with **React[29]** and **TypeScript [30]** are used to create an interactive and maintainable web application with fast rendering and dynamic data handling.

**Tailwind CSS [28]** : Simplifies styling with utility-first classes, enabling quick development of responsive and visually appealing designs.

**Backend**

**FastAPI**: A modern Python framework for building high-performance APIs. It supports asynchronous request handling, validates input, and auto-generates API documentation, making development and debugging easier [25].

**Machine Learning:**

**PyTorch**: Used for implementing the SAINT model, specifically designed for complex tabular data tasks [13]. and will also be used for implementing GANs to generate synthetic, realistic data based on our average data as described in section 3.2.1 .

**XGBoost**: A library for gradient boosting, included to handle tabular data problems where tree-based models perform better than deep learning.[18]

**Pandas**: Handles data preprocessing and manipulation and preparing user inputs for ML models.

**Why This Architecture?**

Next.js and React provide a robust structure for building dynamic, interactive UIs, while TypeScript ensures code maintainability by reducing runtime errors. Tailwind CSS accelerates styling workflows.

FastAPI offers high performance, scalability, and ease of use for integrating machine learning models. TensorFlow and PyTorch provide flexibility and reliability for ML tasks, and Pandas efficiently preprocesses data.

This architecture ensures fast response times, robust data handling, and smooth integration between the frontend and backend.

**Overall Process**

This project predicts life expectancy based on user input through the following steps:

1. User Input: Users submit data via a form on the frontend.
2. Data Transmission: The frontend sends validated data to the backend API.
3. ML Prediction: The backend processes the data, applies preprocessing, and uses machine learning models to predict life expectancy.
4. Response: The backend sends the prediction back to the frontend, where it is displayed to the user interactively and intuitively.

# 6. Expected Result

**Synthetic Data Generation**: Utilize a Generative Adversarial Network (GAN) to build synthetic, realistic datasets for life expectancy prediction. This synthetic data will be leveraged for further studies and experimentation to enhance model performance, handle imbalanced data scenarios, and improve the robustness of predictive insights.

**Model Evaluation:** Evaluate the XGBoost machine learning model, alongside the deep learning model SAINT, to analyze the dataset comprehensively. The objective is to select the most accurate and reliable model for predicting LE, based on performance metrics such as RMSE, MAE, and R², as well as generalizability across validation and test datasets.

**Accuracy Improvement:** Strive to surpass the 89% accuracy benchmark reported in previous studies [21]. This will involve developing a dynamic and adaptable predictive model capable of handling various sets of indicators with high precision. The focus is on optimizing the chosen models to improve both accuracy and robustness, ensuring reliable life expectancy predictions across diverse data scenarios.

# References:

**[1] Lipesa, B. A., Okango, E., Omolo, B. O., & Omondi, E. O. (2023). An application of a supervised machine learning model for predicting life expectancy. *SN Applied Sciences, 5*, Article 189.** [**https://doi.org/10.1007/s42452-023-05404-w**](https://doi.org/10.1007/s42452-023-05404-w)

**[2] Shwartz-Ziv, R., & Armon, A. (2021). Tabular data: Deep learning is not all you need. *arXiv.*** [**https://arxiv.org/pdf/2106.03253**](https://arxiv.org/pdf/2106.03253)

**[3] Stokanović, S., Đukić, D., & Miljković, N. (2023). Robustness of XGBoost algorithm to missing features for binary classification of medical data. *IEEE Xplore.*** [**https://ieeexplore.ieee.org/document/10495929**](https://ieeexplore.ieee.org/document/10495929)

**[4] Wang, H., Naghavi, M., Allen, C., Barber, R. M., Bhutta, Z. A., Carter, A., … & Murray, C. J. L. (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the Global Burden of Disease Study 2015. *The Lancet, 388*(10053), 1459–1544.** [**https://doi.org/10.1016/S0140-6736(16)30175-1**](https://doi.org/10.1016/S0140-6736(16)30175-1)

**[5] O’Donovan, G., Petermann-Rocha, F., Ferrari, G., Lee, I.-M., Hamer, M., Stamatakis, E., Sarmiento, O. L., Ibáñez, A., & Lopez-Jaramillo, P. (2024). Associations of the ‘weekend warrior’ physical activity pattern with all-cause, cardiovascular disease and cancer mortality: The Mexico City Prospective Study. *British Journal of Sports Medicine, 58*(7), 359–367.** [**https://bjsm.bmj.com/content/58/7/359**](https://bjsm.bmj.com/content/58/7/359)

**[6] Rentería, E., Jha, P., Forman, D., & Soerjomataram, I. (2016). The impact of cigarette smoking on life expectancy between 1980 and 2010: A global perspective. *Tobacco Control, 25*(5), 551–557.** [**https://doi.org/10.1136/tobaccocontrol-2015-052265**](https://doi.org/10.1136/tobaccocontrol-2015-052265)

**[7] Wood, A. M., Kaptoge, S., Butterworth, A. S., Willeit, P., Warnakula, S., Bolton, T., … & Danesh, J. (2018). Risk thresholds for alcohol consumption: Combined analysis of individual-participant data for 599,912 current drinkers in 83 prospective studies. *The Lancet, 391*(10129), 1513–1523.** [**https://doi.org/10.1016/S0140-6736(18)30134-X**](https://doi.org/10.1016/S0140-6736(18)30134-X)

**[8] Xu, L., & Veeramachaneni, K. (2018). Synthesizing Tabular Data Using Generative Adversarial Networks. *arXiv preprint arXiv:1811.11264*. Retrieved from** [**https://arxiv.org/abs/1811.11264**](https://arxiv.org/abs/1811.11264)

**[9] Mathur, N. (2024, January 26). The importance of education in improving life expectancy. *News-Medical*. Retrieved from** [**https://www.news-medical.net/news/20240126/The-importance-of-education-in-improving-life-expectancy.aspx**](https://www.news-medical.net/news/20240126/The-importance-of-education-in-improving-life-expectancy.aspx)

**[10] Crimmins, E. M. (2015). Lifespan and healthspan: Past, present, and promise. *The Gerontologist, 55*(6), 901–911.** [**https://doi.org/10.1093/geront/gnv130**](https://doi.org/10.1093/geront/gnv130)

**[11] Georgiev, V., Hadzhikoleva, S., & Hadzhikolev, E. (2024). Impact of global country indicators on life expectancy. *Contemporary Scientific and Industrial Research, 1*(1), Article 11.** [**https://doi.org/10.70862/CSIR.2024.0101-04**](https://doi.org/10.70862/CSIR.2024.0101-04)

**[12] Kang, J. J., & Adibi, S. (2018). Systematic predictive analysis of personalized life expectancy using smart devices. *Technologies, 6*(3), Article 74.** [**https://doi.org/10.3390/technologies6030074**](https://doi.org/10.3390/technologies6030074)

**[13] Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., & Goldstein, T. (2021). SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv.*** [**https://arxiv.org/pdf/2106.01342**](https://arxiv.org/pdf/2106.01342)

**[14] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.** [**https://dl.acm.org/doi/pdf/10.1145/2939672.2939785**](https://dl.acm.org/doi/pdf/10.1145/2939672.2939785)

**[15] Little, C., Elliot, M., Allmendinger, R., & Samani, S. S. (2021). Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study. *arXiv preprint arXiv:2112.01925*. Retrieved from** [**https://arxiv.org/abs/2112.01925**](https://arxiv.org/abs/2112.01925)

**[16]** [**Generative Adversarial Networks**](https://medium.com/@marcodelpra/generative-adversarial-networks-dba10e1b4424)

**[17]** [**Boosting Methods**](https://medium.com/@senihberkay/boosting-methods-57b74300235b)

**[18]** [**https://xgboost.readthedocs.io/en/stable/install.html**](https://xgboost.readthedocs.io/en/stable/install.html)

**[19]** [**Generative Adversarial Networks (GANs): Revolutionizing Deep Learning with Synthetic Data…**](https://medium.com/@siddharthapramanik771/generative-adversarial-networks-gans-revolutionizing-deep-learning-with-synthetic-data-9559fd248a99)

**[20]** [**Don’t die wondering: apps may soon be able to predict your life expectancy, but do you want to know?**](https://theconversation.com/dont-die-wondering-apps-may-soon-be-able-to-predict-your-life-expectancy-but-do-you-want-to-know-129068)

**[21]** [**What Really Drives Higher Life Expectancy.” Towards Data Science, 2021,**](https://towardsdatascience.com/what-really-drives-higher-life-expectancy-e1c1ec22f6e1)

**[22] dataset :** [**https://gist.github.com/aishwarya8615/89d9f36fc014dea62487f7347864d16a**](https://gist.github.com/aishwarya8615/89d9f36fc014dea62487f7347864d16a)

**[23] dataset:** [**Life Expectancy (WHO). Kaggle**](https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who)

**[24]** [**https://medium.com/@stronganalytics/deep-learning-for-tabular-data-an-overview-3a742d72246b**](https://medium.com/@stronganalytics/deep-learning-for-tabular-data-an-overview-3a742d72246b)

**[25]** [**https://fastapi.tiangolo.com/**](https://fastapi.tiangolo.com/)

**[26]** [**https://www.kaggle.com/code/alexisbcook/xgboost**](https://www.kaggle.com/code/alexisbcook/xgboost)

**[27]** [**https://nextjs.org/**](https://nextjs.org/)

**[28]** [**https://tailwindcss.com/**](https://tailwindcss.com/)

**[29]** [**https://react.dev/**](https://react.dev/)

**[30]** [**https://www.typescriptlang.org/**](https://www.typescriptlang.org/)