Politécnico de Tomar
Polytechnic University

Escola Superior de Tecnologia de Tomar

Time-series GAN for Longitudinal Tabular Data: Validating Synthetic Data Utility in Multi-Class Health Prediction

Master's Dissertation

Paulo Humberto Saragga de Melo Banha

Master of Science in Business Intelligence and Analytics
Deep Learning

Tomar, December 2025

Escola Superior de Tecnologia de Tomar

Time-series GAN for Longitudinal Tabular Data: Validating Synthetic Data Utility in Multi-Class Health Prediction

Dissertação de Mestrado

Paulo Humberto Saragga de Melo Banha

Orientada por:
Professora Doutora Sandra Maria Gonçalves Vilas Boas Jardim - Instituto Politécnico de Tomar

Professor Doutor Rolando Lúcio Germano Miragaia - Instituto Politécnico de Tomar

Júri
Professora Coordenadora Sandra Maria Gonçalves de Vilas Boas Jardim
Professor Coordenador José Carlos Bregieiro Ribeiro (Arguente)
Professor Adjunto João Manuel Mourão Patrício

Dissertação apresentada ao Instituto Politécnico de Tomar para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Analítica e Inteligência Organizacional

Resumo

Este trabalho aborda as limitações fundamentais da fragilidade dos modelos de previsão clínica longitudinal e os enviesamentos causados pela escassez de dados e pelo desequilíbrio entre classes, inerentes aos registos sequenciais de doentes, recorrendo especificamente a um conjunto de dados limitado a apenas 117 registos. Apresenta uma metodologia integrada e uma validação empírica rigorosa que utiliza uma Rede Adversária Generativa para Séries Temporais (TimeGAN) para criar uma população sintética de doentes de elevada fidelidade, mitigando esta restrição extrema de dados. A implementação do TimeGAN preservou com sucesso a fidelidade estrutural e temporal, confirmada através de análises estatísticas (teste de Kolmogorov-Smirnov) e estruturais (projecção por PCA), um passo crucial para o treino de classificadores multi-classe robustos para a previsão de resultados em saúde.

As trajetórias longitudinais sintéticas geradas, melhoradas por uma estratégia de balanceamento de classes imperfeita, mas benéfica, constituíram a base de um estudo comparativo que contrastou de forma rigorosa o desempenho de generalização de duas famílias de modelos: o Multi-Layer Perceptron (MLP) estático versus a rede sequencial Long Short-Term Memory (LSTM). Avaliados exclusivamente num conjunto de teste de doentes reais não vistos e reservado (T1), os resultados experimentais demonstraram que o modelo de base treinado apenas com dados reais limitados falhou de forma catastrófica (Macro F1-Score: 15,17%), validando empiricamente a necessidade de aumento de dados sintéticos através do TimeGAN. Em contraste, o modelo LSTM (M4), treinado com sequências sintéticas balanceadas, alcançou a melhor generalização em contexto real (Weighted F1-Score: 74,67%); os modelos MLP sobre ajustaram significativamente às características sintéticas estáticas, resultando numa generalização substancialmente inferior e confirmando a sua incapacidade de explorar a causalidade sequencial.

Estes resultados validam a utilidade do TimeGAN na geração de dados tabulares longitudinais sintéticos fiáveis e estabelecem uma dependência arquitetónica crucial: a previsão multi-classe robusta requer não só um aumento de dados essencial, mas também um modelo sensível à sequência (LSTM) para explorar plenamente a fidelidade temporal preservada pelo enquadramento TimeGAN. Esta investigação oferece uma metodologia validada, de ponta a ponta, para acelerar a investigação em domínios clínicos de elevado valor e com dados limitados.

Abstract

This work addresses the fundamental limitations of longitudinal clinical prediction model fragility and dangerous bias caused by severe data scarcity and class imbalance inherent in sequential patient records, specifically utilizing a dataset limited to just 117 records. It presents an integrated methodology and rigorous empirical validation utilizing a Time-series Generative Adversarial Network (TimeGAN) to create a high-fidelity synthetic patient population, mitigating this extreme data constraint. The TimeGAN implementation successfully preserved structural and temporal fidelity, confirmed through statistical (Kolmogorov-Smirnov test) and structural (PCA projection) analysis, a crucial step for training robust multi-class classifiers for health outcome forecasting.

The generated synthetic longitudinal trajectories, enhanced by an imperfect but beneficial class balancing strategy, formed the basis for a comparative study rigorously contrasting the generalization performance of two model families: the static Multi-Layer Perceptron (MLP) versus the sequential Long Short-Term Memory (LSTM) network. Evaluated solely on an unseen, reserved real patient test set (T1), experimental results demonstrated that the baseline model trained exclusively on limited real data failed catastrophically (Macro F1-Score: 15.17%), empirically validating the necessity of synthetic data augmentation using TimeGAN. In contrast, the LSTM model (M4) trained on balanced synthetic sequences achieved the highest real-world generalization (Weighted F1-Score: 74.67%); MLP models significantly overfit static synthetic features, resulting in significantly reduced generalization and confirming their inability to utilize sequential causality.

These findings validate TimeGAN's utility in generating reliable synthetic longitudinal tabular data and establish a crucial architectural dependency: robust multi-class prediction requires not only essential data augmentation but also a sequence-aware model (LSTM) to fully leverage the temporal fidelity preserved by the TimeGAN framework. This research offers a validated, end-to-end methodology for accelerating research in high-value, data-limited clinical domains.

Acknowledgements

First and foremost, I would like to thank my supervisors, Professor Doutor Rolando Miragaia and Professora Doutora Sandra Jardim. Their patient guidance, insightful critiques, and unwavering encouragement provided the direction necessary to shape and execute this research. Their expertise was essential in navigating the technical challenges of the project.

I am also grateful to Engenheiro Carlos Mora, for his valuable insights and the suggestion for this study.

This work was conducted under the formal authority of the master's in business Intelligence and Analytics program at the Instituto Politécnico de Tomar, to which I am thankful for providing the foundational knowledge and resources.

This academic journey would not have been possible without persistent love and support from my family. To my father, thank you for your patience, true belief in me, and constant encouragement; it finally paid off. To my beloved mother, who sadly left before I could finalize this academic step. This achievement is dedicated to her memory.

Finally, to my future spouse and our future Luena, your patience, sacrifices, and unconditional love provided the stable foundation and the future to look forward to, making all the effort worthwhile.

INDEX

Table of Contents

List of Acronyms

Acronym	Full Term	Context / Notes
CDF	Cumulative Distribution Function	Used for K-S test visualization and distribution fidelity.
D	Discriminator	Component of TimeGAN architecture (Distinguishes real vs. synthetic sequences).
DAAE	Dual Adversarial Autoencoder	Synthetic data generation methodology (Related Work).
E	Embedder	Component of TimeGAN architecture (Maps real data to latent space).
EHR	Electronic Health Record(s)	General clinical data context (Source data)
EHR-M-GAN	GANs for synthesizing Mixed-type longitudinal EHR data	Related Work on generating heterogeneous EHRs.
G	Generator	Component of TimeGAN architecture (Generates sequences in latent space).
GAN	Generative Adversarial Network	General family of generative models
GRU	Gated Recurrent Unit	Recurrent unit utilized in TimeGAN architecture.
K-S	Kolmogorov-Smirnov	Statistical test for marginal distribution fidelity
L_D	Discriminator Loss	TimeGAN training objective.
L_G	Generator Loss	TimeGAN training objective.
L_R	Reconstruction Loss	TimeGAN training objective (Embedder/Recovery loss).
L_S	Supervised Loss	TimeGAN training objective for temporal coherence
L_U	Unsupervised Loss	TimeGAN training objective (Adversarial loss in latent space).
LSTM	Long Short-Term Memory	Sequential classification model (M4)
M1	MLP Real Data Model	Experimental Model Configuration (Baseline).
M2	MLP Synthetic Data Imbalanced Model	Experimental Model Configurations
M3	MLP Synthetic Data Balanced Model	Experimental Model Configurations
M4	LSTM Synthetic Data Balanced Model	Experimental Model Configuration (Core Test).

MGU	Minimal Gate Unit	Deep Learning Architectures (Recurrent Unit, Related Work).
MLP	Multi-Layer Perceptron	Static classification model (M1, M2, M3) architecture.
NaN	Not a Number(s)	Missing value representation in data cleaning.
PC	**P**rincipal **C**omponent	Individual dimension created by PCA.
PCA	Principal Component Analysis	Dimensionality reduction technique
PDF	Probability Density Function	Used for distribution visualization
PFI	Permutation Feature Importance	Explainability technique discussed in Future Work.
PK	Primary Key	Used for imputation strategy (e.g., part_id)
R	Recovery	Component of TimeGAN architecture (Reconstructs data from latent space).
RGAN	Recurrent Generative Adversarial Network	Related Work.
RCGAN	Recurrent Conditional Generative Adversarial Network	Related Work.
RNN	Recurrent Neural Network	General term for recurrent architectures
SC-GAN	Sequentially Coupled Generative Adversarial Network	Related Work.
SDG	Synthetic Data Generation	General process.
SHAP	SHapley Additive exPlanations	Explainability technique discussed in Future Work.
T	Time / Time-steps	Fixed sequence length for patient visits (T=4).
T1	Test Set 1	All Real Data Test Set (Used for TSTR evaluation).
T2	Test Set 2	Synthetic Never-Seen Test Set.
T3	Test Set 3	Combined Test Set (Synthetic + Real).
TimeGAN	Time-series Generative Adversarial Network	Core generative framework
TP	True Positive	Metric used in Confusion Matrix analysis.
TRTR	Train Real, Test Real	General experimental methodology term.

TSTR	Train Synthetic, Test Real	Gold standard evaluation methodology for synthetic data utility.
t-SNE	t-distributed Stochastic Neighbor Embedding	Dimensionality reduction technique for visualization.
VPM	Virtual Patient Model	Concept derived from synthetic data
XAI	Explainable Artificial Intelligence	General field of research.

List of Figures

List of Tables

1 Introduction

The rise of Electronic Health Records (EHRs) has provided researchers with vast repositories of patient data. However, translating this wealth of information into reliable predictive models is often complicated by two critical and mutually reinforcing challenges: data scarcity and severe class imbalance. In the context of specialized, high-stakes clinical predictions such as identifying distinct multi-class stages of patient health decline, the number of available real-world patient trajectories is frequently limited. This limitation results from the cost and sensitivity associated with collecting longitudinal monitoring data. Furthermore, the few available examples are typically concentrated in common, low risk health states (the majority class), leaving rare or critical high-risk states underrepresented. Models trained on such constrained and skewed datasets invariably exhibit poor generalization and a harmful bias toward common outcomes, rendering them statistically unreliable and clinically unsafe for objective decision support. This study utilizes a small, severely imbalanced, longitudinal patient dataset to directly address the challenge of creating a robust, generalizable multi-class health prediction model.

This dissertation addresses this challenge directly by integrating state-of-the-art deep learning techniques. Our approach leverages Time-series Generative Adversarial Networks (TimeGAN) for the critical task of synthetic data generation and sequential neural networks for subsequent predictive modeling. The core motivation is to rigorously determine if a specialized generative model can create a high-fidelity Virtual Patient Model (VPM) reliable enough to train classifiers for robust, multi-class prediction, thereby advancing and democratizing data-intensive research in limited data medical domains. The research is specifically guided by two primary goals: first, to validate the reliability and fidelity of synthetic patient trajectories generated by TimeGAN for predicting a multi-class outcome; and second, to identify the optimal deep learning architecture necessary to fully exploit the inherent temporal features and sequential causality embedded within that synthetic data, ultimately leading to superior generalization on real-world patient data.

To achieve the comprehensive solution required by this research goal, four specific objectives were pursued and met. The investigation first required a successful demonstration of the generative intervention through the TimeGAN Implementation and Fidelity objective, which mandated the training of the TimeGAN framework to synthesize patient trajectories that accurately maintain the structural and temporal coherence of the original data. This intervention then required immediate validation under the Synthetic Data Validation (Necessity) objective, ensuring that models trained on this synthetic data significantly outperform the fragile baseline model (M1) that failed due to data scarcity. Essentially, addressing the longitudinal nature of the data required the Architectural Dependency Validation objective, involving a rigorous comparative analysis between a static Multi-Layer Perceptron (MLP) and a sequential Long Short-Term Memory (LSTM) model to empirically identify the superior architecture for utilizing the sequential causality. This structured approach culminated in the Optimal Model Identification objective, which validated the final, best-performing model configuration (M4) for robust health rate prediction in this resource-constrained setting. The successful achievement of these objectives provides a comprehensive solution for overcoming data limitations in sensitive longitudinal clinical datasets.

The remainder of this dissertation is organized as follows:

- **Chapter 2: Background and Related Work** establish the theoretical context, reviewing the challenges of longitudinal data, the principles of Generative Adversarial Networks (GANs) with a focus on TimeGAN, and the role of deep learning architectures in time-series classification.

- **Chapter 3: Methodology** details the data processing steps, the implementation of the TimeGAN framework, the specific creation of the synthetic datasets, and the training and partitioning strategies for the four experimental models (M1 to M4).

- **Chapter 4: Results and Findings** present comprehensive empirical outcomes, including data fidelity evaluations, hyperparameter optimization results, and performance metrics for all experimental models, focusing on the architectural comparison and the final validation results on the real test set.

- **Chapter 5: Discussion** interprets the key findings, evaluates the effectiveness of TimeGAN and the architectural choices against the research objectives, contextualizes the results within the existing literature, and addresses the study's limitations.

- **Chapter 6: Conclusion and Future Work** summarize the main conclusions of the study, outlines the specific contributions to the field, and proposes concrete directions for future research aimed at clinical implementation.

2 Background and Related Works

This chapter provides the foundational literature review necessary to understand and justify the research methodology presented in Chapter 3. Predictive modeling using longitudinal Electronic Health Record (EHR) data is constrained by fundamental challenges related to data structure, volume, and complexity. Research has consistently identified three major barriers to achieve robust clinical prediction: data scarcity, class imbalance, and the inherent sequential nature of patient trajectories. The literature reviewed here first establishes the distinct structural properties of EHR data, differentiating between static (tabular) and sequential (time-series) formats. Following this, the chapter reviews the state-of-the-art in overcoming these limitations, detailing the evolution of Synthetic Data Generation (SDG) from simple augmentation to sophisticated generative models like Time-series Generative Adversarial Network (TimeGAN), which is uniquely designed to preserve temporal fidelity in synthesized sequences. Finally, it reviews the established architectural consensus in deep learning, comparing static classification models (like Multi-Layer Perceptron or MLP) with sequential models (like Long Short-Term Memory networks or LSTMs) to confirm the appropriate predictive tools for longitudinal tasks. Ultimately, this review identifies a critical gap regarding the necessity of architectural alignment when exploiting TimeGAN's unique outputs for real-world generalization.

2.1 EHR Data Structure: Tabular vs. Time-Series

The foundation for clinical prediction models rests upon Electronic Health Record (EHR) data, which is complex and often heterogeneous in its structure [1]. A thorough review of this domain identifies several key opportunities including disease detection, sequential prediction, and concept embedding alongside major challenges such as data privacy, model complexity, and rigorous performance evaluation [1]. The application of deep learning architectures to EHR data is examined, analyzing its use in tasks like multi-label prediction and data augmentation [1]. This review specifically addresses the special challenges inherent in modeling EHR data, notably its high dimensionality and temporal complexity [1].

The EHR can be primarily structured and presented in two formats that commonly coexist within a single patient record: Tabular and Time-Series data. The tabular format typically stores static, non-sequential patient encounter information, such as essential demographic or gender features, often representing a single snapshot in time. While this format is crucial for baseline risk assessment and cross-sectional analysis, it fundamentally fails to capture the dynamic progression of disease or the evolution of the health condition over time.

In contrast, Time-Series Data is fundamentally a record of data points indexed in time order. It is the format used to capture the essence of a patient's medical journey, from fine-grained vital sign fluctuations to large-scale disease progression over multiple visits (longitudinal data). This sequential nature is critical for modeling the clinical trajectories that determine long-term outcomes. Advanced generative models are necessary to create synthetic time-series that preserve sequential integrity and complex, non-linear dependencies within the data. Techniques capable of generating such synthetic trajectories must capture complex structures and temporal dynamics, outperforming existing methods in fidelity and utility [2]. The variables recorded can be classified as Discrete, Categorical, or Continuous.

The inherent challenge in using EHR data for machine learning lies in the heterogeneous nature of the patient record, which demands specialized modeling and analysis techniques capable of handling sequential and non-sequential information simultaneously [1]. EHR data, in the form of multivariate numeric time-series, is inherently high-dimensional, sparse, and irregularly distributed across time [3] [4]. This structural complexity, which includes clinical events that are irregularly distributed over time, poses a significant challenge for standard learning algorithms [4]. Moreover, these time-series often have different lengths and are measured at irregular intervals, as each patient trajectory is unique [4]. The irregularity and sparsity of observations contain valuable semantic information about clinical hypotheses and treatment choices, meaning robust models must leverage the data's structure in both time and event dimensions [3]. Models that naively apply sequence-based methods only across the time dimension risk losing important relationships along the event dimension, thereby limiting their ability to capture dependencies between different types of medical observations.

Modeling this temporality remains an important research question. Some approaches rely on extracting single values from time-series; however, this leads to the loss of potentially valuable sequential information [3]. When transforming the raw EHR data into a tabular format that can be handled by standard machine learning algorithms, the core problem becomes generating features that effectively represent these time-series. Research has focused on methods like temporal abstraction or temporal logic to define patterns that describe temporal relationships among multiple time-series, such as finding that the occurrence of clinical event A precedes a drop of clinical event B. Other methods transform time-stamped data points into symbolic time intervals, discovering frequent interval-related patterns used to induce a classifier [3]. Furthermore, to compare similarities between time-series, approaches like using subsequences of time-series have been proposed, which require finding local, rather than global, similarities for class separation [3]. This study explicitly tackles the added complexity of dealing with heterogeneous time-series of different lengths and irregular intervals.

2.2 Challenges in Synthetic Data Generation

While Synthetic Data Generation (SDG) offers significant advantages for mitigating data scarcity and protecting privacy, its effective application in the clinical domain is constrained by several limitations that must be addressed to ensure utility and reliability, including:

- **Preservation of Temporal Dependencies and Quality:** The principal challenge in synthesizing EHR data is the preservation of temporal dependencies and the overall data quality. Unlike static tabular data, time-series EHRs require generative models to accurately capture long-range correlations across multiple time steps and the conditional distributions between sequential events [5]. Inadequate temporal preservation can degrade the utility of synthetic data, leading to classifiers that fail to capture the dynamic progression of disease. Furthermore, the synthetic data itself must be statistically robust; biases or inaccuracies introduced during generation, such as generating noise or overlapping instances, can propagate to the downstream predictive model, leading to unreliable and potentially harmful clinical outcomes.

- **Generalization and Tailoring (Fidelity vs. Utility):** A persistent constraint in SDG is navigating the trade-off between fidelity (how closely the synthetic data matches the statistical properties of the real data) and utility (how useful the synthetic data is for a

downstream task, like prediction). There is a risk that the augmentation process may be too tailored to the specific characteristics of the real training data, limiting the model robustness and generalization ability in novel clinical situations [6]. In contrast, if the synthetic data deviates too much from real distribution, the model trained on it will not generalize well to the real-world test data. This tension necessitates the use of comprehensive evaluation frameworks that go beyond simple statistical comparisons to assess utility, such as the Train Synthetic, Test Real (TSTR) method [7] and clinically relevant validation [8].

- **Data Privacy and Computational Constraints:** In sensitive clinical domains, adherence to data privacy standards and ethical constraints poses a major limitation for SDG adoption [9]. While GANs are frequently used to mitigate disclosure risk, robust privacy metrics [10] and compliance techniques like Differential Privacy [11] are essential but often add complexity. Furthermore, advanced iterative models like TimeGAN [12] and its extensions [5] are computationally intensive and time-consuming. The complex training scheme, involving multiple recurrent neural network (RNN) components (Embedder, Generator, Discriminator) and a multi-part loss function, can require significant resources, limiting the scalability of these techniques across large or resource-constrained healthcare systems.

These challenges are significantly magnified when dealing with the highly specific structural and temporal irregularities found in longitudinal patient records, which are detailed in the following section.

2.3 The Challenge of Longitudinal Data in Health Prediction

Predictive modeling in clinical settings faces significant challenges due to the unique characteristics of longitudinal data, necessitating specialized approaches for robust and reliable outcomes [13]. Longitudinal studies offer a unique opportunity to characterize individual human lifespan trajectories, providing a major advantage over cross-sectional studies by unlocking information on an individual's development and aging trajectory [14].

Clinical data is, by definition, longitudinal, resulting in the creation of complex sequential and temporal interdependencies between features. For any predictive model to be successful, it requires a dataset that maintains the temporal fidelity of realistic patient progression. A critical challenge for maintaining model fidelity is the pervasive issue of missing data and irregular sampling inherent in real-world Electronic Health Records (EHRs). This data irregularity requires complex imputation techniques. Recognizing this, recent state-of-the-art research has focused on unifying time-series imputation and generation within a single framework to simultaneously tackle both the missing data problem and data scarcity [15]. This combined approach aims to create highly accurate, dense synthetic time-series data where the fidelity is preserved across both measured and imputed time steps. Longitudinal EHRs contain crucial patient trajectories, making the synthesis of such time-series vital for enabling new clinical applications related to the status of disease progression [15]. The challenges observed in EHR modeling are mirrored in other complex biomechanical domains, such as Gait stability assessment, which is similarly challenged by limited data availability and measurement complexity [16].

A fundamental justification for using longitudinal approaches is that inter-individual variation (differences between people) is not the same as intra-individual variation (changes within a single person over time) [14]. The key point is that individuals do not always follow the group average trajectory. In fact, it can be proven that inter-individual variation observed across a group of different individuals at one time point is not methodologically interchangeable with measuring change within a single individual over a period of time [14]. Since processes like learning, development, and disease progression are dynamic and unique, cross-sectional group averages cannot reliably replace the individual trajectory data [14]. The longitudinal setup is therefore essential because it allows every individual to act as their own control, detecting subtle, meaningful changes in patient state over time [14].

To effectively model these complex patient trajectories, a range of deep learning architectures have been developed for clinical decision support, including applications such as medication recommendation, health risk prediction, and disease progression understanding [17]. Models have evolved from Recurrent Neural Networks (RNNs) used for tasks like medication recommendation and health risk prediction via time-aware attention mechanisms to more sophisticated sequential models [18]. The emergence of Attention-based Transformer models (e.g., HiTANet) [19], and deep state-space models (like the Attentive Deep Markov Model) [20] offers superior capabilities for handling intricate sequential data and tracing patients latent status [17]. Moreover, Transformer architecture is critical not only for high-performance prediction but for modeling the continuous evolution of health trajectories [21]. Advanced approaches are now focused on generating updated predictions at every time point (e.g., via causal attention masks) to provide continuous insight into early predictors and aid in the early detection of changes in a person's health status [21].

A major practical barrier is the severe issue of data scarcity and class imbalance inherent in niche clinical datasets [22]. When studying rare outcomes, minority classes are highly underrepresented [23], leading to models that are biased toward predicting the majority class. The integration of Synthetic Data Generation (SDG) techniques [13] has emerged as a promising approach to improve the accuracy and robustness of pattern modeling by addressing data scarcity, enabling dataset augmentation, and mitigating privacy concerns [24]. The longitudinal setup itself poses additional challenges, including data quality issues such as missing data and attrition rates which can introduce bias toward healthier individuals in later waves as well as the high economic cost and time required for multiple waves of data acquisition [14].

2.4 The Evolution of Synthetic Data Generation for EHRs

The core of data augmentation research has moved beyond simple feature generation to handling temporal dynamics [15]. The landscape of GAN-based synthesis for EHRs has been comprehensively reviewed [25]. While generating tabular EHR data is useful [24], it fails to capture the essential dynamics and changes recorded in time-series data. To address this, specialized frameworks have been developed including the SynTEG Framework [22], , which focuses on generating timestamped diagnostic events using a self-attention layer and Wasserstein GAN [23], the Dual Adversarial Autoencoder (DAAE) [26], which synthesizes sequences of set-valued medical records; Recurrent GANs (RGAN and RCGAN) [27], which use Long Short-Term Memory (LSTM) units to handle continuous time-series EHRs; and

Sequentially Coupled GAN (SC-GAN) [28] [29], , developed to coordinate the generation of mutually influential data types.

The heterogeneous nature of EHRs requires generative models to capture complex dependencies across varied data types simultaneously [25]. This has led to the development of models for:

- **Tabular Heterogeneity**: Models have been developed to account for constraints and preserve relationships by incorporating a penalization term for constraint violation during GAN training [30].
- **Longitudinal Mixed-Type Data:** GANs for synthesizing Mixed-type longitudinal EHR data (EHR-M-GAN) [31] utilize a dual variational autoencoder to generate a shared latent space representation and a sequentially coupled generator implemented with bilateral LSTM to capture temporal correlations.
- **Constraint-Based SDG for Interpretability:** Research employing constraint-based SDG has achieved exceptional data fidelity and enhances model interpretability by realigning feature attribution [16].

The generation of synthetic EHRs is most accomplished through Generative Adversarial Networks (GANs) [2] [27] [28] [29] and Variational Autoencoders (VAEs) [24] [26]. The recent advancements in these generative AI models for synthetic longitudinal data have shifted focus toward generating richer, more complex patient histories [15]. To improve the quality and utility of synthetic time-series data, current research has centered on leveraging the capabilities of both recurrent networks and attention mechanisms [17] [18] [29]. Sophisticated temporal generative models have been introduced specifically to address the challenge of capturing complex sequential dependencies and properties, aiming to improve the fidelity of the data generated compared to simpler recurrent models [12].

A critical component of this field's evolution is the development of robust metrics to evaluate the quality of synthetic EHRs. Synthetic EHRs are only useful if they maintain the statistical and structural properties of real data while preserving patient privacy [32]. A systematic review confirmed that deep learning approaches, particularly GANs (used in 22 out of 28 reviewed papers) [6], are the dominant generation method used for both static and time-series data. The evaluation is classified into three fundamental dimensions: Fidelity, where the resemblance to real data statistics and distributions, assessed via methods like correlation matrices [33] [34] [35], distribution measures [9] [35], and distance metrics [36]; Utility, where the usefulness for downstream tasks, assessed via the Train on Synthetic, Test on Real (TSTR) framework [27] [37] and Dimension-wise Predictions [30] [38]; and Privacy where the measure of information leakage risk, assessed via disclosure metrics [10] [39] and the effect of Differential Privacy [11] [27] [6].

Despite the development of numerous metrics, significant gaps remain, primarily impending the accurate assessment of complex EHRs. The most crucial finding is the temporal evaluation gap: most existing methods fail to account for time dependencies between patient encounters, instead evaluating time-series EHRs as static snapshots [40]. This leads to challenges, as visual analysis (e.g., t-SNE [41]) often removes the time aspect when bundling sequences, and standard correlation heatmaps may show completely different results compared to real

data [6]. Furthermore, a large portion of fidelity measures, such as Kullback–Leibler Divergence (KLD) [35] and Jaccard Similarity [42] [43], are applied in a univariate manner, making it challenging to confirm if the models preserve the crucial multivariate relationships and dependencies that exist in clinical data [10]. Other limitations include the high subjectivity of dimensionality reduction visualizations [44] [41] [45] [46], the lack of standard reporting for parameter choices in distribution measures (e.g., bin selection in histograms) [35], and a notable under-reliance on clinical validation by domain experts [11] [8]. The experimental results confirmed the inconsistency, showing that different metrics can contradict one another (e.g., Jaccard Similarity may indicate high fidelity while Cosine Similarity suggests dissimilarity) and that utility can sometimes be high (e.g., low MAE in sequence prediction) even when statistical fidelity is low [6]. Addressing these limitations requires placing a strong emphasis on developing comprehensive evaluation frameworks and specifically targeting methodologies for evaluating temporal data [6].

2.5 Time-series GAN for Temporal Data Augmentation

One of today's most vital resources is data, yet its accessibility is often restricted due to issues like confidentiality, Personally Identifiable Information compliance, and the high cost of manual annotation [47]. Synthetic data, generated via simulations or algorithms, offers a crucial solution, particularly in fields like pharmaceuticals and healthcare, by mitigating privacy risks, accelerating product testing, and enabling the training of robust machine learning algorithms on diverse, large datasets [47]. While different data types can be synthesized, creating synthetic time-series data is far trickier than tabular data due to the data being dispersed across multiple time sequences, making it challenging to learn long-term features [47]. While many GAN variants target specific aspects of EHR generation, the Time-series GAN (TimeGAN) framework [48] is the state-of-the-art solution chosen, specifically for its superior ability to maintain the fine-grained temporal and cross-feature correlations of continuous longitudinal data.

TimeGAN was designed to overcome the limitations of standard GANs when modeling the inherent temporal complexity of sequence data [5]. It integrates an autoencoder structure (consisting of the Embedder and Recovery components) with adversarial training (Generator and Discriminator) within a latent space, which is critical for enabling the generation of synthetic data that accurately resembles real medical records [49] [47].

The architecture is defined by the interaction of four core RNN components (often Gated Recurrent Units or Long Short-Term Memory): the Embedder (E), Recovery (R), Generator (G), and Discriminator (D). The Embedder maps real data into the latent space, and the Recovery component reconstructs it back, ensuring the latent space preserves structural characteristics. The Generator operates in this latent space, while the Discriminator distinguishes between real and synthetic sequences, also in the latent space [12].

The relationship between these four components is illustrated in Figure 2-1 (Adapted from Yoon et al. [12] [7]).

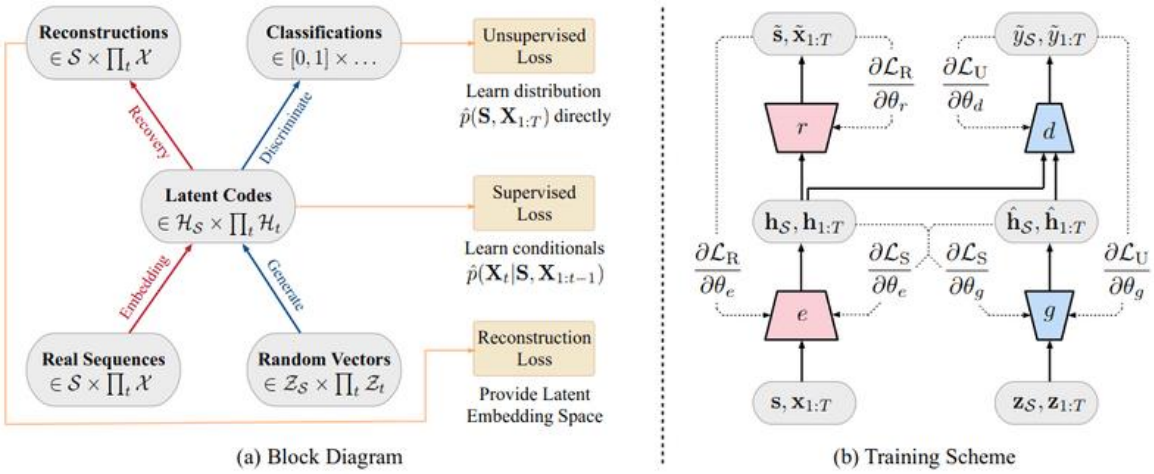


Figure 2-1 TimeGAN architecture (Embedder, Recovery, Generator, Discriminator)

The specialized nature of TimeGAN is codified in its unique, three-part loss function, designed to balance structural, temporal, and adversarial fidelity. The interaction of these three loss components during the training scheme is visualized in Figure 2-2 (Adapted from Yoon et al. [12] [7]).

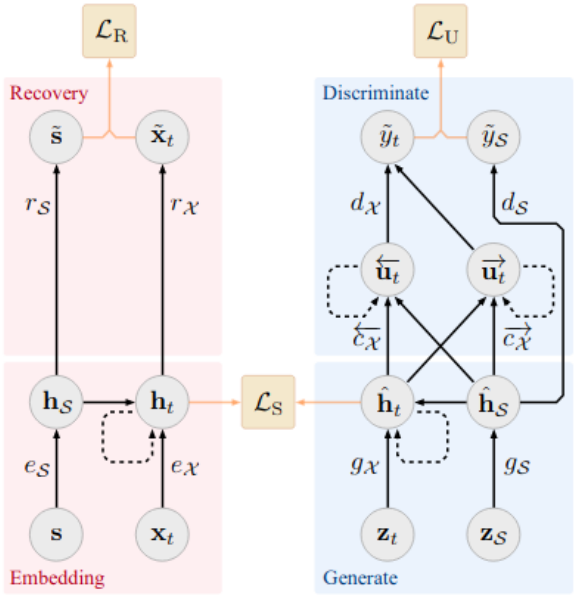


Figure 2-2 TimeGAN loss functions

The model is trained using a specialized three-part loss function based on the TimeGAN framework [12] [7]. This approach is designed to balance adversarial learning with data fidelity by introducing two key non-adversarial objectives into the training process. To ensure high-fidelity synthetic data generation, the following loss components are utilized:

- **Reconstruction Loss (L_R):** This is a key innovation within the framework. It ensures the latent space preserves the structural characteristics of the data by minimizing the difference between the input sequence and the sequence recovered from the latent space. It is calculated using an embedding network Encoder (E) and a recurrent

network Recovery (R) to ensure the latent space effectively captures the static features of the data, providing stability and interpretability.

- **Supervised Loss (L_S):** This represents a crucial temporal loss that forces the Generator to accurately predict the next step in the sequence based on previous steps, preserving the progression dynamics [12]. This explicitly penalizes the Generator if it fails to learn the underlying temporal dynamics and conditional dependencies inherent in the time-series.

- **Unsupervised Loss (L_U):** This is the standard adversarial loss, applied in the latent space, which ensures the synthetic sequences are statistically indistinguishable from real-world observations.

The total generator loss (L_G) and discriminator loss (L_D) are formally defined as a combination of these components. Prediction models for clinical outcomes have been shown to perform significantly better when training data was augmented with high-fidelity GAN-generated timeseries [50].

The total generator loss (L_G) and discriminator loss (L_D) are formally defined as:

$$L_G = L_U(G) + \lambda_S \times L_S(G) + \lambda_R \times L_R(E,R) \text{ and } L_D = L_U(D)$$

In this formulation, (L_D) represents the standard adversarial objective where the Discriminator focuses solely on classifying data as "real" or "fake" (L_U), without relying on reconstruction or supervised components.

For the Generator (L_G), the adversarial goal is balanced by the weighted penalties:

$\lambda_S \times L_S(G)$ addresses poor temporal prediction.

$\lambda_R \times L_R(E,R)$ addresses poor data reconstruction and structure preservation.

The λ_S and λ_R are hyperparameters that must be tuned during model training to weighing the importance of temporal dynamics and reconstruction fidelity against the adversarial objective.

To address the challenge of effectively learning dependencies over longer time steps, an enhancement involves introducing a multi-head self-attention mechanism layer after the Gated Recurrent Unit (GRU) structure in the Recovery module [48]. This mechanism enhances the model's ability to learn dependencies between time steps and selectively weighing key information [48]. The robustness of the GAN framework is validated by its effectiveness for imputing missing data in sequential records with superior accuracy [50]. Furthermore, research confirms that the quantity of generated data must be optimized, not simply maximized, as over-augmentation can lead to a decline in the prediction model's generalization ability on real-world test data [48]. The utility of the generated synthetic data is most rigorously validated using the Train Synthetic Test Real (TSTR) method, which assesses how effectively a predictor model, trained solely on the synthetic data, performs when evaluated on the original real-world data [7].

Despite its advantages, TimeGAN is recognized to have limitations, particularly concerning stability and fully capturing multivariate factors [5] [47]. When dealing with complex multivariate time-series data, TimeGAN was found to be unable to perform adequately, often showing low autocorrelation scores on lengthy sequences and being vulnerable to mode collapse [47]. This necessitates more advanced approaches to properly handle the complex relationships across time and between variables [47]. To address these limitations, the MTS-TGAN model, an extension of TimeGAN, was proposed to tackle the multivariate challenge by using six GRU layers per module and incorporating a feature selector as a pre-processing layer to filter noise and focus on important features [47]. Another critical improvement is the SeriesGAN framework, developed to overcome the issues of TimeGAN's suboptimal convergence and information loss [5]. This framework utilizes two discriminators (one in the latent space and one in the feature space) to enhance fidelity and reduce information loss. It integrates an autoregressive supervisor network using a teacher forcing approach to explicitly model the conditional distributions, strengthening the capture of temporal dynamics. Finally, SeriesGAN employs Least Squares GANs (LSGANs) to mitigate vanishing gradients and includes loss components to explicitly match the Mean and Variance of the real and synthetic data, ensuring consistent, optimal results [5].

Based on a comprehensive review of the generative landscape, the Time-series Generative Adversarial Networks (TimeGAN) framework is the primary solution adopted for this study [12] [51]. While acknowledging the theoretical benefits of advanced models, TimeGAN provides a robust, validated, and efficient methodology that is highly effective for synthesizing sequential data where the core utility lies in capturing accurate temporal dynamics [7]. The selection is motivated by TimeGAN's ability to capture the complex time conditional distributions essential for preserving the clinical meaning of EHR records. Its use of an Embedding network for dimensionality reduction is key to achieving a more stable training process, and its capacity to handle both static attributes and sequential features makes it inherently suitable for clinical records [7]. Furthermore, practical demonstrations confirm that TimeGAN produces synthetic data that achieves an almost perfect visual overlap with real data when analyzed via techniques like PCA and t-SNE [7]. By adopting TimeGAN, this research utilizes a framework that is both academically recognized and practically validated for generating high-fidelity, time-coherent synthetic data.

2.6 Handling Imbalanced Datasets in EHR

The issue of class imbalance is a fundamental challenge when analyzing Electronic Health Records (EHR) and structured medical data, where the minority class (e.g., patients with a specific disease) is significantly underrepresented compared to the majority class (healthy patients). The literature addresses this problem through three major methodological pathways: data-level resampling, learning-level algorithmic adjustments, and the adoption of hybrid or combined techniques [52].

Data-level methods focus on adjusting the distribution of the training data. Oversampling is the most prevalent technique, with variants of the Synthetic Minority Oversampling Technique (SMOTE) being the most common. While simple duplication risks overfitting, advanced SMOTE variants integrate data distribution specificities, such as KNSMOTE (k-means clustering-based SMOTE) [53] used by Xu et al. [54] to filter noise and overlapping samples, achieving superior performance on datasets like the Haberman and Pima Diabetes datasets. Alternatively, under-sampling reduces the majority class, often by integrating clustering (e.g., K-means) to remove non-informative or noisy samples, as demonstrated by Babar and Ade

[55] and Jain et al. [56], with high efficacy on datasets like Parkinson's Disease and Chronic Kidney Disease.

Learning-level methods modify the training process itself. Cost-sensitive learning is a highly effective strategy that assigns specific, higher penalties for the misclassification of minority instances. This approach has yielded optimal results in diagnostics, such as the cost-sensitive XGBoost model for breast cancer [57] and a cost-sensitive version of Multiple Layer Perceptron used for IBD detection [58]. Beyond cost-sensitivity, the complexity of learning models is increasing using ensemble learning (e.g., AdaBoost, Random Forest) and deep learning algorithms (e.g., CNN, GRU), which are often combined with optimization techniques like Genetic Algorithms (GA) to enhance parameters [59]. Simple classifiers, when paired with robust postprocessing (hyperparameter tuning) or effective feature selection, can still achieve significant performance, highlighting that simple, interpretable models should not be overlooked [60] [61].

The most sophisticated and often top-performing solutions are hybrid techniques, which combine data-level and learning-level strategies. These commonly pair SMOTE variants with under-sampling methods like Tomek Links or ENN [62] [63] to clean up synthetic data noise. Newer trends in this combined space include integrating sampling with Generative Adversarial Networks (GANs) for generating high-quality synthetic data [64] [65] or embedding feature selection into the pipeline. Crucially, the literature confirms that strategic post-synthesis filtering is necessary, as demonstrated by the success of methods like KNSMOTE which use clustering to refine the quality of synthetic samples [54].

Despite these methodological advancements, key challenges persist. A critical gap remains in achieving a reliable compromise between sensitivity (correctly identifying the diseased) and specificity (correctly identifying the non-diseased), a trade-off often ignored in favor of maximizing minority class detection [52]. Furthermore, the progressive complexity of algorithms necessitates a greater focus on interpretability and explainability (Explainable Artificial Intelligence XAI) [52] to foster trust and adoption in clinical settings. The research community also faces persistent standardization issues, including the lack of a universal metric for quantifying imbalance severity and inconsistent reporting of methodological details, which complicates the comparative analysis of findings across varied medical datasets [52].

2.7 Deep Learning Architectures for Sequential Prediction

The effectiveness of predictive modeling on longitudinal EHR data, especially data synthesized by time-aware frameworks like TimeGAN, depends on the selection of an architecturally aligned classifier. Traditional static classifiers (e.g., MLP) are fundamentally incapable of leveraging the temporal dependencies preserved in the synthetic time-series.

The Long Short-Term Memory (LSTM) Network is the specialized deep learning architecture best suited for this task [66]. LSTMs are Recurrent Neural Networks (RNNs) that represent the preferred choice for sequence data due to their sophisticated internal gating mechanisms (Input, Forget, and Output gates). This mechanism allows them to retain and manage information across sequences, enabling them to capture long-range temporal dependencies across multiple time steps, as shown in the cell architecture **Figure 2-3** (Adapted from [66]).

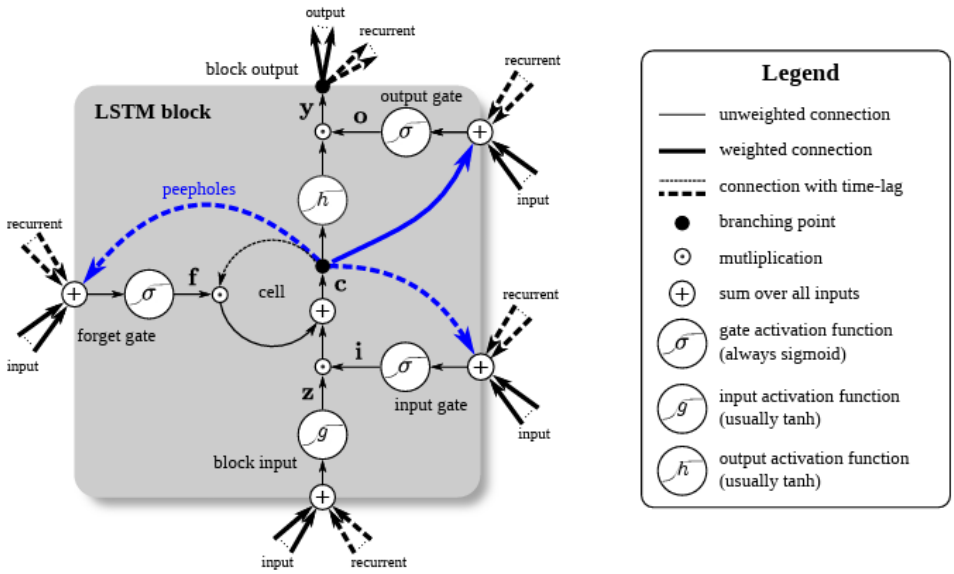
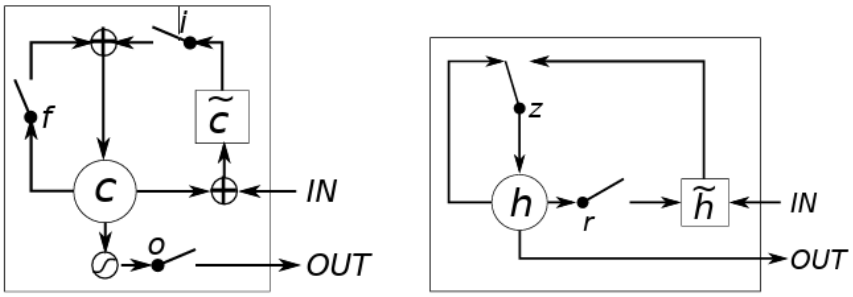


Figure 2-3 Long Short-Term Memory (LSTM) cell architecture

The LSTM's performance on synthetic data directly measures whether the temporal coherence enforced by TimeGAN translates into superior real-world predictive power.

Recent research has explored various simplified variants of the LSTM architecture, aiming to maintain performance while reducing computational complexity:

- **Gated Recurrent Unit (GRU):** Proposed by Cho et al. [67], the GRU replaces the LSTM's three gates with an update gate and a reset gate. The simplified structure of the GRU cell is detailed in Figure 2-4. (Adapted from Chung et al. [68])



(a) Long Short-Term Memory (b) Gated Recurrent Unit

Figure 2-4 LSTM and GRU cell architecture

Performance comparisons observed that the GRU performed comparably or even exceeded the LSTM [68]. The core components of both architectures are described as follows [68]:

"Illustration of (a) Long Short-Term Memory and (b) gated recurrent units. (a) *i*, *f* and *o* are the input, forget and output gates, respectively. *c* and $\tilde{c}$ denote the memory cell and the new memory cell content. (b) *r* and *z* are the reset and update gates, and h and $\tilde{h}$ are the activation and the candidate activation." [68]

- **Minimal Gate Unit (MGU):** Zhou et al. [69] proposed the MGU, which has a minimum of one gate (the forget gate).

- **Critical Components:** Extensive architectural explorations [66] and evaluations [51] have confirmed that key elements, particularly the forget gate and the output activation are critical components for effective sequence learning.

By selecting appropriate sequential architecture, this study ensures that the predictive model can fully exploit the high-fidelity temporal output generated by the TimeGAN framework.

2.8 Performance Evaluation Metrics

To evaluate the predictive power of the trained models, a suite of performance metrics is used, focusing on the classification task. These metrics are particularly critical when dealing with EHR data, which is often subject to class imbalance.

Evaluating the performance of any classification model, especially when dealing with potentially imbalanced medical datasets, requires a set of precise and informative metrics. These metrics are derived from the fundamental counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Beyond simple Accuracy, the metrics of Precision and Recall are essential, as they provide insight into the model's reliability and its completeness in identifying positive cases, respectively. The F1-score then serves as a single measure to balance these two critical trade-offs. The specific definitions and formulas for these core classification metrics are presented in Table 2.1:

Table 2.1 Classification Metrics

Metric	Description	Formula
Accuracy	The most straightforward metric. It represents the proportion of total predictions that were correct across all classes.	$a = \frac{TP + TN}{TP + TN + FP + FN}$
Precision (p)	Also known as Positive Predictive Value. It measures the reliability of a positive classification. It answers: Out of all cases the model predicted as positive for a class, how many were correct?	$p = \frac{TP}{TP + FP}$
Recall (r)	It is also known as Sensitivity or True Positive Rate. It measures the ability of the model to find all the relevant cases for each class. It answers: Out of all the actual positive cases for a class, how many did the model correctly identify?	$r = \frac{TP}{TP + FN}$
F1-score	The harmonic Mean of Precision and Recall. It is useful because it punishes models that achieve high Precision but low Recall, or vice versa, providing a single score balancing both metrics.	$F1 = 2 \times \frac{p \times r}{p + r}$

The predictive models are evaluated using average metrics to provide a comprehensive view of performance across all classes, which is essential given the class imbalance inherent in EHR datasets. The averaging methodology drastically affects the interpretation of the results, particularly for Precision, Recall, and the resulting F1-score. It is important to understand the macro and weighted averaging:

- **Macro Average (Unweighted)** The Macro average calculates the metric (Precision, Recall, or F1-score) independently for each class, and the final Macro score is the simple arithmetic average of those class scores. This approach treats all classes equally, regardless of their size (its support). A strong Macro score indicates the model is not biased toward the majority class and performs robustly in the smallest, rarest classes. This is the most critical metric for assessing success in mitigating the imbalance of high-stakes, low-incidence clinical outcomes.
- **Weighted Average** the Weighted average calculates the metric for each class and then averages them, weighing each class score by the number of samples belonging to that class (its support). This metric reflects the expected performance of the model on the full, original, imbalanced dataset. It is naturally dominated by the majority class and is generally of higher value. It provides a measure of overall prediction quality but

should be considered alongside the Macro scores to fully understand the model's performance in the minority classes.

2.9 Conclusion

This chapter established the critical foundation for predictive modeling using Electronic Health Records (EHRs), confirming that the central challenges are the data's sequential nature, severe class imbalance, and issues of data scarcity and temporal fidelity. The literature review confirms that Time-series Generative Adversarial Networks (TimeGAN) represent the state-of-the-art solution for synthesizing high-fidelity sequential data, capable of preserving the crucial temporal dependencies required for clinical prediction [12]. Furthermore, it was established that sequential prediction tasks necessitate specialized, recurrent deep learning architectures, such as LSTM and GRU networks [66], as traditional static models are fundamentally inadequate for exploiting time-dependent features.

This comprehensive review identifies a significant gap in current literature: the empirical evaluation of the architectural alignment necessity. Specifically, while TimeGAN is proven to generate superior temporal features, there is a lack of rigorous, comparative evidence demonstrating that a prediction model trained in this output must utilize sequential (e.g., GRU) architecture to maximize utility, as opposed to a simple static (e.g., MLP) architecture. The following chapter, Methodology, is therefore dedicated to designing and executing an experiment that leverages TimeGAN's high-fidelity synthetic data to rigorously test this gap. It will define the data pipeline, the implementation of the sequential TimeGAN framework, the sampling strategy used to mitigate class imbalance, and the comparative experimental setup necessary to validate the essential role of architectural alignment in maximizing predictive performance on real-world longitudinal EHR data.

3 Methodology and Experiments

This chapter details the specific, empirical steps taken to conduct the research, including dataset description, data cleaning preprocessing, synthetic data generation, and the final model training and evaluation protocols.

3.1 Data Preparation and Preprocessing

This section details the initial handling and preparation of the raw patient data, ensuring data quality, consistency, and suitability for sequential modeling. The complete list and detailed description of all features utilized in this study are provided in Appendix A - Feature Set Overview.

The core dataset utilized is a longitudinal cohort study of patient health trajectories, characterized by a small sample size of 117 total records across multiple visits [70]. This inherent data scarcity and its longitudinal structure (sequential visits over time) are the primary justifications for the use of synthetic data augmentation. The target classification variable was established as (*health_rate*), which contains five distinct, ordinal classes. The distribution of this target class in the original dataset confirmed a severe class imbalance, with the majority class (Class 4 - Good) dominating the distribution (see Table 4.3 in Chapter 4).

The raw dataset, characterized by its longitudinal structure and heterogeneous data types (numerical, categorical, and ordinal), required preparation for neural network training.

- **Feature Encoding and Scaling:** Categorical features were converted using One-Hot Encoding, and ordinal features were converted using Ordinal Encoding. All numerical features were normalized using Min-Max Scaling to confine their values within the range [0, 1]. Min-Max Scaling was specifically chosen over Standardization (Z-score) because it preserves the precise bounding nature of clinical variables, aligning with the requirements of the Time-series Generative Adversarial Network (TimeGAN) architecture for stable boundary conditions. The final data frame consisted of 56 key columns, after removing the *mna_total* feature due to excessive missing values (>95%).
- **Feature Integrity and Imputation:** Feature integrity was strictly maintained through targeted outlier detection followed by imputation. Outliers were detected exclusively using the Z-score method (threshold=3.0). Missing values were handled using an Imputation within the Primary Key (PK) approach to maintain longitudinal consistency. This Group-based Strategy is crucial because it utilizes the mean/median/mode derived from the individual patient's records (*part_id*). Crucially, if a patient's group data was insufficient to calculate a reliable statistic, the imputation process included a fallback to the global mean/median/mode from the entire dataset. This strategy preserves the uniqueness of individual patient trajectories while ensuring all missing values are addressed.
- **Dimensionality Reduction:** To refine the feature space and reduce noise input for the TimeGAN, Principal Component Analysis (PCA) was performed on the cleaned and encoded dataset. This method was applied separately to the numerical and categorical feature sets (post-One-Hot Encoding). The goal of this process was to select the

optimal number of components for each set that maximized the cumulative variance retention, ensuring an efficient feature space was created without sacrificing critical information. The detailed quantitative analysis justifying the component selections is provided in Appendix C - Correlation and PCA Analysis for dimensions reduction.

3.2 Synthetic Data Generation via TimeGAN

This section details the transformation of the prepared data into synthetic sequences using the Time-series Generative Adversarial Network (TimeGAN).

The prepared data was transformed into the sequential structure required by TimeGAN: sequences and were segmented into a fixed length T = 4 visits and filled using zero-padding. The final training input was a 3D tensor (Number of Sequences, T = 4, D = *num_features*).

The TimeGAN architecture was implemented using the TimeSeriesSynthesizer. All four core components (Embedding, Generator, Recovery, and Discriminator) were constructed using Gated Recurrent Units (GRUs), selected for their efficiency in time-series modeling [68]. The training process relies on minimizing a multi-component loss function: Reconstruction Loss, Adversarial Loss and Supervisory Loss.

The specific parameters for the TimeGAN architecture and the training process are summarized in Table 3.1 below, with a high epochs value (2,000) necessary to ensure adequate convergence of the adversarial components.

Table 3.1 TimeGAN Model and Training Hyperparameters

Parameter Category	Parameter	Value	Description
Recurrent Unit Type	RNN Unit	GRU	Used in all four networks (E, G, R, D) to capture temporal dynamics.
Model Capacity	layers_dim	256	The dimension of the hidden state for the recurrent units, defining the model's capacity.
Latent Dimension	latent_dim	32	The dimension of the compressed, feature-rich latent space [12].
Noise Dimension	noise_dim	64	The dimension of the random noise vector input to the Generator.
Discriminator Loss Weight	gamma	1.0	The weight applied to the Discriminator's objective function.
Sequence Length (T)	sequence_length	T_max	The length of the sequences used in training, equal to the maximum number of visits (4 in this study).
Batch Size	batch_size	128	Number of sequences processed per training step.
Training Epochs	epochs	2000	Total number of full passes over the training dataset.
Learning Rate	lr	5×10^{-4}	Constant learning rate used for the Adam optimizer across all networks.

TimeGAN generation was conducted in two distinct phases to support the comparative experimental design:

1. **Phase I (Imbalanced Baseline):** The model was trained on the raw data to replicate the original dataset's native class imbalance, creating the Synthetic Imbalanced Dataset.

2. **Phase II (Balanced Experiment):** A targeted post-generation rejection sampling strategy was implemented on a large, generated set to create the Synthetic Balanced Dataset, thereby mitigating severe class imbalance in the *health_rate* target variable prior to model training.

Following the generation of synthetic data, the 3D data array underwent Denormalization and Decoding, followed by Dynamic Temporal Reconstitution. This reconstitution step is critical because TimeGAN synthesizes the sequence of values but not the timing between visits. The visit date (*q_date*) was reconstituted by randomly sampling a starting date and iteratively adding a random time delta (constrained between 6 and 9 months), thereby preserving the observed inter-visit frequency and clinical plausibility. The generated data's fidelity was validated using the Table Evaluator tool (including t-SNE/PCA visualizations). This validation confirmed that the synthetic data successfully replicated the marginal and joint distributions of the original data, validating its use as a high-fidelity training resource.

3.3 Experimental Design and Classifier Training Protocol

This section outlines the data partitioning and the comparative classification protocols used to test the core hypothesis: the necessity of architectural alignment when modeling TimeGAN-generated sequential data.

The experimental design utilizes a partitioning strategy based on the data source:

- **Real Data Partitioning:** The original real dataset was split into a 70% training set (exclusively for M1) and a dedicated 30% test set (T_M1).
- **Synthetic Data Partitioning:** For the synthetic experiments (M2, M3, M4), training was conducted on a corresponding 70% split of the respective synthetic datasets.

Evaluation was performed against four distinct test sets to rigorously assess generalization and fidelity:

- **Test Set T_M1 (M1 Baseline Test):** The 30% portion of the original real data, used exclusively to test the control Model M1's performance on never-seen real data.
- **Test Set T1 (All Real Data):** 100% of the original real data, used to measure synthetic-trained models (M2, M3, M4) real-world generalization (Train Synthetic, Test Real - TSTR).
- **Test Set T2 (Synthetic Never Seen):** The remaining 30% of the synthetic data, used to measure synthetic fidelity for models M2, M3, M4.
- **Test Set T3 (Combined Test Set):** A concatenation of T1 and T2 to measure overall robustness for models M2, M3, M4.

Four classification models were trained and compared to test the core hypothesis of architectural alignment and summarized in table 3.2:

Table 3.2 Classification Models

Model	Training Data Used	Classifier Type	Core Hypothesis Test
M1 (Control)	70% Real Data (Imbalanced)	Static MLP	Baseline performance on original data, representing the failure state due to data scarcity.
M2	70% Synthetic Imbalanced Data	Static MLP	Measures synthetic utility without balancing or sequence exploitation.
M3 (Intermediate)	70% Synthetic Balanced Data	Static MLP	Measures the impact of synthetic balancing using a classifier not designed for sequence data.
M4 (Core Test)	70% Synthetic Balanced Data	Sequential LSTM	Measures the combined effect of balancing and architectural alignment with TimeGAN's sequential output.

The Static Multi-Layer Perceptron (MLP) classifier (M1, M2, M3) was implemented with a multi-layer architecture utilizing Dense layers and Rectified Linear Unit (ReLU) activation. The Sequential Long Short-Term Memory (LSTM) classifier (M4) utilized a single LSTM layer

followed by a final Dense layer. The specific unit counts for all models were determined via hyperparameter optimization using the Optuna framework (detailed in Table 4.7 in Chapter 4). Both classifiers were trained using an adaptive optimization algorithm (Adam or RMSprop as optimizer) and the Binary Cross-Entropy loss function.

The comparison of M3 (Static) vs. M4 (Sequential LSTM) will directly test the necessity of sequential architecture for maximizing predictive performance on high-fidelity time-series data.

This chapter established the rigorous data preparation pipeline, defined the TimeGAN architecture and parameters, and outlined the four comparative classification experiments (M1 through M4) along with their corresponding four evaluation test sets (T_M1, T1, T2 and T3). The following chapter, Chapter 4 Results and Findings, presents the empirical data derived from these experiments, including the detailed data distributions, the results of the hyperparameter optimization, and the final comparative performance metrics of all four models across the defined test sets.

The complete, version-controlled source code for this dissertation can be accessed at: https://github.com/paulobanha-web/VSCODE

4 Results and Findings

This chapter presents the empirical outcomes derived from the data preparation pipeline, the synthetic data generation process, the definition of the training and evaluation datasets, the hyperparameter optimization, and the final performance metrics for the developed models. Specifically, the performance of the four primary classification models (M1 to M4) is presented, demonstrating the comparative effectiveness of synthetic data augmentation and temporal modeling. These findings establish the foundation for the discussion and interpretation presented in the subsequent chapter.

4.1 Definition of Experimental Models and Evaluation Sets

Due to the scarcity of the original dataset, the experimental design utilized four distinct models and four unique test sets to rigorously evaluate the impact of synthetic data and sequence modeling.

The four classifier models, M1 through M4, are defined by their architecture and the type of data used for training (70% split of the respective dataset) as follows in table 4.1:

Table 4.1 Classifier Models

Model Name	Architecture	Training Set Size	Training Data	Purpose in Study
M1	MLP	81	70% Real Records (Baseline)	Control group; establishes the performance ceiling/floor without augmentation.
M2	MLP	10592	70% Synthetic Imbalanced Records	Assesses synthetic data utility while retaining native class bias.
M3	MLP	11132	70% Synthetic Balanced Records	Assesses synthetic data utility with targeted class imbalance mitigation.
M4	LSTM	11132 (2783 Patients)	70% Synthetic Balanced Sequences	Assesses the benefit of temporal modeling (LSTM) using balanced synthetic data.

Evaluation was conducted using four distinct test sets, T_M1 for the baseline M1 model, and T1, T2, and T3 for the augmented models (M2, M3, M4) as follows in table 4.2:

Table 4.2 Evaluation Test Sets

Test Set Name	Data Composition	Size Imbalanced	Size Balanced	Data Source	Purpose
T_M1 (M1 Test Split)	30% Real Records	36 records	n.a.	Held-out real data	Measures M1's generalization on real data never seen during training.
T1 (All Real Data)	100% Real Records	117 records	117 records 27 patients	All original real data	Measures models {M2-M4's real-world generalization (Train Synthetic, Test Real).
T2 (Synthetic Never Seen)	30% Synthetic Data	4,540 records	4,772 records 1,193 patients	Held-out synthetic data	Measures M2-M4's fidelity, or how well the models learned the TimeGAN output.
T3 (Combined Test)	T1 + T2	4,657 records	4,889 records 1,220 patients	All Real + Never-seen Synthetic	Measures the overall robustness and prediction quality on the augmented distribution.

4.2 Data Integrity, Feature Engineering and Dimensionality Reduction

This section presents the empirical outcomes of data cleaning, feature engineering, and preparation pipeline applied to the raw longitudinal dataset, critical for establishing a high-quality feature space prior to TimeGAN training and downstream classification.

4.2.1 Data Consistency and Target Class

The initial check for data consistency, using the composite key of (*part_id* and *clinical_visit*), confirmed no duplicate data records were present. One feature, (*mna_total*), was subsequently removed due to an extremely high proportion of missing values (>95%) and insufficient variance. The target classification variable was established as (*health_rate*), which contains five distinct, ordinal classes, whose distribution in the original dataset confirmed a significant class imbalance (Table 4.3), which directly necessitated the synthetic data augmentation strategy.

Table 4.3Target Class Distribution in Original Dataset

Health Rate	Count
5 - Excellent	5
4 - Good	71
3 - Medium	31
2 - Bad	6
1 - Very bad	4
Total	117

4.2.2 Outlier Detection and Remediation Strategy

Univariate outlier detection was performed using the Z-Score method across all numerical features. A hybrid remediation strategy was implemented to distinguish between biologically impossible errors and plausible clinical extremes. Physiologically implausible values were identified and replaced with null values (NaN) to be handled during imputation; these errors included missing data codes in physical performance metrics, as well as clear measurement or entry errors (e.g., 999.0) were assessed using box plot and histogram charts to address them and take a decision based on the evidence. The figure 4-1 shows clearly the outlier detection on feature *gait_get_up*.

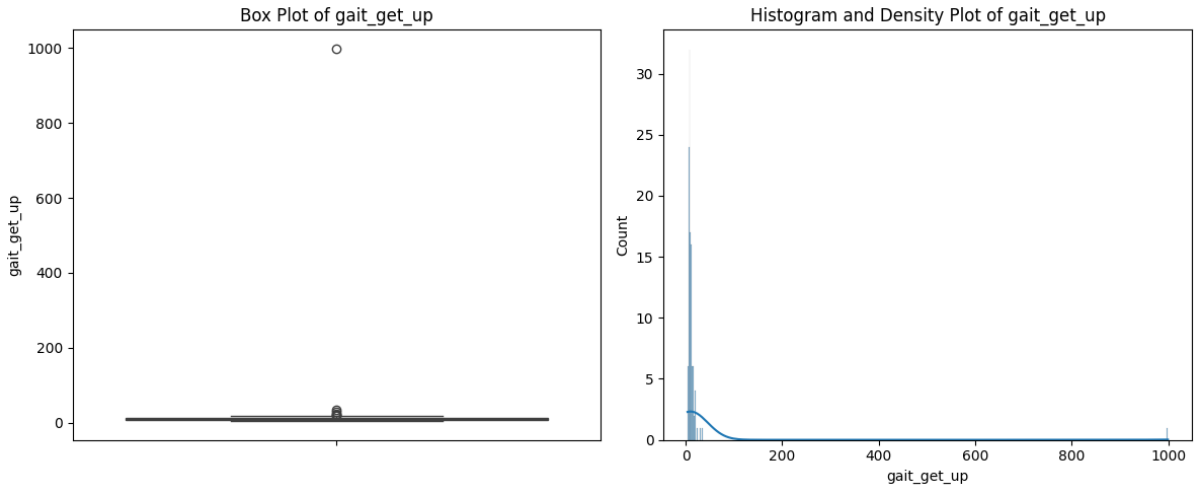


Figure 4-1Outlier visualization of gait_get_up

The value 999.0 functions as a sentinel value, likely representing 'test not performable' or missing data. As Z-score detection would flag this value as a high numerical outlier if unaddressed, it must be explicitly converted to null (NaN) to facilitate robust analysis and proper imputation

The detailed empirical justification, including visualizations for key features, is provided in Appendix B - Outliers Analysis. In contrast, extreme but biologically possible values were

retained to ensure the synthetic data model learned the full, clinically relevant spectrum of the cohort.

4.2.3 Imputation of Missing Values

Following data cleaning and the outlier remediation step, which converted certain erroneous values into null values (NaN), the remaining missing values (including both original NaN and newly created nulls) across both numerical and categorical features were imputed. This process utilized a Group-based Strategy that prioritized consistency within the individual patient's trajectory, identified by the unique composed primary key (*part_id* and *clinical_visit*). Specifically, for numerical features, imputation was performed by calculating the Mean or Median of the non-missing values specific to the individual patient group. When an individual patient group contained insufficient data for this calculation, the process fell back to using the global statistic (Mean or Median) from the entire dataset, which was not necessary since there were enough clinical visits per patient with plausible values that were used for the group strategy. Similarly, for categorical features, imputation was performed using the Mode (most frequent value) derived from the individual patient's records, with a fallback to the global mode if the patient-specific mode could not be determined. The specific imputation feature assignments are detailed in Table 4.4, ensuring a complete feature set while preserving longitudinal integrity.

Table 4.4 Imputation Strategy by Feature Type

Feature Type	Feature Name	Imputation Strategy
Numerical	bmi_body_fat, lean_body_mass, waist	Mean
Numerical	gait_get_up, raise_chair_time	Median
Categorical	activity_regular, cognitive_total_score, comorbidities_most_important, house_suitable_participant, house_suitable_professional, leisure_club, memory_complain, sleep, stairs_number	Mode

4.2.4 Final Feature Space and Dimensionality Reduction

The initial step in defining the model-ready feature set involved two sequential analyses: correlation analysis to identify and exclude highly redundant variables, followed by Principal Component Analysis (PCA) to generate a compact, information-rich feature space for the TimeGAN architecture.

The original cleaned and encoded dataset comprised 7 numerical variables and 48 categorical variables (post-One-Hot Encoding), with a total of 55 features excluding *q_date* feature. A preliminary correlation analysis was executed to identify potential linear multicollinearity, which can destabilize deep learning models. The correlation heatmap for the initial 7 numerical features is presented in Figure 4-2, while the complete correlation matrix for the 48 categorical features is provided in Appendix C. This analysis confirmed that no feature pairs exhibited linear multicollinearity above the stringent set threshold ($r > 0.95$), ensuring that the subsequent PCA operated on variables that contributed distinct, non-redundant information.

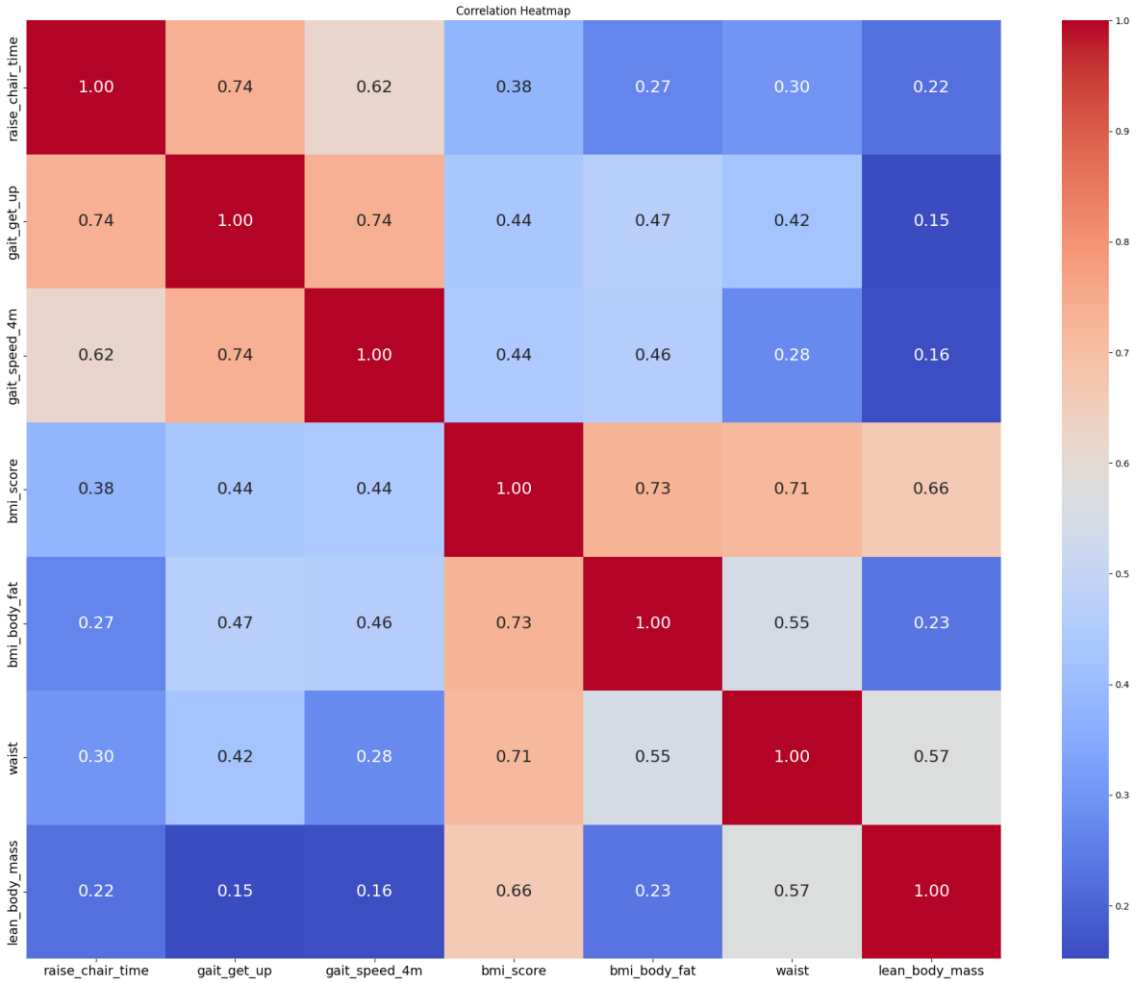


Figure 4-2 Correlation Heatmap Numerical Features

PCA was then applied to the numerical and categorical feature matrices to maximize variance retention while drastically reducing the input feature dimensionality (D). This process identifies the components that explain the most variance, guiding the selection of the most informative features.

- Numerical Feature Reduction (D = 7 to 5): For the 7 numerical features, the analysis determined that 5 principal components were sufficient to capture most of the cumulative variance. The Cumulative Explained Variance plot (Figure 4-3) visually supports this decision, justifying the selection of the 5 most discriminative features.

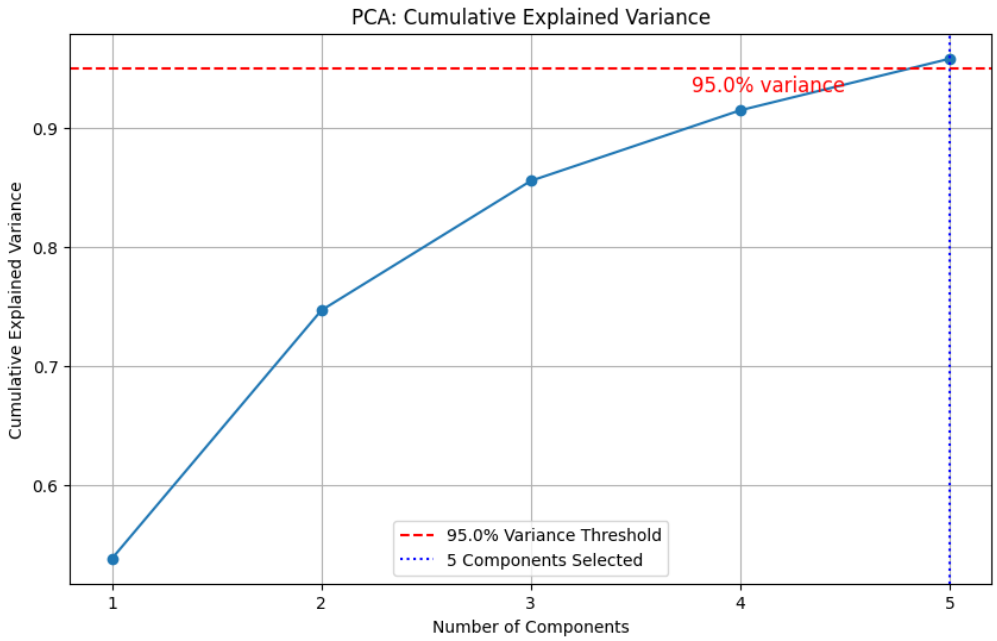


Figure 4-3Cumulative Variance Numerical Features

- Categorical Feature Reduction (D = 48 to 34): For the 48 categorical features, the analysis required retaining 34 components to meet the variance threshold. The detailed Cumulative Explained Variance plot supporting this decision is provided in Figure 4-4.

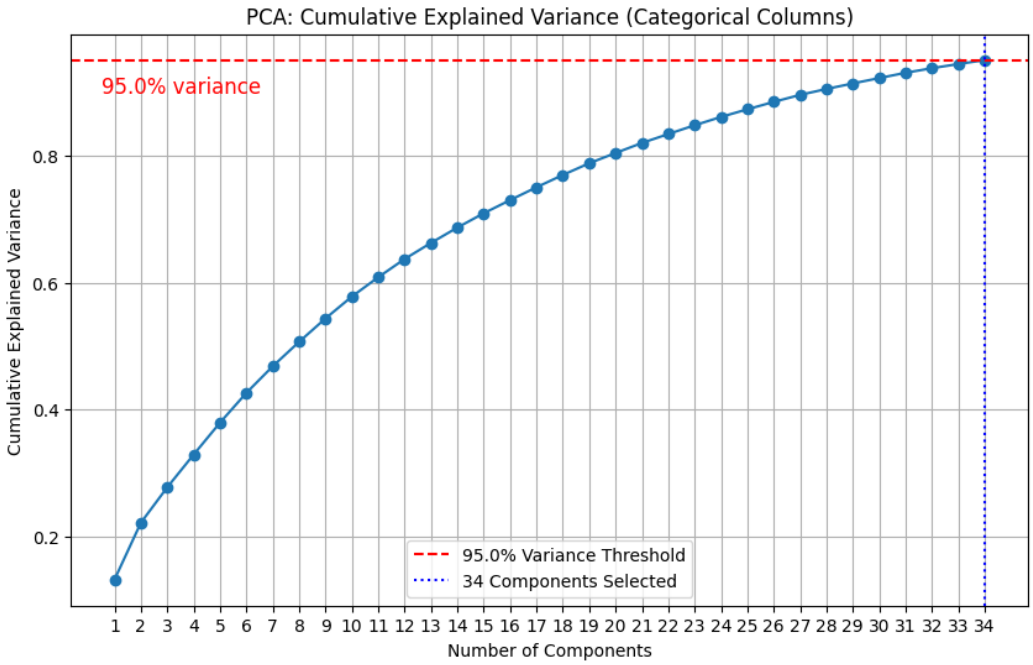


Figure 4-4 Cumulative Variance Categorical Features

This process resulted in a final feature space comprising 5 numerical components and 34 categorical components, yielding a total feature space of 39 components for sequence encoding and TimeGAN modeling, which effectively dropped 17 redundant or less discriminative variables.

4.3 Generation of Synthetic data

This section details the two distinct phases of the synthetic data generation process using the Time-series Generative Adversarial Network (TimeGAN). The goal was to produce high-fidelity longitudinal patient trajectories that could augment the original real dataset. Phase I focused on preserving the natural class imbalance for baseline comparison, while Phase II employed a rejection sampling strategy to achieve a near-balanced target distribution.

4.3.1 TimeGAN Synthetic Data Generation (Imbalanced Strategy)

The primary goal of this initial phase was to generate high-fidelity synthetic time-series data using the TimeGAN model while preserving the natural class imbalance. The TimeGAN model was successfully trained for 2,000 epochs using pre-processed and scaled real patient data. A total of 15,132 synthetic patient visits were ultimately generated. Post-processing revealed a notable shift in the majority class: while the original data was dominated by Class 4 (Good), the synthetic output was dominated by Class 5 (Excellent) The resulting distribution is shown in table 4.5:

Table 4.5 Imbalanced Health Rate data distribution

Health Rate Class	Synthetic Data Count	Synthetic Data Proportion
5 - Excellent	7882	52.09%
4 - Good	2831	18.71%
3 - Medium	3192	21.09%
2 - Bad	649	4.29%
1 - Very bad	578	3.82%
Total	15132	100.00%

This result confirms the efficacy of TimeGAN as a method for generating realistic time-series, even capturing the inherent bias of the original data distribution, although shifting the peak to the highest health category.

4.3.2 TimeGAN Synthetic Data Generation (Targeted Balance Strategy)

This phase utilized unconditional TimeGAN generation followed by a rejection sampling process to attempt to balance the distribution. The key-finding here is the failure to achieve true class balance. Despite the filtering logic, the resulting synthetic dataset remained highly skewed toward the dominant class, "5 - Excellent." The extreme minority classes (1 and 2) remained under sampled, suggesting the TimeGAN model struggled to generate diverse, high-quality sequences for these underrepresented categories. The resulting improved, yet still skewed, distribution is presented in Table 4.6:

Table 4.6 Balanced Health Rate data distribution

Health Rate Class	Synthetic Data Count	Synthetic Data Proportion
5 - Excellent	7615	47.88%
4 - Good	1952	12.27%
3 - Medium	2775	17.45%
2 - Bad	1657	10.42%
1 - Very bad	1905	11.98%
Total	15904	100.00%

The attempt at class balance merely reduced the overall dominance of the majority class while still yielding a highly skewed distribution. This actual finding is a critical result that influences the interpretation of any downstream model trained on this data. The normalized distributions are visually compared in Figure 4-5 to confirm the residual skew.

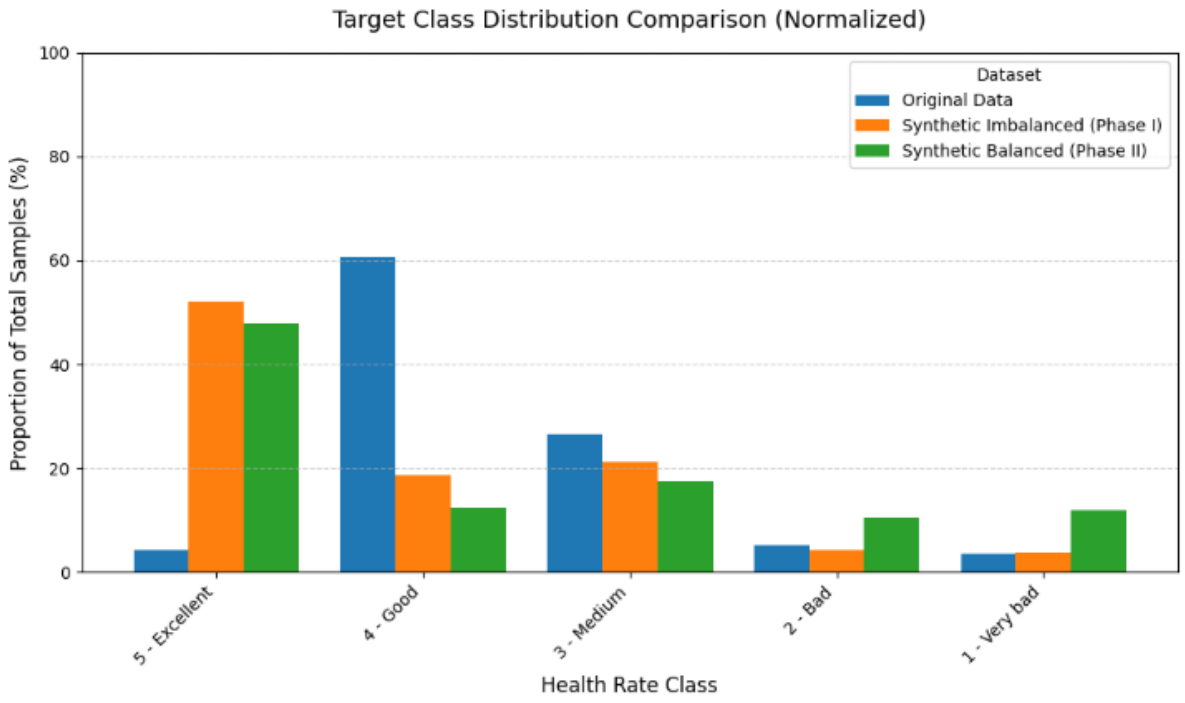


Figure 4-5 Normalized Target Class Distribution Comparison across Datasets

4.4 Data Quality and Evaluation

The fidelity of the data generated was rigorously assessed using the Table Evaluator framework, quantifying the similarity between the marginal and joint distributions of the real and synthetic feature sets.

4.4.1 Statistical Fidelity (Marginal Distribution)

The Kolmogorov-Smirnov (K-S) test showed that the Imbalanced Synthetic Data (for M2 Training Set) achieved a high degree of statistical overlap with the real dataset. This finding is visually confirmed by the Cumulative Distribution Function (CDF) in Figure 4-6 and the Probability Density Function (PDF) in Figure 4-7, which show the distributions closely aligning.

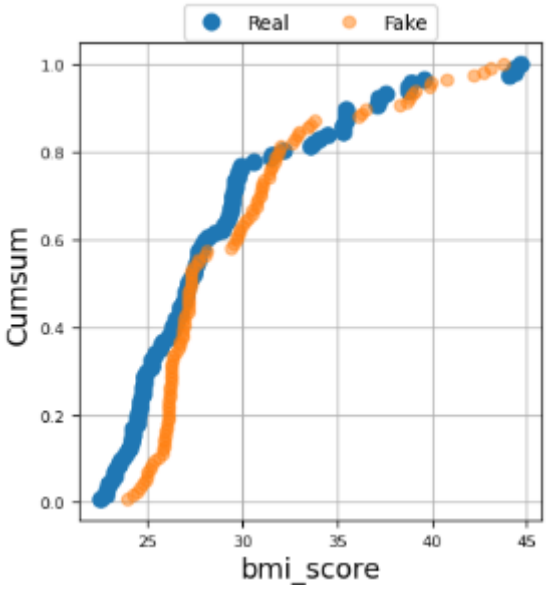


Figure 4-6 CDF Imbalanced

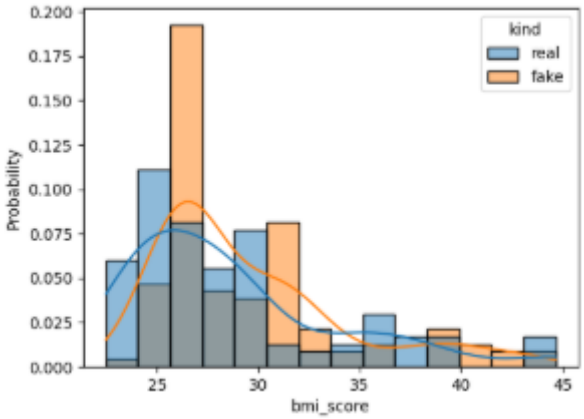


Figure 4-7 PDF Imbalanced

In contrast, the fidelity assessment of the Balanced Synthetic Data (for M3/M4 Training Set) showed less statistical, resulting in a measurable gap in the CDF (Figure 4-8) and PDF lines (Figure 4-9).

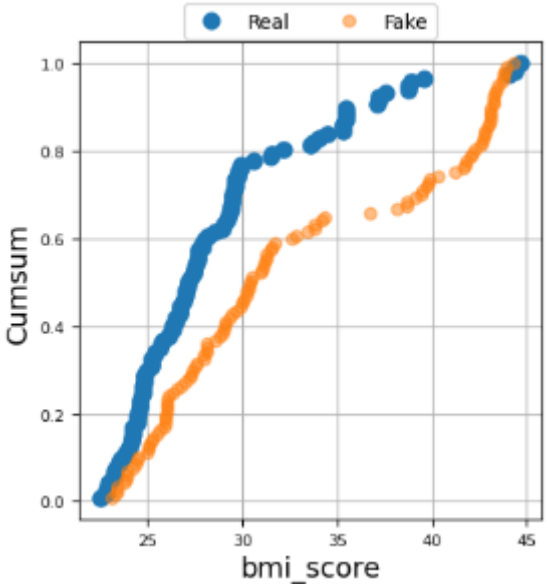


Figure 4-8 CDF Balanced

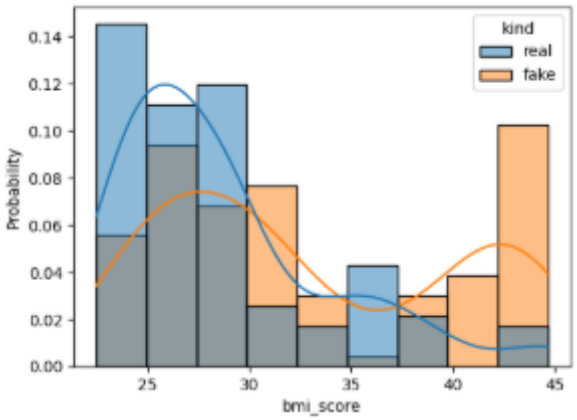


Figure 4-9 PDF Balanced

This discrepancy supports the conclusion that the rejection sampling strategy used to equalize the class distribution, while serving its intended purpose, introduced a minor distortion to the statistical properties of the synthetic feature space relative to the original real data.

The complete assessment is presented in Appendix D.

4.4.2 Structural Fidelity (Joint Distribution and Feature Dependencies)

Structural fidelity was assessed using dimensionality reduction techniques and correlation matrix analysis to confirm that the TimeGAN captured the complex, non-linear relationships between features.

Principal Component Analysis (PCA) was utilized to project the high-dimensional feature space onto two dimensions. The results for the Imbalanced Synthetic Data (Figure 4-10) demonstrate that the TimeGAN successfully preserved the underlying manifold shape of the real dataset. As observed in the real data, the synthetic points remained concentrated around the origin (x=0), with a consistent vertical spread along the Y-axis (ranging from -20 to 20) and

a sparse scattering of outliers. This confirms that without external constraints, the model correctly replicated the majority-dominated geometry of the real feature space.

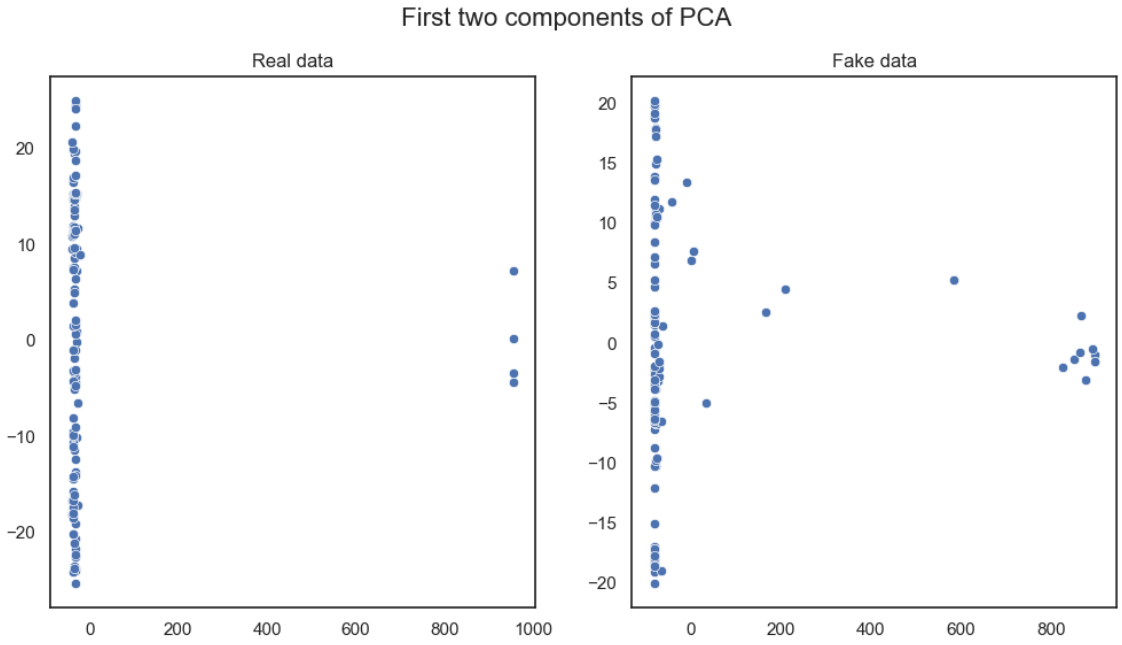


Figure 4-10 First 2 Components of PCA Imbalanced Synthetic Data

Significantly, the PCA projection of the Balanced Synthetic Data (Figure 4-11) reveals a marked distributional shift rather than a simple overlay.

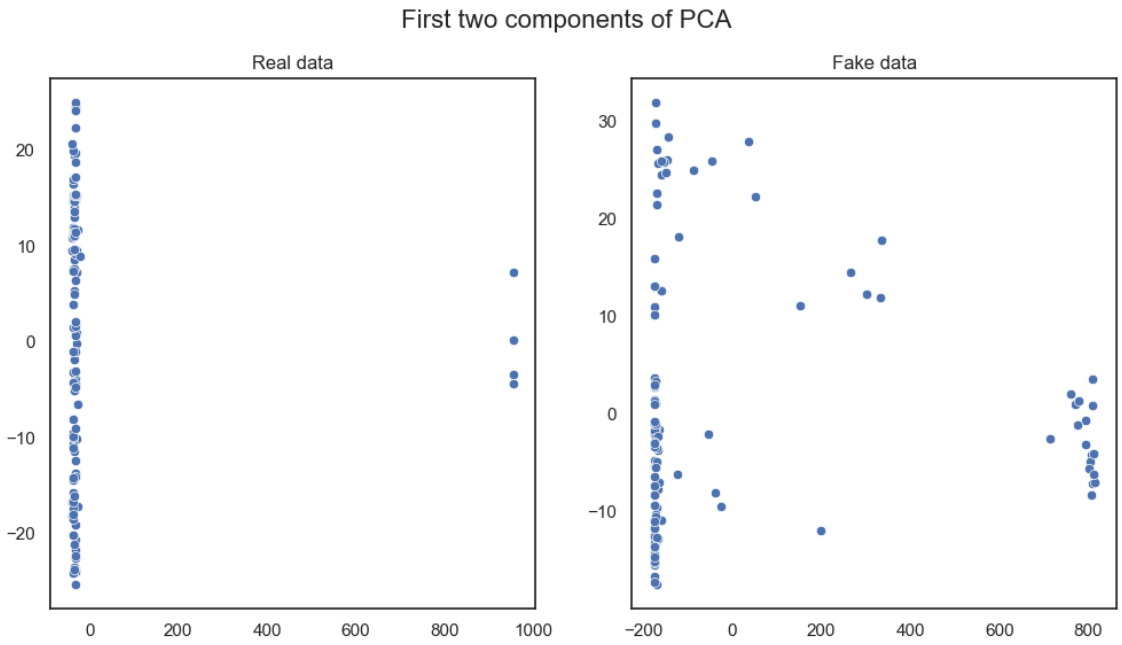


Figure 4-11 First 2 Components of PCA Balanced Synthetic Data

Three distinct structural changes confirm the impact of the balancing strategy:

1. **Centroid Shift:** The dense cluster originally centered at x=0 in the real data shifted leftward to approximately x=-200.

2. **Scale Expansion:** The scale of the first principal component (X-axis) expanded substantially (from a range of roughly 0 to 1000 in the Imbalanced set, to -200 to 800 in the Balanced set).

3. **Minority Signal Amplification:** The sparse outliers originally observed at the extremes (near x=1000 in real/imbalanced data) evolved into a distinct, denser cluster of approximately 30 scatter points near x=800.

This structural divergence is not a failure of fidelity, but rather an expansion of the feature space. Because the Balanced dataset artificially inflates the prevalence of minority classes (Classes 1 and 2), the statistical center of the dataset necessarily moves away from the "healthy" majority profile. The emergence of the denser cluster at x=800 and the wider spread indicates that the TimeGAN successfully synthesized new, diverse variations of the rare, extreme clinical profiles that were barely visible in the original dataset.

Finally, feature dependency was analyzed using correlation heatmaps. Figure 4-12 demonstrates that the Imbalanced Synthetic Data largely preserves the correlation structure of the real data, maintaining the magnitude and direction of key relationships. However, a specific artifact was observed: a hyper-correlation between low_physical_activity and alcohol_units that exceeded the levels seen in the real data.

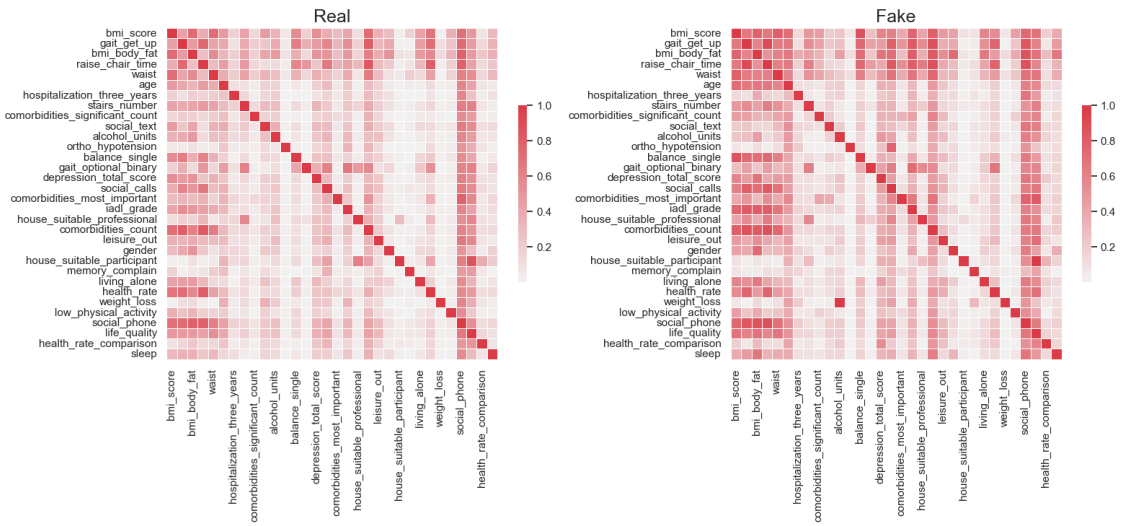


Figure 4-12Correlation Heat Map Imbalanced Real vs Fake

In the Balanced Synthetic Data (Figure 4-13), the overall topology of the correlation matrix remains consistent with the real data. Notably, the previously observed artifact between physical activity and alcohol consumption is attenuated, suggesting that the balancing process may have regularized specific features over-dependencies.

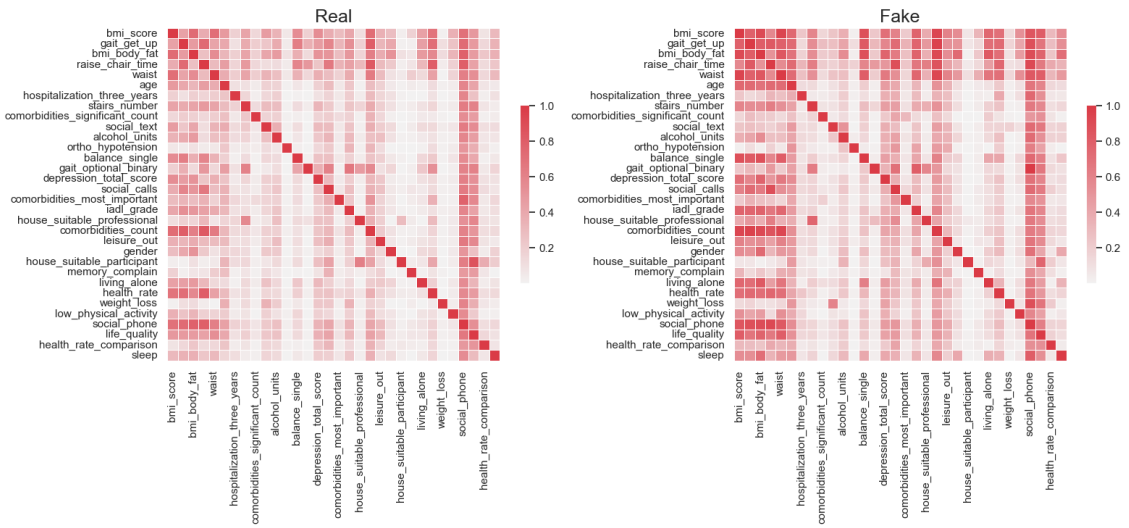


Figure 4-13 Correlation Heat Map Balanced Real vs Fake

The Difference Maps (Figure 4-14 and Figure 4-15) highlight the element-wise divergence between real and synthetic correlations. While the similarity is less visually evident in the Difference plots due to the stochastic nature of GAN generation, the preservation of the macroscopic structure in the primary heatmaps confirms that the TimeGAN effectively learned the essential clinical dependencies required for robust modeling.

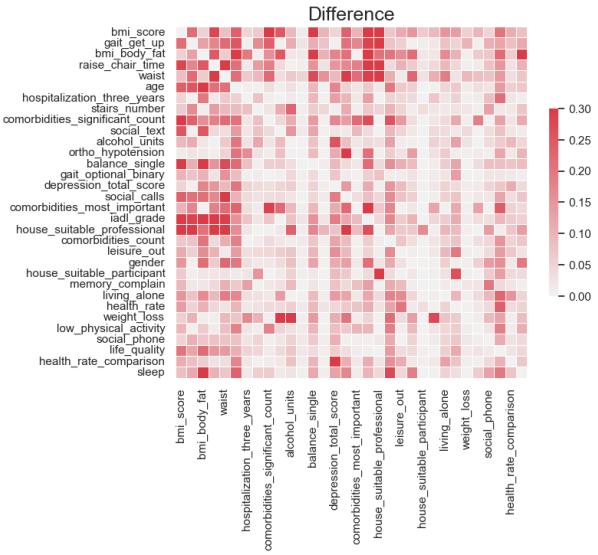


Figure 4-14 Correlation Heat Map Difference between Imbalanced Real and Fake

Figure 4-15 Correlation Heat Map Difference between Balanced Real and Fake

4.5 Hyperparameter Optimization (Optuna) Results

The four downstream classification models (M1, M2, M3, and M4) required extensive hyperparameter tuning using the Optuna framework.

The Optuna process utilized the 70% training split of the respective dataset for each model to ensure a fair search space:

- M1 (MLP) was tuned using the 70% training split of the Real dataset.
- M2 and M3 (MLPs) were tuned using the 70% training split of the Synthetic dataset.
- M4 (LSTM) was tuned using the sequences extracted from the 70% training split of the Synthetic dataset for training and validation within the Optuna trials.

The objective for all four models (M1, M2, M3, and M4) during the Optuna tuning process was to find the optimal configuration that maximizes the best validation accuracy. The optimal parameters are summarized in Table 4.7:

Table 4.7 Optimal Hyperparameters Determined by Optuna for Classification Models

Model Name	Architecture	Hyperparameter	Optimal Value (Optuna)
M1	MLP	n_neurons	18
		learning_rate	0.001568
		patience	24
		optimizer	RMSprop
		n_layers	3
M2	MLP	n_neurons	36
		learning_rate	0.000615
		patience	24
		n_layers	3
M3	MLP	n_neurons	68
		learning_rate	0.000409
		patience	30
		n_layers	3
M4	LSTM	lstm_units	40
		learning_rate	0.005608
		patience	24
		optimizer	Adam
		dropout_rate	0.0650

The optimal parameters confirm that models trained on synthetic data (M2, M3) required significantly higher neural capacity (36 and 68 neurons) compared to the control model M1 (18 neurons), which was forced to adopt a smaller architecture to mitigate severe overfitting due to data scarcity. This increase in capacity confirms that the synthetic data successfully enabled the exploration of more complex, higher-performing model architecture. The temporal model M4 utilized an efficient recurrent layer of 40 units, demonstrating effective temporal learning without needing excessive capacity.

4.6 Comparative Analysis of Model Performance

This section presents the comprehensive results from the four classification experiments, confirming the necessity of synthetic data and identifying the optimal model architecture for predicting patient health rates.

4.6.1 Final Comprehensive Model Performance Metrics

The models were either trained on 70% Real Data (M1) or 70% Synthetic Data (M2, M3, M4), and then rigorously evaluated on four distinct, never-seen test sets (T_M1, T1, T2, T3).

Table 4.8 summarizes the performance of all four classification models across their specific evaluation sets.

Table 4.8 Final Comprehensive Model Performance Metrics

Model	Evaluation Set & Size (n)	Accuracy	F1-Score (Weighted)	F1-Score (Macro)
M1	T_M1 (n=36)	61.11%	46.36%	15.17%
M2	T1 (n=117)	40.17%	51.56%	51.12%
	T2 (n=4540)	93.79%	93.78%	91.44%
	T3 (n=4657)	92.44%	92.43%	90.37%
M3	T1 (n=117)	47.01%	58.65%	60.11%
	T2 (n=4772)	89.27%	89.26%	88.89%
	T3 (n=4889)	88.26%	88.24%	87.98%
M4	T1 (27 patients)	66.67%	74.67%	74.10%
	T2 (1193 patients)	86.17%	86.17%	85.54%
	T3 (1220 patients)	85.74%	85.73%	85.31%

4.6.2 Validation of Synthetic Data Necessity (M1 Baseline)

The performance of the control model, MLP Real Data Model (M1), serves as the critical validation point for the need for data augmentation. Model M1, trained on only 81 real records, achieved a Macro F1-Score of just 15.17% on the held-out real test set (T1_M1).

This outcome, summarized alongside other metrics in Figure 4-16, confirms that the model suffered from catastrophic failure due to data scarcity and severe bias towards the majority class.

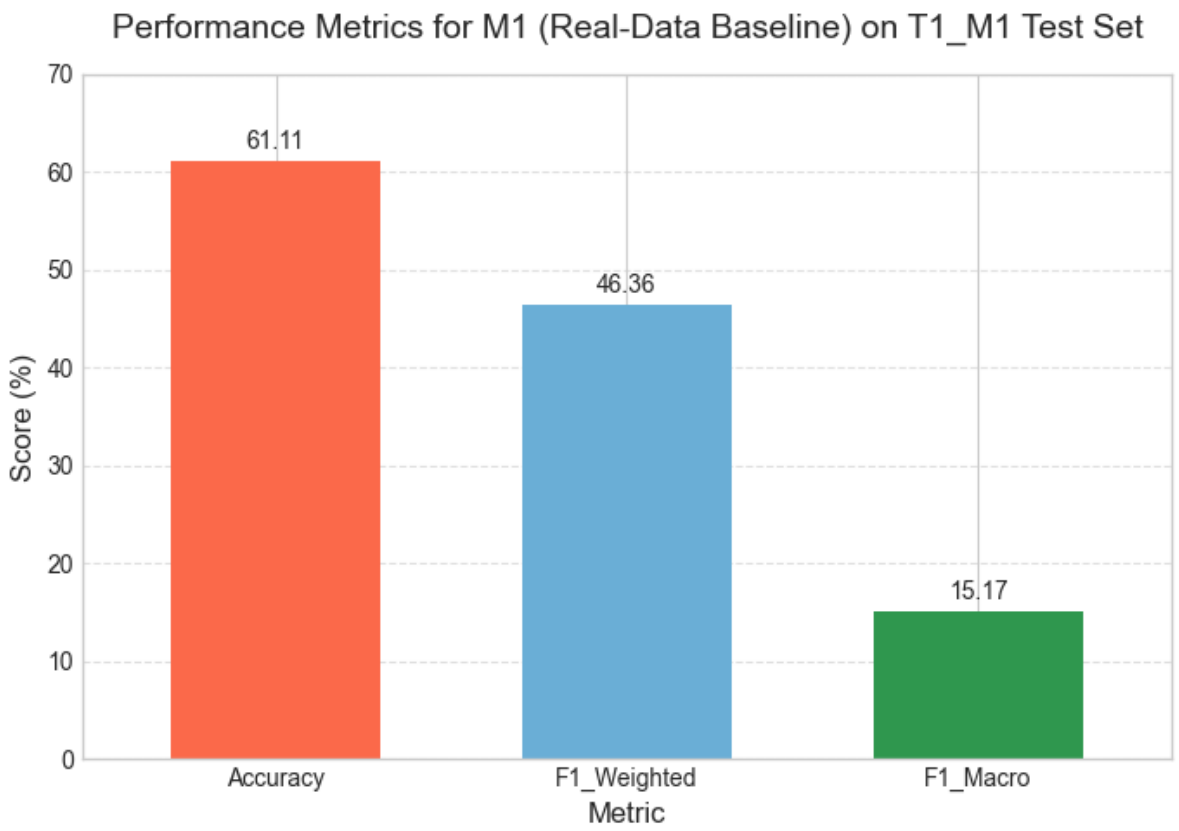


Figure 4-16 Performance Metrics for M1 on T1_M1

The pronounced difference between its relatively high Accuracy (61.11%) and its severely low Macro F1-Score (15.17%) is the clearest evidence of this failure. The model achieved high overall accuracy by finding the statistical center of the data (the majority class), but the low Macro F1-Score proved it failed to learn and generalize any of the rare, critical minority classes.

The resulting confusion matrix presented in figure 4-17 visually confirms that Model M1 failed to predict any class other than the majority class (Class 4 - Good) for all instances, including true minority classes.

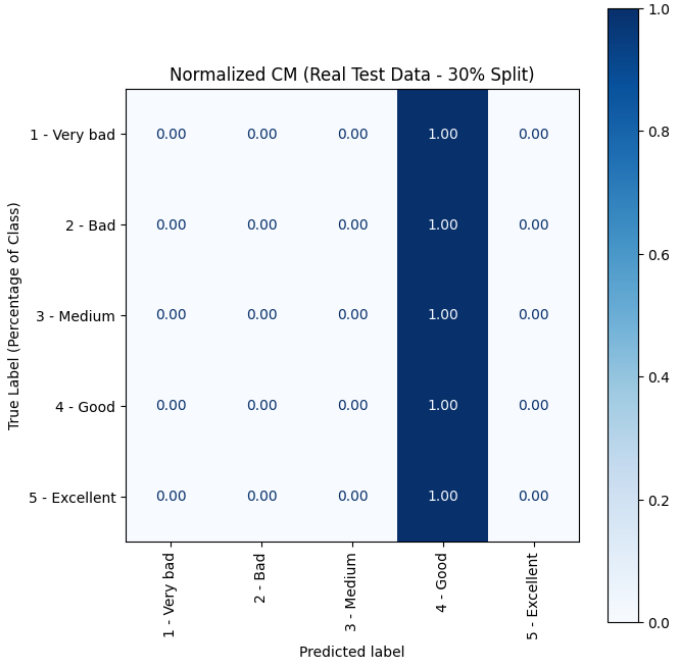


Figure 4-17 Confusion Matrix T_M1 M1

This result strongly affirms the core hypothesis of the study: data augmentation using TimeGAN is mandatory to enable models to learn the feature space necessary for robust multi-class health prediction. This severe performance gap between the real data baseline and the augmented models is empirically validated on both F1-Macro and F1-Weighted in Figure 4-18.

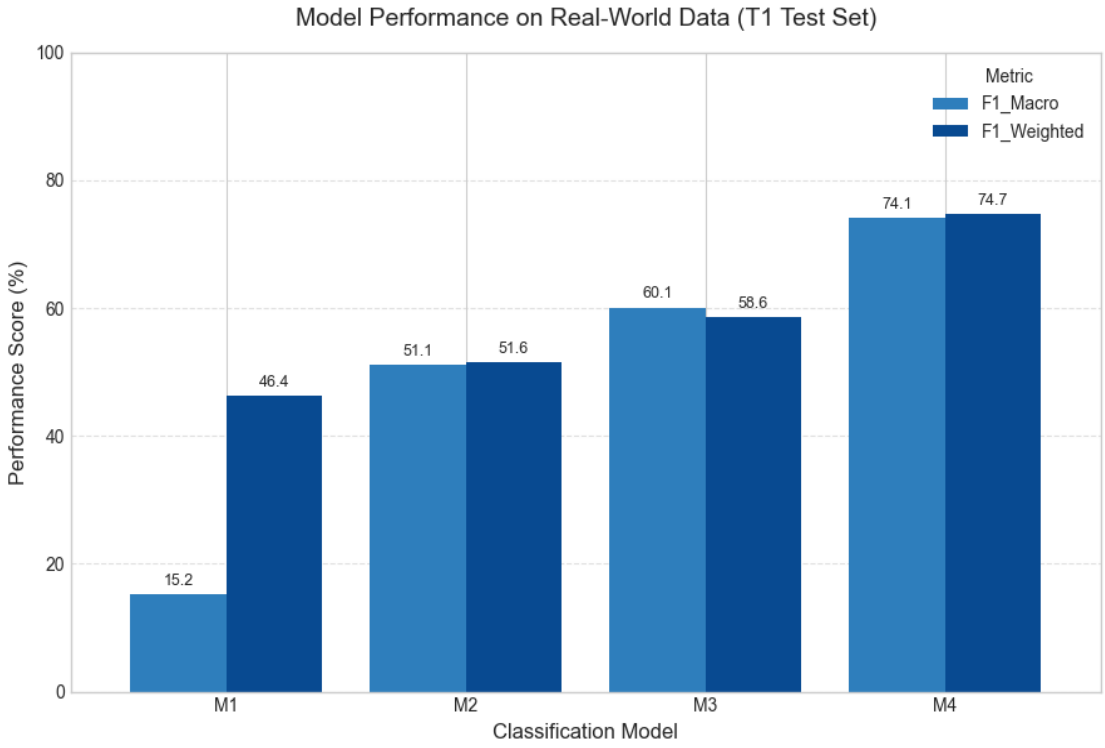


Figure 4-18 F1-Score Generalization Gap between Models in Real World Dataset

Figure 4-18 illustrates the severe performance gap between the Real Data baseline (M1) on the real-world test set (T1_M1) and all TimeGAN augmented models (M2, M3, M4) on the real-world test set (T1), empirically validating the necessity of synthetic data.

4.6.3 Detailed Confusion Matrix Analysis and Class Performance

While the aggregated F1-Macro score quantifies the overall performance across all classes, the analysis of the normalized Confusion Matrices provides critical, fine-grained insight into how each synthetic-trained model handles the minority classes across various test sets, and where specific prediction errors occur. The primary goal of this analysis is to quantify the generalization gap (comparing T2 to T1) and validate the impact of both the balancing strategy and architectural alignment.

1. Comparative Analysis of Synthetic Data (T2)

The performance on the Synthetic Never-Seen Test Set (T2) measures the intrinsic fidelity of the generated data and the model's ability to learn the patterns within the synthetic feature space. The M2 model, trained on imbalanced synthetic data, achieves the highest accuracy (93.79%), demonstrating its ability to perfectly learn the statistical properties of the TimeGAN output. However, its confusion matrix (Figure 4-19) confirms this fidelity comes with an underlying bias in favor of the majority classes.

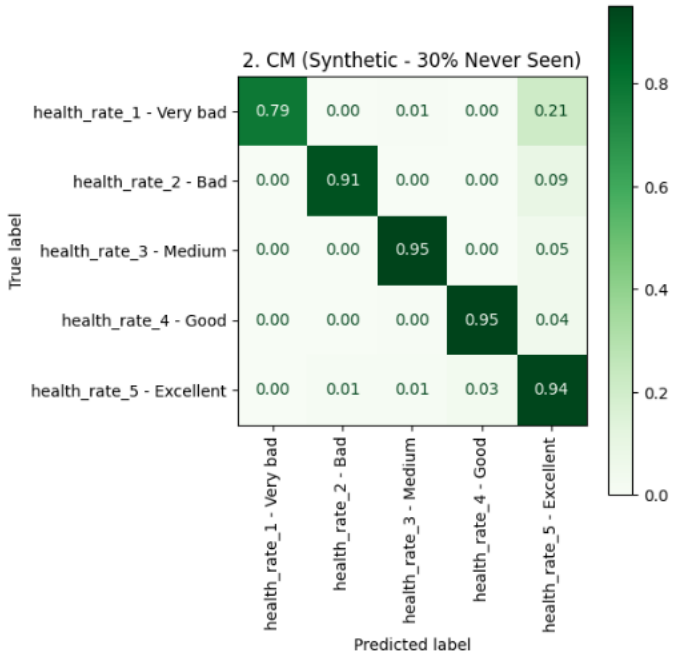


Figure 4-19 Confusion Matrix T2 M2

In comparison, the M3 model (Balanced MLP) achieves a slightly lower but still robust accuracy (89.27%). Its corresponding matrix (Figure 4-20) shows a significantly more uniform True Positive rate across the minority classes, confirming that the balancing step successfully distributed the learning focus without substantially sacrificing data fidelity.

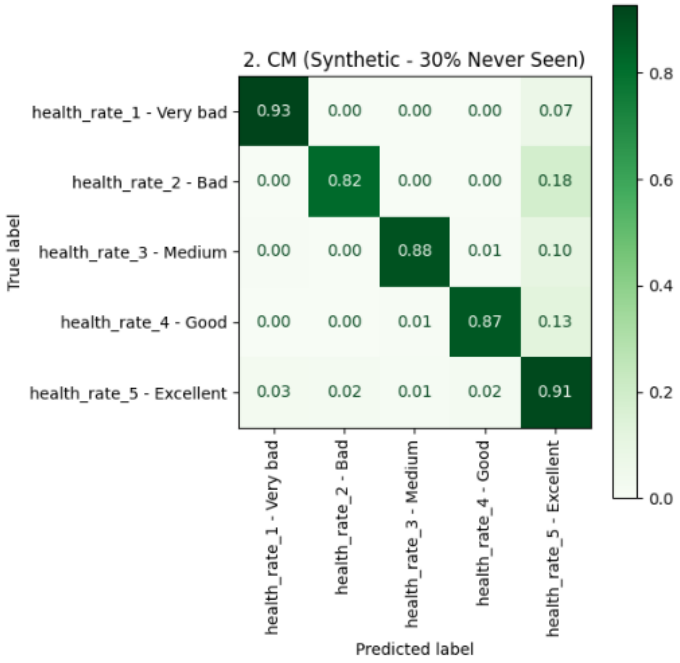


Figure 4-20 Confusion Matrix T2 M3

Finally, the M4 model (Balanced LSTM) registers the lowest accuracy on the synthetic set (86.17%), as evidenced in its confusion matrix (Figure 4-21). This suggests that the sequential LSTM architecture, while robust, may introduce minor smoothing or regularization effects when processing the fixed, generated sequence structure. Collectively, the high accuracy of all models on T2 fundamentally validates that the TimeGAN successfully produced high-fidelity data suitable for deep learning architectures.

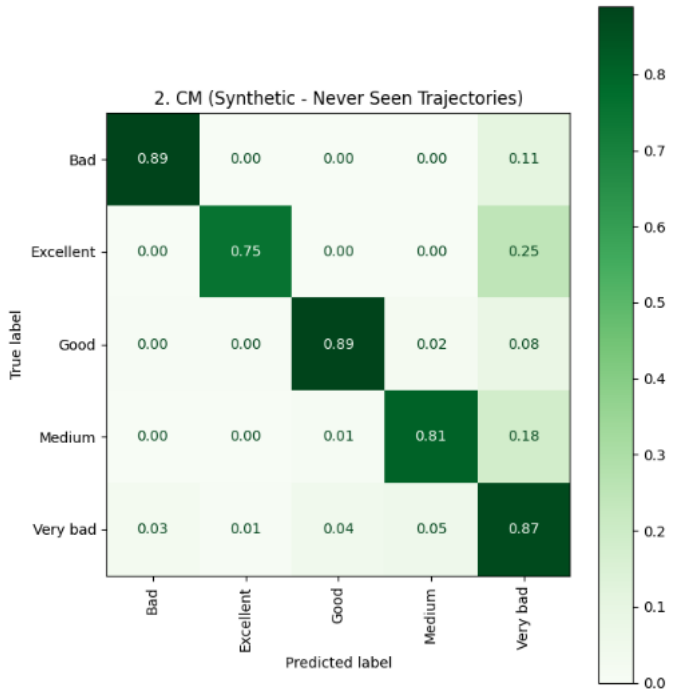


Figure 4-21 Confusion Matrix T2 M4

2. Comparative Analysis in Real Data (T1)

The performance on the All-Real Data Test Set (T1) is the Train Synthetic, Test Real (TSTR) gold standard, measuring the model's real-world generalization from the synthetic patient trajectories. This comparison reveals a marked difference in architectural utility. The M2 model suffers a catastrophic generalization failure, with accuracy dramatically declining to 40.17%. Its normalized confusion matrix (Figure 4-22) shows severe off-diagonal error, demonstrating the MLP failed to generalize the imbalanced synthetic patterns to the noisy, unpredictable real data.

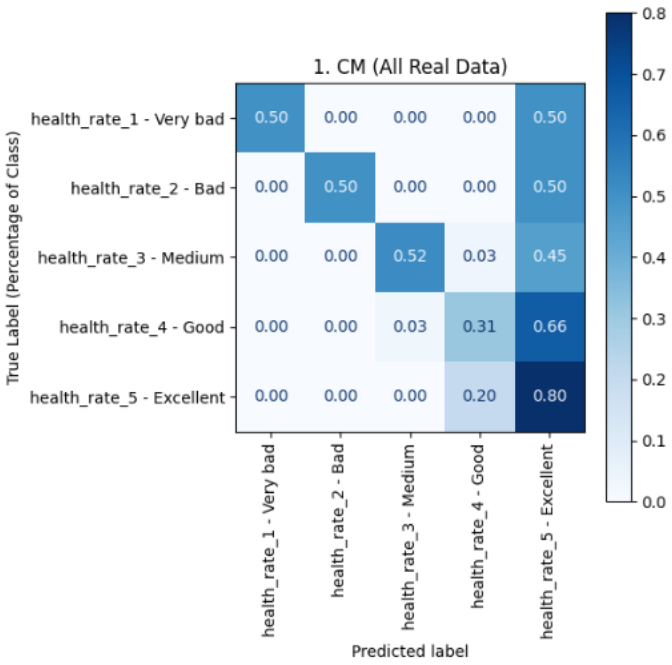


Figure 4-22 Confusion Matrix T1 M2

The M3 model (Balanced MLP) sees a moderate improvement in accuracy to 47.01%. Its matrix (Figure 4-23) confirms the balancing helps mitigate the extreme bias seen in M2 and leads to better performance on minority classes; however, the static MLP still struggles significantly with real-world complexity, unable to leverage the sequential structure.

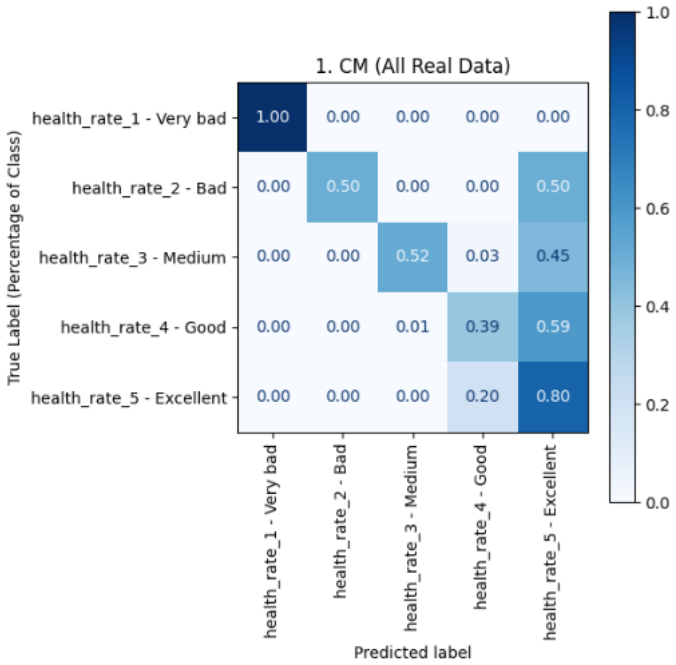


Figure 4-23 Confusion Matrix T1 M3

Significantly, the M4 model (Balanced LSTM) demonstrates superior generalization, with accuracy rising sharply to 66.67%. The normalized matrix for M4 (Figure 4-24) exhibits the cleanest diagonal trend, confirming that the LSTM effectively utilized the temporal dependencies captured by TimeGAN to create a stable decision boundary that generalized robustly to the real patient data. This comparison confirms that the sequential LSTM architecture is demonstrably superior at translating TimeGAN's high-fidelity output into real-world predictive utility.

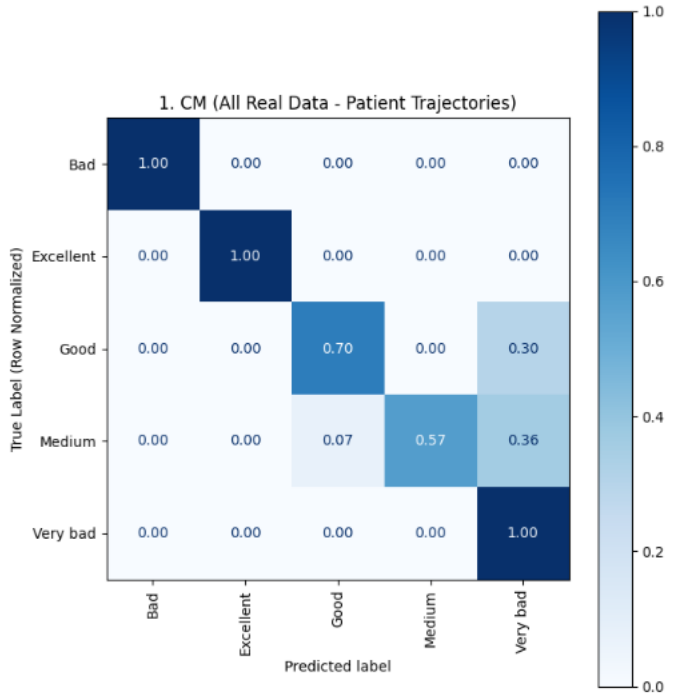


Figure 4-24 Confusion Matrix T1 M4

3. Comparative Analysis on Combined Data (T3)

The Combined Test Set (T3) measures the model's overall stability and robustness across a heterogeneous mix of synthetic and real data. The results largely mirror the trends observed on T2 due to the significantly larger synthetic sample size in the combined set. The M2 model maintains high accuracy (92.44%), but its matrix (Figure 4-25) shows the residual bias toward the majority classes persist, highlighting the structural flaw of ignoring imbalance.

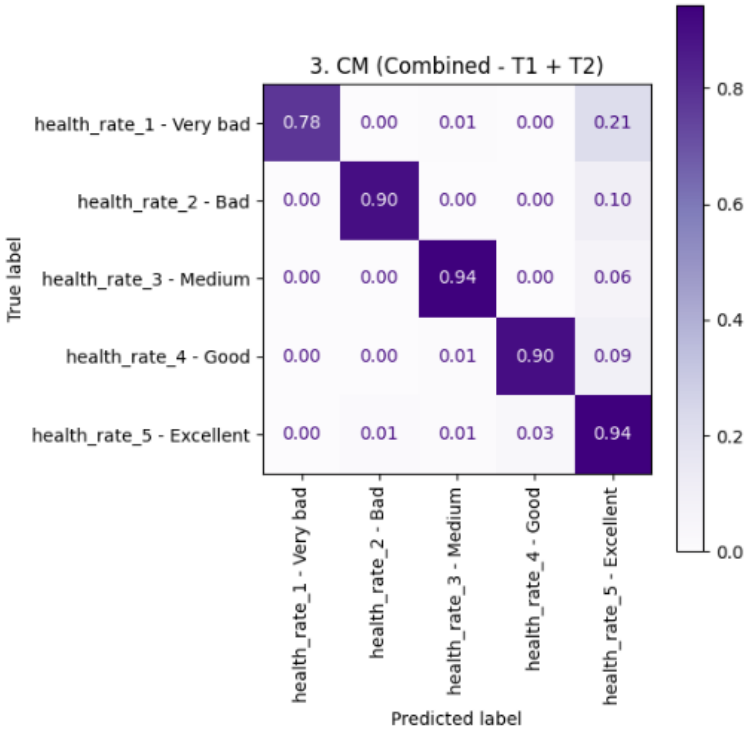


Figure 4-25 Confusion Matrix T3 M2

The M3 model demonstrates stable performance with 88.26% accuracy, and its matrix (Figure 4-26) confirms that the balancing strategy ensures a more even distribution of True Positives, validating the overall robustness of the balancing approach when high-fidelity data is present.

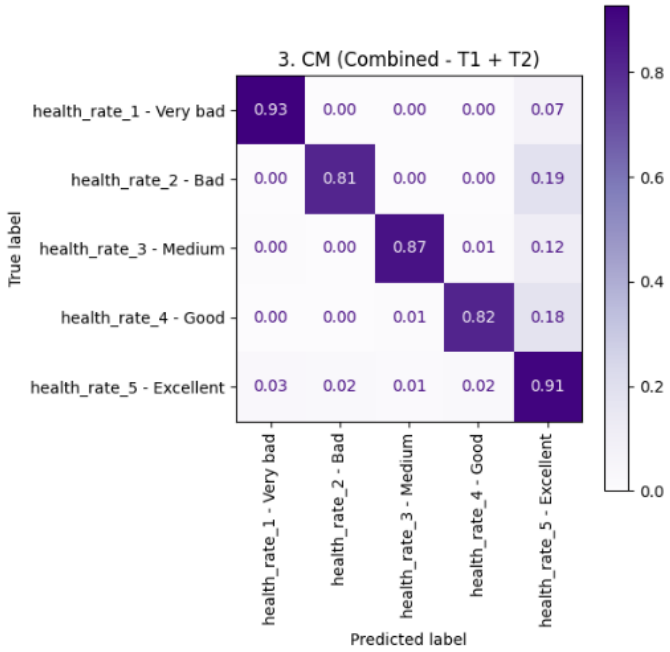


Figure 4-26 Confusion Matrix T3 M3

Finally, the M4 model (LSTM) shows consistent reliability across the combined set, achieving 85.74% accuracy. Its confusion matrix (Figure 4-27) confirms LSTM's consistent ability to model the sequential data, with the overall accuracy being slightly lower than the MLPs, reflecting its greater focus on temporal structure and generalization over numerical peak accuracy on static data points.

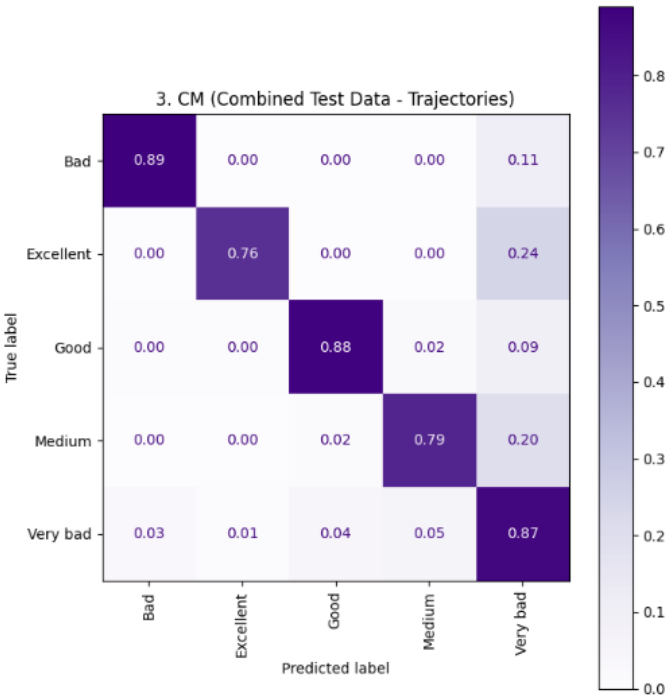


Figure 4-27 Confusion Matrix T3 M4

4.6.4 Impact of Architecture and Temporal Modeling

Comparing the performance of the synthetic-trained models (M2, M3, M4) on the Real Data (T1) highlights the effectiveness of incorporating temporal modeling. The LSTM SynthData Balanced Model (M4), which uniquely processed the data as sequences of patient trajectories, demonstrated the best performance on real-world data. It achieved the highest balanced metrics (Weighted F1-Score of 74.67% and Macro F1-Score of 74.10%). This success indicates that the LSTM effectively utilized the temporal dependencies captured by the TimeGAN, allowing it to generalize more accurately from the synthetic training set to the real patient population.

In contrast, both MLP models (M2 and M3) demonstrated excellent performance on the synthetic test sets (T2), confirming they successfully learned the TimeGAN output (with approximately 93% and 89% accuracy, respectively). However, their generalization to the Real Data (T1) was significantly poorer, yielding the lowest accuracies (40.17% and 47.01%) among the synthetic-trained models, as shown in Figure 4-28.

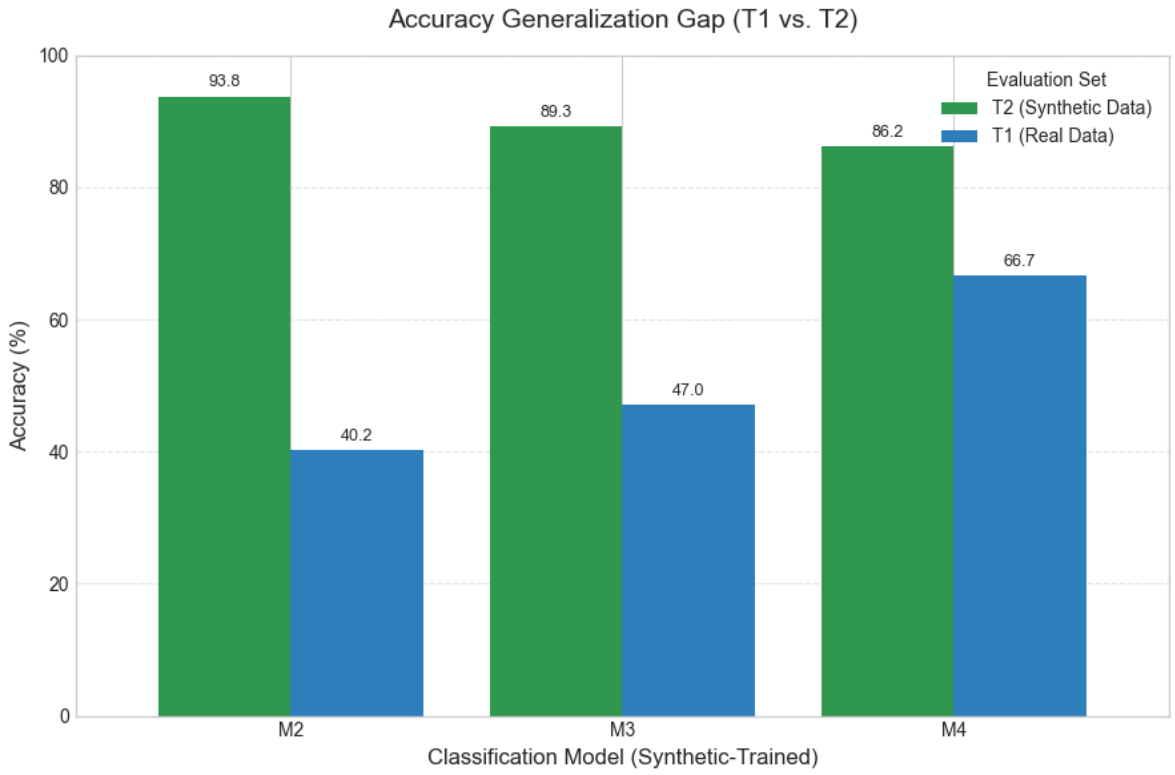


Figure 4-28 Accuracy Generalization Gap of Models (M2-M4)

Figure 4-28 illustrates this pronounced difference. This discrepancy between high performance on synthetic data and poor performance on real data is a strong indicator of overfitting to the specific statistical patterns introduced by the TimeGAN output, which the simpler MLP architecture, lacking sequential modeling capability, failed to make robust for real-world applications. In contrast, the LSTM architecture (M4) maintained a much smaller performance drop, indicating that its ability to model temporal dependencies was key to successful generalization from the synthetic patient trajectories to the real data.

This discrepancy between high performance on synthetic data and poor performance on real data is a strong indicator of overfitting to the specific statistical patterns introduced by the TimeGAN, which the simpler MLP architecture failed to make robust for real-world application. The superior generalization LSTM is quantified by the difference in performance in real test set, as shown in Figure 4-29.

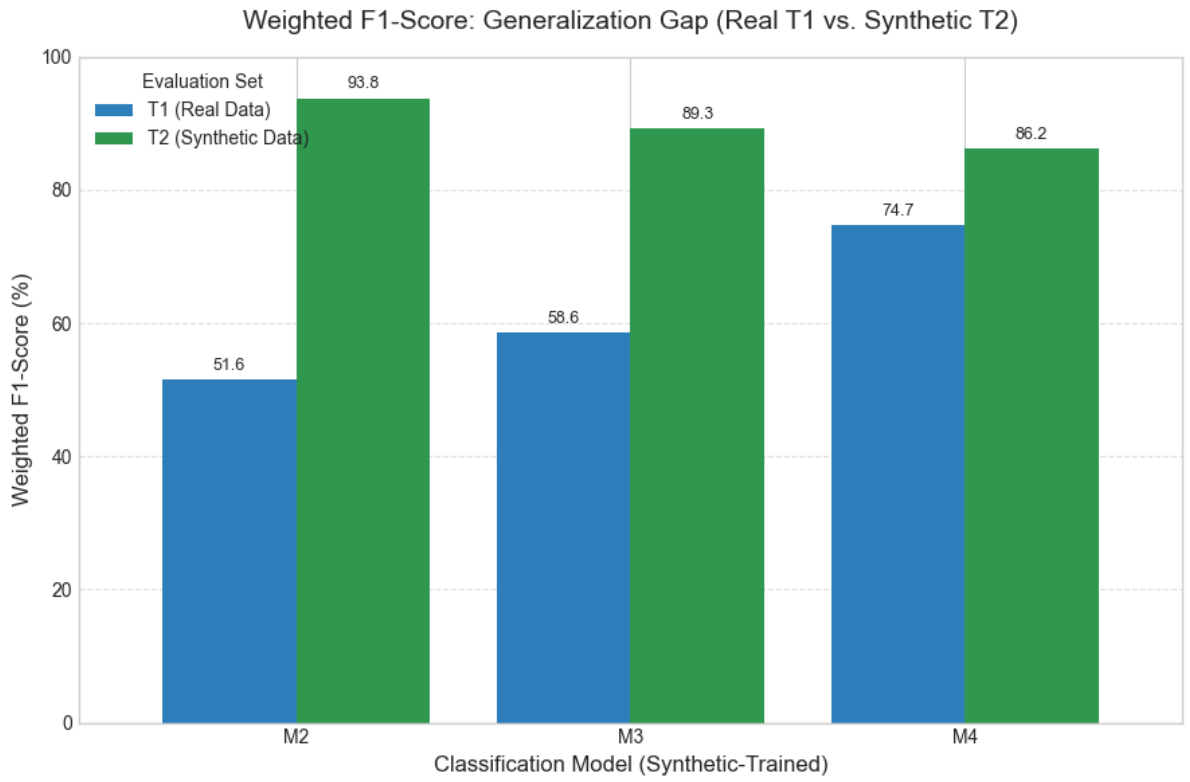


Figure 4-29 Weighted F1 Score Generalization Gap of Models (M2-M4)

Figure 4-29 visualizes the Generalization Gap by comparing the Weighted F1-Score of the synthetic-trained models (M2, M3, M4) on the synthetic test set (T2) versus the complete real test set (T1) and quantifies the architectural dependency, showing the large drop-off for static MLPs (M2, M3) compared to the smaller, more robust generalization gap exhibited by the sequential LSTM (M4).

4.6.5 Effect of Balancing Strategy (M2 vs. M3)

The direct comparison between the two MLP models trained on the imbalanced (M2) and balanced (M3) synthetic datasets yielded two key observations. The balanced model (M3) demonstrated a small but notable improvement in real-world generalization (Macro F1: 60.11%) compared to the imbalanced model (M2) (Macro F1: 51.12%). This suggests that synthetic data balancing, while imperfect, has provided the model with enough exposure to minority classes to achieve better, more balanced prediction on the real data. Furthermore, both models maintained high and comparable performance on the Synthetic Test Set (T2), indicating that the synthetic data itself possessed high fidelity regardless of the balancing step.

The overall findings strongly support the use of synthetic data augmentation combined with temporal deep learning architecture as the most effective strategy for multi-class health prediction in this challenging, data-scarce environment.

5 Discussion

The discussion chapter provides an interpretation of the empirical findings presented in Chapter 4, evaluating the results within the context of the study's primary objectives: mitigating data scarcity using Time-series Generative Adversarial Network (TimeGAN) and comparing the effectiveness of sequential versus static deep learning architectures for longitudinal clinical prediction, primarily assessed via Macro F1-score, Weighted F1-score, and Accuracy.

5.1 The Necessity of Synthetic Data

The most compelling finding of this study is the critical requirement for data augmentation via TimeGAN to achieve a viable predictive model, fundamentally addressing the problem of data scarcity in this specialized clinical domain.

The performance of the Multi-Layer Perceptron (MLP) Real Data Model (M1), which was trained exclusively on the limited 81 real patient records, serves as the definitive baseline. Its Macro F1-Score of 15.17% on the real test set (T_M1), as documented in Table 4.8, confirms the study's initial premise: the raw, scarce, and imbalanced clinical data is insufficient to train a model capable of multi-class generalization. The poor Macro F1-Score of 15.17% contrasted with its high Accuracy (61.11%), is a clear indicator that the model exhibited severe bias toward the majority class, effectively failing to predict any instance of the minority classes (Health Rates 1 and 2). This renders it clinically unusable for identifying at-risk patients.

In contrast, the immediate viability of all synthetic-trained models (M2, M3, M4) which achieved significantly higher Macro F1-Scores on real data demonstrates that TimeGAN successfully provided the necessary feature density and class representation to learn the underlying predictive structure. This outcome is visually summarized by the severe performance gap shown in Figure 4.18, which validates the utility of Time-series GAN for Longitudinal Tabular Data as an effective, data-agnostic solution for mitigating data scarcity in specialized clinical research environments.

5.2 Architectural Evaluation

The comparison between the static Multi-Layer Perceptron (MLP) and the sequential Long Short-Term Memory network (LSTM) provides the critical validation for the suitability of the TimeGAN framework for sequence prediction.

The LSTM SynthData Balanced Model (M4) achieved the best overall performance on the real patient test set (T1), with a Weighted F1-Score of 74.67% and a Macro F1-Score of 74.10%. This superior generalization can be directly attributed to the LSTM's inherent ability to capitalize on the temporal coherence preserved by TimeGAN’s training objectives, specifically the Supervised Loss (L_S). The LSTM is uniquely positioned to interpret the sequential dependencies synthesized by TimeGAN, leading to robust generalization.

In contrast, the MLP models (M2 and M3) showed a significant performance gap on real data (Weighted F1-scores between 51.56% and 58.65%) despite achieving high scores on the synthetic test set (T2) (up to 93.78%). This indicates the models overfit to the static features

of the synthetic data. The MLP, unable to leverage the sequential context, learned non-robust correlations, confirming that for longitudinal health outcomes, temporal modeling (LSTM) is essential for realizing the full predictive potential of TimeGAN-generated data.

5.3 Analysis of the Balancing Strategy and Data Fidelity

The outcomes of the imperfect balancing strategy applied to the synthetic data provide insights into the fidelity limits of the generative process, particularly concerning severe minority classes.

The comparison between the imbalanced (M2) and balanced (M3) MLP models revealed that the improvement in multi-class performance gained by the balancing step was substantial: 9 points increase in Macro F1-score (60.11% for M3 vs. 51.12% for M2). This confirms that the filtering successfully provided better, if still limited, exposure to minority classes, which is key to improving generalization. However, this gain was marginal compared to the architectural advantage of the LSTM (M4).

The root cause for the imperfect class equalization and marginal performance improvement is likely a limitation in the generation quality: TimeGAN struggles to synthesize high-quality, diverse sequences for the severe minority classes (Health Rates 1 and 2). This challenge highlights a residual difficulty in generating truly novel and distinct time-series for rare clinical states, despite the overall success of the method.

5.4 Alignment with Related Work, Clinical Relevance, and Future Work

These findings align with related work that confirm that LSTMs are inherently superior to MLPs for time-series forecasting due to their architecture's superior ability to model dependencies. Fundamentally, this validates the use of the supervised loss (L_S) within TimeGAN as a critical component for clinical relevance, confirming that the synthesized data is reliable for training sequence-aware decision support systems.

The final model, M4 (LSTM with TimeGAN-Balanced Data), provides a prediction capability that was unattainable using the original scarce dataset. Achieving a Macro F1-Score of 74.10% on real, unseen patient data represents a viable starting point for a system designed to predict frailty progression over time. Future work should focus on refining the conditional generation process within TimeGAN itself to improve the diversity and fidelity of the minority classes, potentially pushing the prediction metrics into a clinically acceptable range.

5.5 Conclusion

In summary, the discussion confirms that synthetic data augmentation using TimeGAN is an important first step for multi-class prediction in this data-scarce longitudinal setting. Furthermore, the performance metrics clearly dictate that when TimeGAN is used, the synthetic sequences must be leveraged by a sequential deep learning architecture (LSTM) to translate the preserved temporal features into robust, real-world predictive generalization.

6 Conclusion and Future Work

This dissertation successfully validated the use of the Time-series Generative Adversarial Network (TimeGAN) for longitudinal tabular data as a reliable and necessary strategy for multi-class health prediction in a severely data-scarce clinical environment. The study rigorously addressed the core problem of model fragility caused by limited and imbalanced patient records through a comparative experimental design. The research conclusively met all primary objectives.

The catastrophic failure of the control model, which achieved only a 15.17% Macro F1-Score on real data, empirically confirmed the mandatory requirement for synthetic data augmentation. The TimeGAN framework successfully generated synthetic datasets that preserved the critical structural and temporal fidelity of the original data, immediately demonstrating viability in all augmented models. Furthermore, the comparative analysis established a clear architectural dependency: the static MLP models exhibited significant overfitting and failed to generalize robustly. The optimal configuration was the LSTM Model (M4), achieving the highest real-world generalization (Weighted F1-Score: 74.67%).

This outcome conclusively proves that sequential architecture (LSTM) is essential for effectively utilizing the temporal coherence encoded by TimeGAN's supervised loss component. In summary, this research concludes that the effectiveness of synthetic data for multi-class prediction is fundamentally dependent on the prediction model's ability to process the data's inherent sequential nature. The coupling of TimeGAN with the LSTM network provides the most robust and performant path forward for developing clinically useful predictive systems in data-limited domains.

This work offers an Empirical Validation of TimeGAN Utility and quantifies the architectural dependency, establishing a validated, end-to-end methodology for transforming severely scarce, imbalanced longitudinal datasets into a viable training resource. While highly successful, the study encountered two primary limitations: imperfect class balancing, due to the TimeGAN Generator's inherent difficulty in producing diverse, high-fidelity samples for the severe minority classes, and the reliance on a fixed four-visit trajectory.

Future work should prioritize Advanced Conditional Generation, exploring conditional TimeGAN or alternative conditional diffusion models to improve minority class fidelity, and should incorporate Variable Sequence Length Testing to further validate the LSTM’s robustness. Finally, techniques like Permutation Feature Importance (PFI) and SHapley Additive exPlanations (SHAP) should be applied to the optimal M4 LSTM model to transition the system from a predictive tool to a diagnostic support tool, enhancing clinical applicability.

References / Bibliography

[1] E. C. J. S. Cao Xiao, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *Journal of the American Medical Informatics Association, Volume 25,* p. 1419–1428, 10 October 2018.

[2] J. K. M. L. S. P. E. C. Eunbyeol Cho, "Generating Multi-Table Time Series EHR from Latent Space with Minimal Preprocessing," 9 July 2025.

[3] P. P. L. A. H. B. Jing Zhao, "Learning from heterogeneous temporal data in electronic health records," *Journal of Biomedical Informatics,* vol. 101, January 2020.

[4] A. P. S. H. Z. E. Y. M. V. T. P. R. G. K. Alex Labach, "DuETT: Dual Event Time Transformer for Electronic Health Records," 2023.

[5] S. M. H. S. F. B. MohammadReza EskandariNasab, "SeriesGAN: Time Series Generation via Adversarial and Autoregressive Learning," 28 October 2024.

[6] K. E. A. S. T. R. Emmanuella Budu, "Evaluation of synthetic electronic health records: A systematic review and experimental assessment," *Neurocomputing,* vol. 603, 28 October 2024.

[7] [Online]. Available: https://ydata.ai/resources/synthetic-time-series-data-a-gan-approach.html. [Accessed 2025].

[8] T. W. E. B. John Weldon, "Generation of Synthetic Electronic Health Records Using a Federated GAN," 6 September 2021.

[9] M. D. L. Ofer Mendelevitch, "Fidelity and privacy of synthetic medical data," 2 June 2021.

[10] R. P. S. B. S. J. C. L. S. A. Goncalves A., "Generation and evaluation of synthetic patient data," vol. 20, no. 108, 7 May 2020.

[11] Y. H. J. X. R. D. G. M. T. M. Z. Q. X. L. Lee D., "Generating sequential electronic health records using dual adversarial autoencoder," vol. 27, no. 9, pp. 1411-1419, September 2020.

[12] J. J. J. v. d. S. M. Yoon, "Time-series Generative Adversarial Networks," in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, 2019.

[13] H. A. M. K. N. M. G. Z. S. B. A. Murtaza, "Synthetic data generation: State of the art in health care domain," *Computer Science Review,* vol. 48, no. May 2023, 100546, 2023.

[14] R. M. B. Hilleke E. Hulshoff Pol, "Unique opportunities and challenges of longitudinal approaches in studying brain health and mental health," *NeuroView,* vol. 113, no. 12, 18 June 2025.

[15] F. P. M. A. &. A. G. Mohammad Loni, "A review on generative AI models for synthetic medical text, time series, and longitudinal data," *npj Digital Medicine,* 15 May 2025.

[16] C. O. C. V. B. N. T. B. R. Mauricio C. Cordeiro1, "Gait stability prediction through synthetic time-series and vision-based data," 2025.

[17] J. M. Q. Y. Z. W. L. B. X. Y. Shuai Niu, "Modelling Patient Longitudinal Data for Clinical Decision Support: A Case Study on Emerging AI Healthcare Technologies," 18 July 2024.

[18] H. Y. Y. N. M. E. H. O. M. F. W. H. B. C. N. L. Feng Xiea, "Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies," *JournalofBiomedicalInformatics,* January 2022.

[19] J. Y. M. X. C. M. F. Luo, "Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records".

[20] M. K. T. H. S. F. Yilmazcan Özyurt, "AttDMM: An Attentive Deep Markov Model for Risk Scoring in Intensive Care Units," 17 February 2021.

[21] V. R. A. V. M. P. S. K. A. G. P. M. Hans Moen, "Towards modeling evolving longitudinal health trajectories with a transformer-based deep learning model".

[22] N. S. N. P. J. U. L. J. A. N. G. L. K. I. P. Ashish Vaswani, "Attention Is All You Need," 2 August 2023.

[23] A. R. A. R. B. e. a. Esteva, "A guide to deep learning in healthcare," vol. 25, 17 October 2018.

[24] "Synthetic data generation for tabular health records: A systematic review," *Neurocomputing,* vol. 493, pp. 28-45, 7 July 2022.

[25] J. L. T. Z. Ghadeer O. Ghosheh, "A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records," vol. 56, no. 6, 2024.

[26] H. Y. X. J. D. R. M. G. M. T. Q. Z. L. X. Dongha Lee, "Generating sequential electronic health records using dual adversarial autoencoder," 1 July 2020.

[27] S. L. H. G. R. Cristóbal Esteban, "Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs," 4 December 2017.

[28] W. Z. &. X. H. Lu Wang, "Continuous Patient-Centric Sequence Generation via Sequentially Coupled Adversarial Learning," in *Database Systems for Advanced Applications* , 2019.

[29] R. V. N. E. N. P. T. Alexandre Yahi, "Generative Adversarial Networks for Electronic Health Records: A Framework for Exploring and Evaluating Methods for Predicting Drug-Induced Laboratory Test Trajectories," 1 December 2017.

[30] Z. Z. S. N. B. A. M. Chao Yan, "Generating Electronic Health Records with Multiple Data Types and Constraints," 23 March 2020.

[31] B. J. C. J. L. T. Z. Jin Li, "Generating Synthetic Mixed-type Longitudinal Electronic Health Records for Artificial Intelligent Applications," 31 January 2023.

[32] M. D. L. Ofer Mendelevitch, "Fidelity and Privacy of Synthetic Medical Data," 2 June 2021.

[33] J. C. B. L. J. e. a. Li, "Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications," *Digital Medicine,* no. Article number: 98, 27 May 2023.

[34] A. T. D. G. Bilici Ozyigit E., "Generation of realistic synthetic validation healthcare datasets using generative adversarial networks," 2020.

[35] J. O. N. W. Z. M. P. T. A. de Benedetti, "Practical lessons from generating synthetic healthcare data with Bayesian networks," in *Communications in Computer and Information Science*, 2020.

[36] S. N. V. I. T. B. S. H. B. A. Venugopal R., "Privacy preserving generative adversarial networks to model electronic health records," *Neural Networks,* vol. 153, September 2022.

[37] T. S. J. E. V. Kieran Chin-Cheong, "Generation of differentially private heterogeneous electronic health records," June 2020.

[38] Y. C. M. D. S. J. M. B. Zhang Z., "Ensuring electronic medical record simulation through better training, modeling, and evaluation," vol. 27, no. 1, pp. 99-108, January 2020.

[39] D. S. D. R. G. I. P. A. B. K. Yale A., "Generation and evaluation of privacy preserving synthetic health data," vol. 416.

[40] Y. C. L. T. S. J. M. B. Zhang Z., "SynTEG: a framework for temporal structured electronic health data simulation," vol. 28, no. 3, pp. 596-604, March 2021.

[41] H. G. Maaten L.v.d., "Visualizing data using t-SNE," 2008.

[42] S. G. J. D. B. M. W. S. J. S. Siddharth Biswal, "Generating Longitudinal Electronic Health Records Using Conditional Variational Autoencoders," December 2020.

[43] A. T. D. G. Bilici Ozyigit E., "Generation of realistic synthetic validation healthcare datasets using generative adversarial networks," *Studies in Health Technology and Informatics,* 2020.

[44] C. J. Jolliffe I.T., "Principal component analysis: a review and recent developments," vol. 374, no. 2065, 13 April 2016.

[45] H. J. M. J. McInnes L., "UMAP: Uniform manifold approximation and projection for dimension reduction," 18 September 2020.

[46] K. J.B., "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," vol. 29, no. 1, pp. 1-27, March 1964.

[47] P. Yadav, M. Gaur, N. Fatima and S. Sarwar, "Qualitative and Quantitative Evaluation of Multivariate," *Applied Sciences,* 24 March 2023.

[48] Z. L. X. W. X. L. M. Peihao Tang, "Time Series Data Augmentation for Energy Consumption Data Based on Improved TimeGAN," *sensors,* vol. 25, no. 2, 16 January 2025.

[49] R. Shamsuddin, B. M. Maweu, M. Li and B. Prabhakaran, "Virtual Patient Model: An Approach for Generating Synthetic Healthcare Time Series Data," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 2018.

[50] A. K. A. M. T. Al-Fakih, "Well log data generation and imputation using sequence based generative adversarial networks," *Scientific Reports,* no. 11000, 31 March 2025.

[51] R. a. Z. W. a. S. I. Jozefowicz, "An Empirical Exploration of Recurrent Network Architectures," vol. 37, July 2015.

[52] D. A. D. O. A. A. &. S. V. Mabrouka Salmi, "Handling imbalanced medical datasets: review of a decade of research," vol. 57, no. 273, September 2024.

[53] F. B. F. L. Georgios Douzas, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," vol. 465, pp. 1-20, October 2018.

[54] D. S. T. N. Y. K. Zhaozhao Xu, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data," vol. 107, July 2020.

[55] R. A. Varsha Babar, "A Novel Approach for Handling Imbalanced Data in Medical Diagnosis using Undersampling Technique," vol. 5, no. 7, July 2016.

[56] S. R. D. K. Anju Jain, "A novel multi-objective genetic algorithm approach to address class imbalance for disease diagnosis," vol. 15, pp. 1151-1166, 2023.

[57] M. Phankokkruad, "Cost-Sensitive Extreme Gradient Boosting for Imbalanced Classification of Breast Cancer Diagnosis," pp. 46-51, 2020.

[58] T. B. T. Q. M. B. H. H. L. T. P. L. a. N. C. T. Hai Thanh Nguyen, "Enhancing Disease Prediction on Imbalanced Metagenomic Dataset by Cost-Sensitive," vol. 11, no. 7, 2020.

[59] K. K. X.-Z. G. &. D. S. R. MadhuSudana Rao Nalluri, "Multiobjective hybrid monarch butterfly optimization for imbalanced disease classification problem," vol. 11, pp. 1423-1451, 2020.

[60] F. B. R. Walid Ksiaa, "Tuning Hyperparameters on Unbalanced Medical Data Using Support Vector Machine and Online and Active SVM," pp. 1134-1144, 2021.

[61] R. W. Y. L. W.-H. L. H. C. J. C. Huan Zhao, "Severity level diagnosis of Parkinson's disease by ensemble K-nearest neighbor under imbalanced data," *Expert Systems with Applications,* vol. 189, no. 116113, March 2022.

[62] T. Q. G. V. P. M. v. d. A. B. D. M. Xi Shi, "A Resampling Method to Improve the Prognostic Model of End-Stage Kidney Disease: A Better Strategy for Imbalanced Data," vol. 9, 2022.

[63] B. Z. F. W. X. L. Min Zeng, "Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data," 2016.

[64] D. M. G. P. S. C. S. Adyasha Rath, "Heart disease detection using deep learning methods from imbalanced ECG samples," *Biomedical Signal Processing and Control,* vol. 68, no. 102820, July 2021.

[65] J. W. L. Yawen Xiao, "Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data," *Computers in Biology and Medicine,* vol. 135, August 2021.

[66] R. K. S. J. K. B. R. S. J. S. Klaus Greff, "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 28, no. 10, 10 October 2017.

[67] B. v. M. C. G. D. B. F. B. H. S. Y. B. Kyunghyun Cho, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," 3 September 2014.

[68] C. G. K. C. Y. B. Junyoung Chung, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," 11 December 2014.

[69] J. W. C.-L. Z. Z.-H. Z. Guo-Bing Zhou, "Minimal Gated Unit for Recurrent Neural Networks," 31 March 2016.

[70] K. Deltouzos, "Aggregated Virtual Patient Model Dataset," 2019.

[71] L. E. E. K. D. L. e. a. Mosquera, "A method for generating synthetic longitudinal health data," *BMC Medical Research Methodology,* no. Article number: 67, 2023.

[72] C.-W. L. S. C. H. Der-Chiang Li, "A learning method for the class imbalance problem with medical data sets," *Computers in Biology and Medicine,* vol. 40, no. 5, 26 March 2010.

[73] G. S. V. V. Pankaj Yadav, "Rebalancing the Scales: A Systematic Mapping Study of Generative Adversarial Networks (GANs) in Addressing Data Imbalance," 23 February 2025.

[74] N. P. M. F. S. e. a. Kuo, "The Health Gym: synthetic health-related datasets for the development of reinforcement learning algorithms," *Scientific Data,* vol. 9, no. 693, 11 November 2022.

[75] D. C. K. C. E. R. W. Zachary C. Lipton, "Learning to Diagnose with LSTM Recurrent Neural Networks," 21 March 2017.

Appendix

Appendix A - Feature Set Overview

This appendix provides a detailed description of the raw longitudinal patient data utilized in this study. The dataset comprises sequential records of patient visits, capturing various facets of health, physical function, cognition, social activity, and lifestyle. The full dataset consists of 58 original features, as detailed below [70].

Feature Name	Description	Data Type
part_id	The user ID, which should be a 4-digit number.	Identifier
q_date	The recording timestamp (YYYY-MM-DDTHH:mm:ss.fffZ format).	Date/Time
clinical_visit	The sequential number indicating which clinical evaluation these measurements refer to.	Discrete Numerical
fried	Ordinal categorization of frailty level according to Fried operational definition of frailty.	Ordinal Categorical
hospitalization_one_year	Number of nonscheduled hospitalizations in the last year.	Discrete Numerical
hospitalization_three_years	Number of nonscheduled hospitalizations in the last three years.	Discrete Numerical
ortho_hypotension	Presence of orthostatic hypotension.	Binary Categorical
vision	Visual difficulty (qualitative ordinal evaluation).	Ordinal Categorical
audition	Hearing difficulty (qualitative ordinal evaluation).	Ordinal Categorical
weight_loss	Unintentional weight loss >4.5 kg in the past year.	Binary Categorical
exhaustion_score	Self-reported exhaustion.	Binary Categorical
raise_chair_time	Time in seconds to perform a lower limb strength clinical test.	Continuous Numerical
balance_single	Single foot station (Balance).	Binary Categorical
gait_get_up	Time in seconds to perform the 3-meter Timed Get Up And Go Test.	Continuous Numerical
gait_speed_4m	Speed for 4 meters' straight walk.	Continuous Numerical
gait_optional_binary	Gait optional evaluation (qualitative evaluation by the investigator).	Binary Categorical
gait_speed_slower	Slowed walking speed.	Binary Categorical
grip_strength_abnormal	Grip strength outside the norms.	Binary Categorical
low_physical_activity	Low physical activity.	Binary Categorical
falls_one_year	Number of falls in the last year.	Discrete Numerical
fractures_three_years	Number of fractures during the last 3 years.	Discrete Numerical
fried_clinician	Fried's categorization according to the clinician's estimation.	Ordinal Categorical
bmi_score	Body Mass Index (in Kg/m^2).	Continuous Numerical
bmi_body_fat	Body Fat (%).	Continuous Numerical
waist	Waist circumference (in cm).	Continuous Numerical
lean_body_mass	Lean Body Mass (%).	Continuous Numerical
screening_score	Mini Nutritional Assessment (MNA) screening score.	Discrete Numerical
cognitive_total_score	Montreal Cognitive Assessment (MoCA) test score.	Discrete Numerical
memory_complain	Memory complaint.	Binary Categorical
mmse_total_score	Folstein Mini-Mental State Exam score.	Discrete Numerical
sleep	Reported sleeping problems (qualitative ordinal evaluation).	Ordinal Categorical
depression_total_score	15-item Geriatric Depression Scale (GDS-15).	Discrete Numerical
anxiety_perception	Anxiety auto-evaluation (visual analogue scale 0-10).	Continuous Numerical
living_alone	Living Conditions.	Binary Categorical
leisure_out	Leisure activities (number of leisure activities per week).	Discrete Numerical
leisure_club	Membership of a club.	Binary Categorical
social_visits	Number of visits and social interactions per week.	Discrete Numerical
social_calls	Number of telephone calls exchanged per week.	Discrete Numerical
social_phone	Approximate time spent on the phone per week.	Continuous Numerical
social_skype	Approximate time spent on videoconference per week.	Continuous Numerical
social_text	Number of written messages (SMS and emails) sent by the participant per week.	Discrete Numerical
house_suitable_participant	Subjective suitability of the housing environment according to participant's evaluation.	Binary Categorical
house_suitable_professional	Subjective suitability of the housing environment according to the investigator's evaluation.	Binary Categorical
stairs_number	Number of steps to access house (without possibility to use elevator).	Discrete Numerical
life_quality	Quality of life self-rating (visual analogue scale 0-10).	Continuous Numerical
health_rate	Self-rated health status (qualitative ordinal evaluation).	Ordinal Categorical
health_rate_comparison	Self-assessed change since last year (qualitative ordinal evaluation).	Ordinal Categorical
pain_perception	Self-rated pain (visual analogue scale 0-10).	Continuous Numerical
activity_regular	Regular physical activity (ordinal answer).	Ordinal Categorical
smoking	Smoking.	Binary Categorical
alcohol_units	Alcohol Use (average alcohol units consumption per week).	Continuous Numerical
katz_index	Katz Index of ADL score.	Discrete Numerical
iadl_grade	Instrumental Activities of Daily Living score.	Discrete Numerical
comorbidities_count	Number of comorbidities.	Discrete Numerical
comorbidities_significant_count	Number of comorbidities which significantly affect the person's functional status.	Discrete Numerical
medication_count	Number of active substances taken on a regular basis.	Discrete Numerical

Source: https://zenodo.org/records/2670048

Appendix B - Outliers Analysis

This appendix details the empirical results and justification for the values identified and converted to null values (**NaN**) during the Outlier Remediation step (Section 4.2.2). The methodology employed a Z-score threshold of 3.0 to identify extreme numerical outliers.

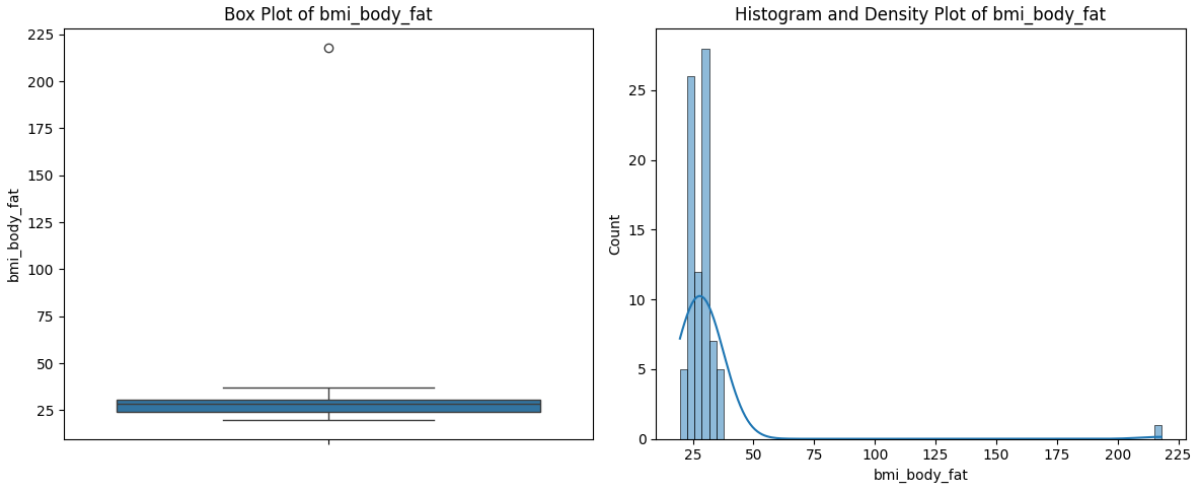
The following table summarizes the identified outliers across the numerical features. Each entry includes the subject identifier (*part_id*), the time point (*clinical_visit*), the affected feature (Outlier Column), and the original value (Outlier Value) that was converted to NaN.

.

Index	part_id	clinical_visit	Outlier Column	Outlier Value
0	2085	3	bmi_body_fat	218.000000
1	1088	2	bmi_score	44.414062
2	1088	3	bmi_score	44.375000
3	1088	4	bmi_score	44.658044
4	1104	1	gait_get_up	999.000000
5	1104	1	gait_speed_4m	22.800000
6	1104	3	gait_speed_4m	20.400000
7	2584	1	gait_speed_4m	18.000000
8	2584	3	gait_speed_4m	18.000000
9	2085	3	lean_body_mass	-78.706000
10	1090	1	raise_chair_time	999.000000
11	1104	1	raise_chair_time	999.000000
12	1104	2	raise_chair_time	999.000000
13	1104	3	raise_chair_time	999.000000
14	1104	4	raise_chair_time	999.000000
15	2087	4	waist	188.000000

The following sections provide a detailed, feature-specific assessment for each distinct outlier identified. Each assessment includes the statistical location, distribution visualization (Boxplot and Histogram), and the contextual patient data sequence necessary for data validation rationale regarding the value's conversion to NaN.

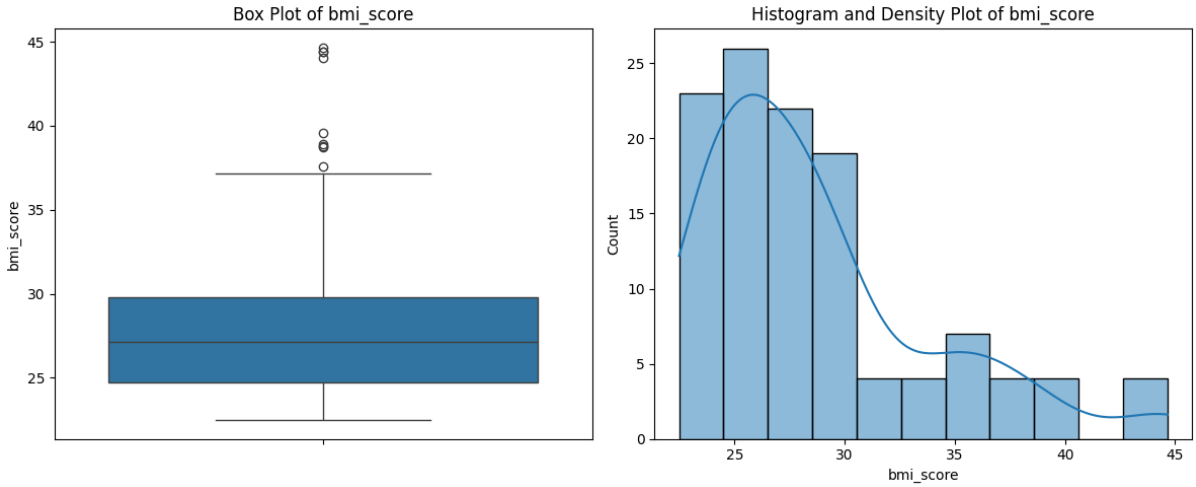
1. Feature: 'bmi_body_fat' (Body Composition)
    - **Outlier Index:** [70]
    - **Found at:** part_id 2085, clinical_visit 3



Index	part_id	clinical_visit	bmi_body_fat	lean_body_mass	bmi_score	health_rate
68	2085	1	NaN	NaN	27.120316	4 - Good
69	2085	2	NaN	NaN	27.325773	3 - Medium
70	**2085**	**3**	**218.0**	**-78.7060**	**27.407955**	4 - Good
71	2085	4	22.9	51.2715	26.978782	4 - Good

*(The value of **218.0** for body fat is physiologically implausible for an adult, and the linked **-78.7060** for lean body mass is a clear error, strongly suggesting data entry/measurement error for index 70.)*

2. Feature: 'bmi_score' (Body Mass Index)
    - **Outlier Indices:** [13, 14, 15]
    - **Found at:** part_id 1088, clinical_visit 2, 3, and 4

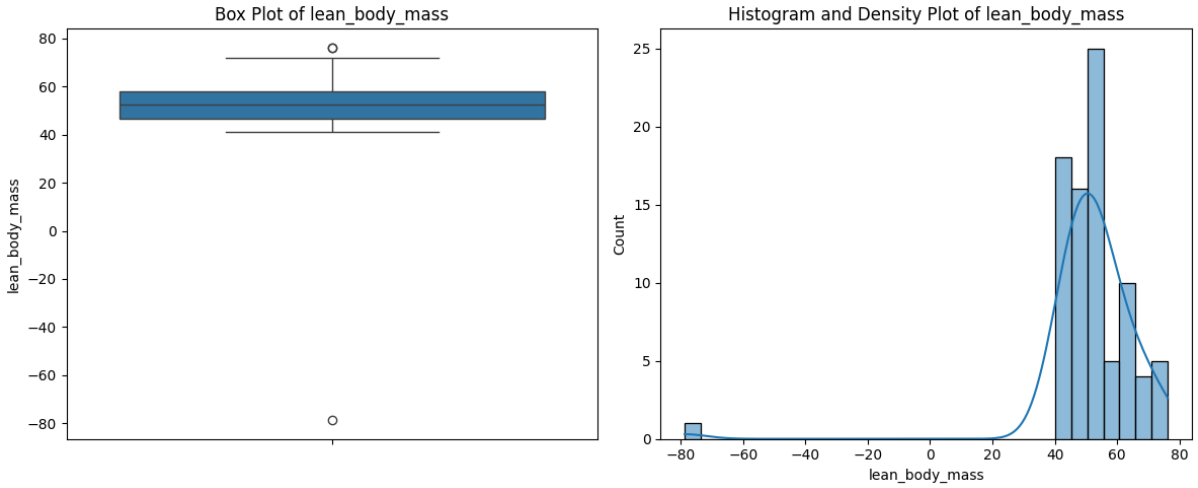


Index	part_id	clinical_visit	bmi_score	lean_body_mass	bmi_body_fat	health_rate
12	1088	1	44.062500	71.5152	36.6	1 - Very bad
13	**1088**	**2**	**44.414062**	71.4036	37.2	1 - Very bad
14	**1088**	**3**	**44.375000**	71.7952	36.8	1 - Very bad
15	**1088**	**4**	**44.658044**	71.0141	37.1	1 - Very bad

(BMI values around 44 are severely obese (Class III). While high, this could be a legitimate biological value, especially given the consistent high measurements across all visits for this participant. It warrants careful consideration before removal.)

3. Feature: 'lean_body_mass' (LBM)

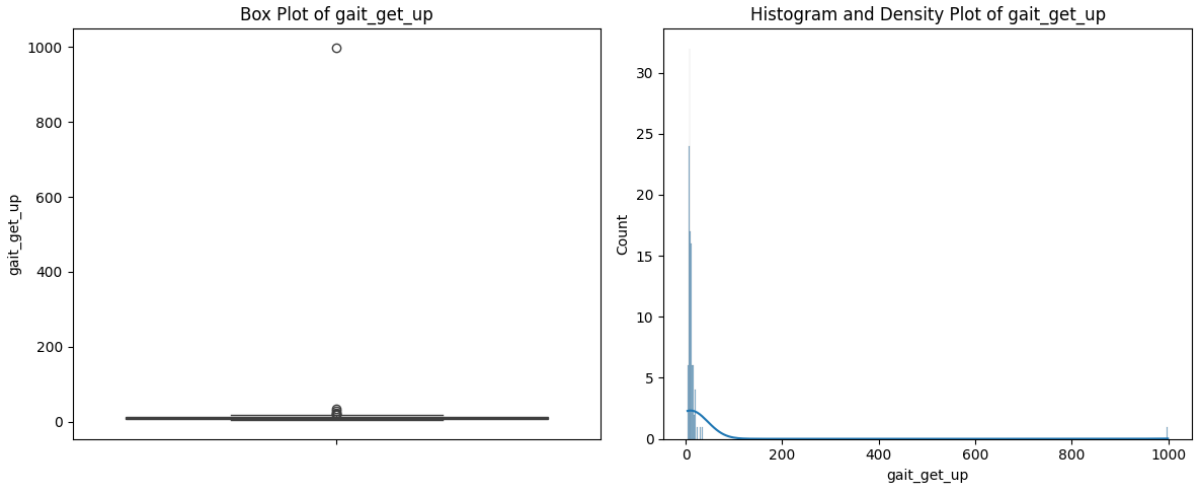
    - **Outlier Index:** [70]
    - **Found at:** part_id 2085, clinical_visit 3



Index	part_id	clinical_visit	lean_body_mass	bmi_score	bmi_body_fat	health_rate
68	2085	1	NaN	27.120316	NaN	4 - Good
69	2085	2	NaN	27.325773	NaN	3 - Medium
70	**2085**	**3**	**-78.7060**	**27.407955**	**218.0**	4 - Good
71	2085	4	51.2715	26.978782	22.9	4 - Good

(The negative value for lean_body_mass is biologically impossible, making index 70 a strong candidate for removal or treatment as missing data, regardless of the bmi_body_fat outlier in the same row.)

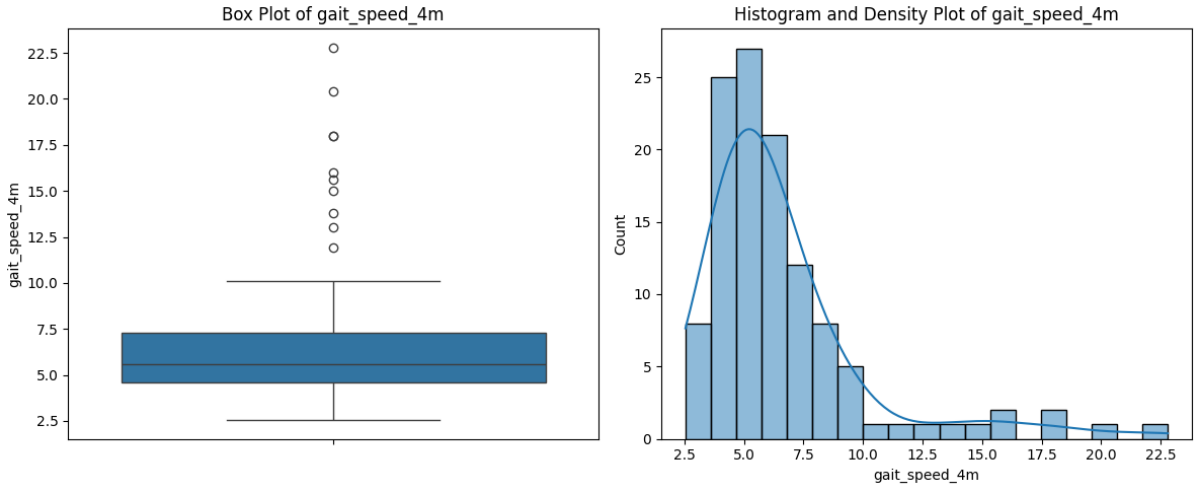
4. Feature: 'gait_get_up' (Timed Up and Go/TUG)
    - **Outlier Index:** [56]
    - **Found at:** part_id 1104, clinical_visit 1



Index	part_id	clinical_visit	gait_get_up	gait_speed_4m	raise_chair_time	health_rate
56	**1104**	**1**	**999.0**	22.8	999.0	2 - Bad
57	1104	2	24.7	13.8	999.0	2 - Bad
58	1104	3	34.7	20.4	999.0	2 - Bad
59	1104	4	30.1	15.6	999.0	2 - Bad

*(The value **999.0** is likely a placeholder for "test not performable" or missing data. If your Z-score detection didn't treat this as a special case, it was flagged as a high numerical outlier. It should be treated as **missing data (NaN)** for robust analysis.)*

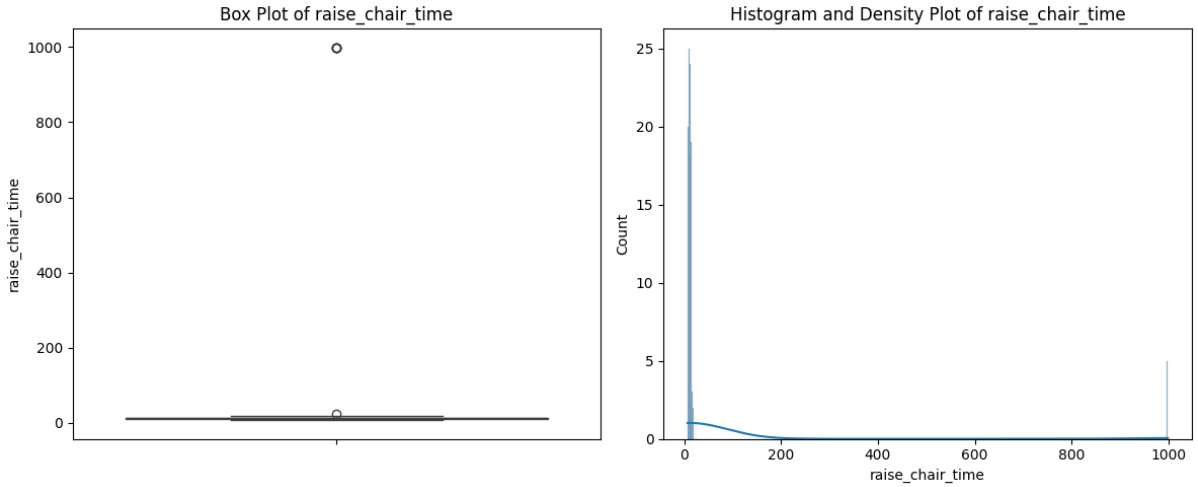
5. Feature: 'gait_speed_4m' (Walking Speed)
    - **Outlier Indices:** [56, 58, 113, 115]
    - **Found at:** part_id 1104 (clinical_visit 1, 3) and part_id 2584 (clinical_visit 1, 3)



Index	part_id	clinical_visit	gait_speed_4m	raise_chair_time	gait_get_up	health_rate
56	**1104**	**1**	**22.8**	999.0	999.0	2 - Bad
57	1104	2	13.8	999.0	24.7	2 - Bad
58	**1104**	**3**	**20.4**	999.0	34.7	2 - Bad
59	1104	4	15.6	999.0	30.1	2 - Bad
113	**2584**	**1**	**18.0**	14.0	15.0	4 - Good
114	2584	2	15.0	14.0	12.0	4 - Good
115	**2584**	**3**	**18.0**	14.0	15.0	4 - Good
116	2584	4	16.0	14.2	12.0	3 - Medium

(Gait speed is typically measured in m/s. Given the maximum healthy adult speed is around 2.0 m/s and these values are much higher, they likely represent the time taken over 4 meters in seconds (T4M). Higher time values (like 22.8 and 20.4) are usually indicative of ***slower*** *speed/poorer performance, but the Z-score indicates they are outliers on the high side of the distribution of values.)*

6. Feature: 'raise_chair_time' (Chair Stand Up Time)
    - **Outlier Indices:** [20, 56, 57, 58, 59]
    - **Found at:** part_id 1090 (clinical_visit 1) and part_id 1104 (all visits)

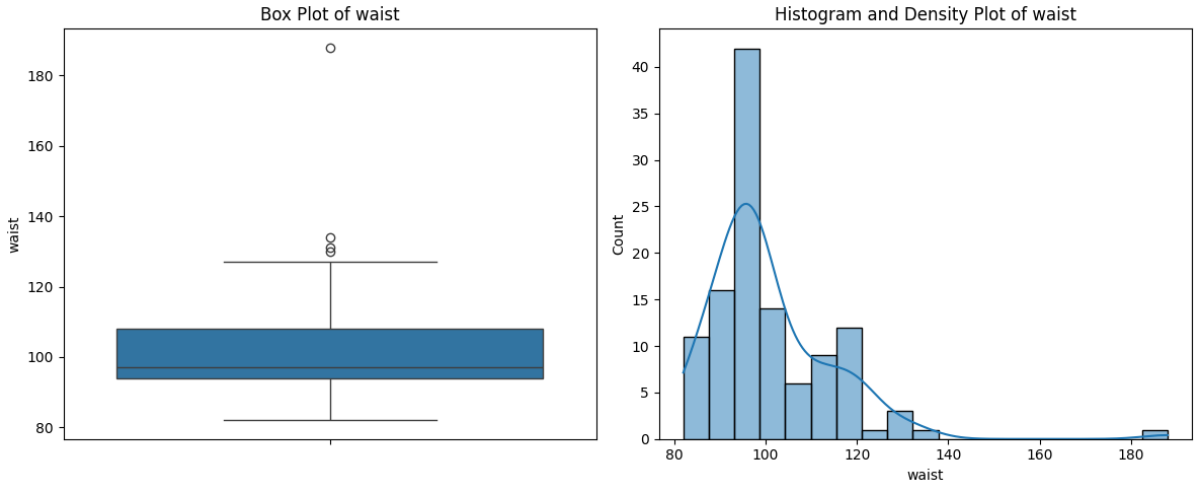


Index	part_id	clinical_visit	raise_chair_time	gait_speed_4m	gait_get_up	health_rate
20	**1090**	**1**	**999.00**	11.90	7.9	4 - Good
21	1090	2	24.31	8.41	15.3	4 - Good
22	1090	3	18.30	10.10	16.2	3 - Medium
23	1090	4	19.20	9.60	16.5	3 - Medium
56	**1104**	**1**	**999.00**	22.80	999.0	2 - Bad
57	**1104**	**2**	**999.00**	13.80	24.7	2 - Bad
58	**1104**	**3**	**999.00**	20.40	34.7	2 - Bad
59	**1104**	**4**	**999.00**	15.60	30.1	2 - Bad

*(The value **999.00** for all visits of participant 1104, and the first visit of 1090, is most likely a missing data code. Similar to gait_get_up, these should be treated as **missing values (NaN)**.)*

7. Feature: 'waist' (Waist Circumference)

    - **Outlier Index:** [79]
    - **Found at:** part_id 2087, clinical_visit 4

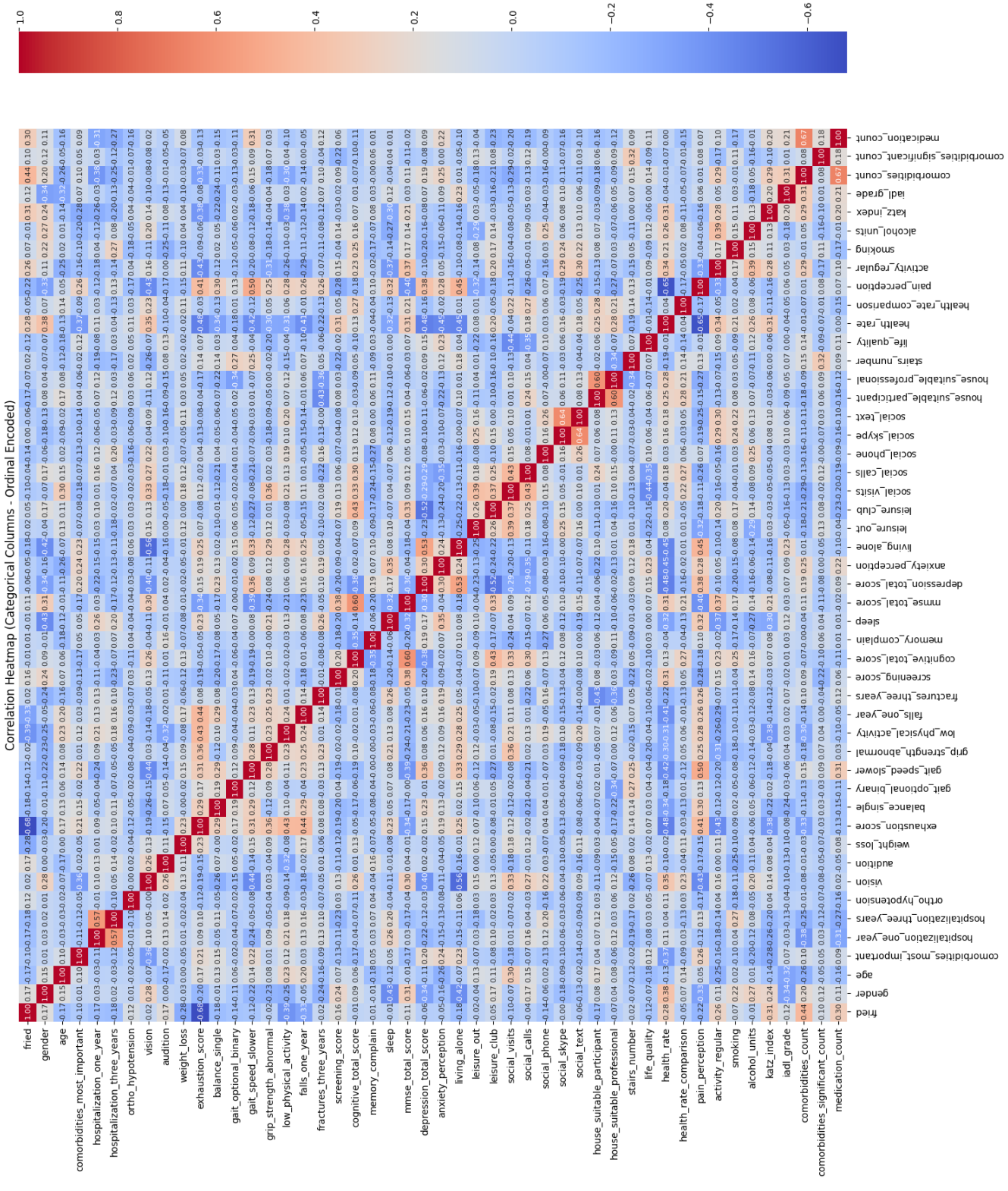


Index	part_id	clinical_visit	waist	health_rate
76	2087	1	89.0	4 - Good
77	2087	2	85.0	4 - Good
78	2087	3	85.0	5 - Excellent
79	**2087**	**4**	**188.0**	4 - Good

*(The value **188.0** cm for waist circumference is exceptionally high, particularly when the previous three visits showed values around 85-89 cm. This large jump suggests a potential **data entry error or measurement mistake** and should be verified or treated as an outlier/missing value.)*

Appendix C - Correlation Heatmap

This appendix provides the heatmap referred on the Final Feature Space and Dimensionality Reduction (Section 4.2.4)



Appendix D - Data quality Assessment (K-S Test and Distribution Overlap)

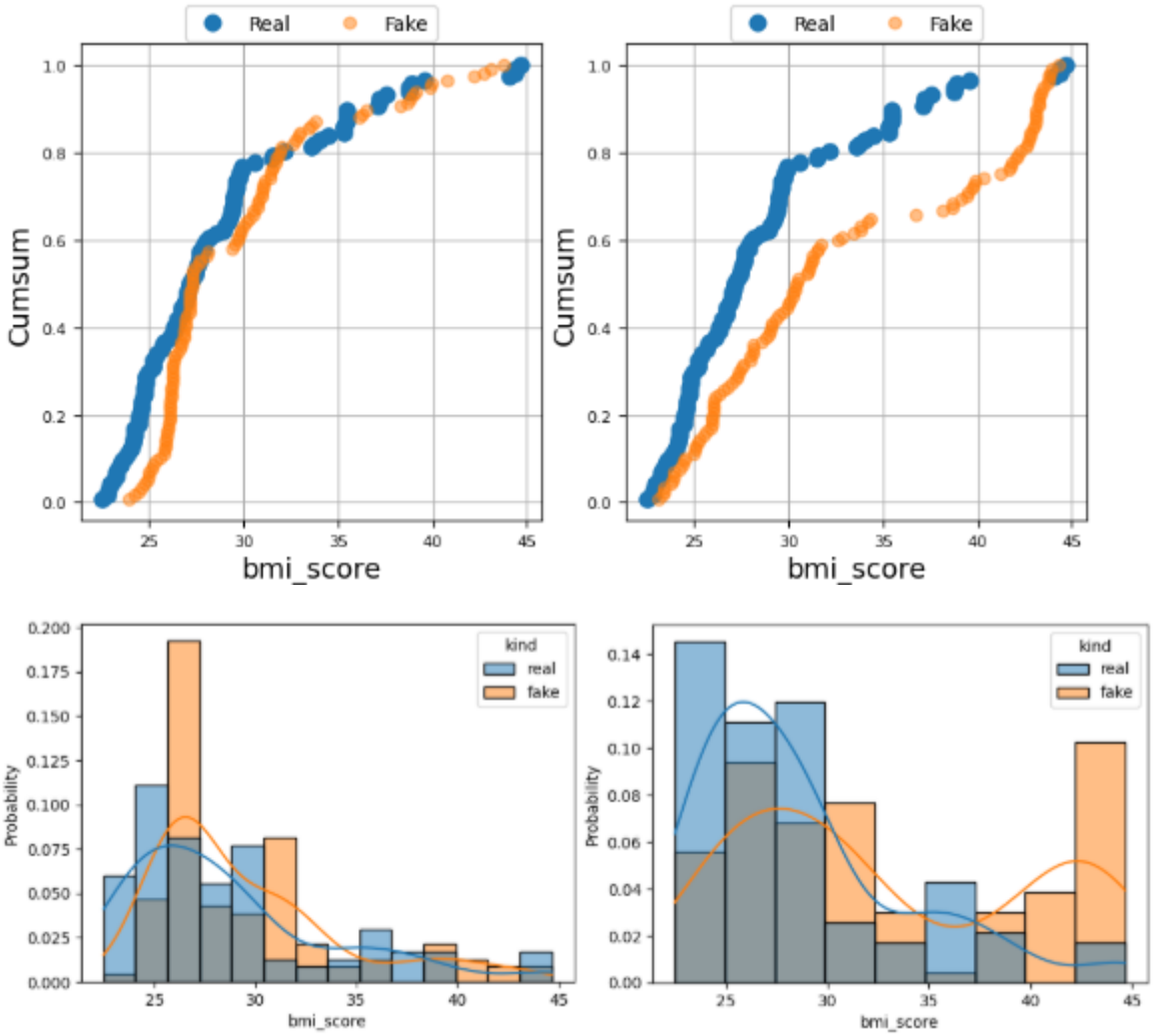
This appendix provides the detailed, feature-by-feature results of the data fidelity assessment, comparing the distributions of the two synthetic datasets (Imbalanced and Balanced) against the original Real Dataset using the Kolmogorov-Smirnov (K-S) test.

The K-S test quantifies the statistical overlap between two distributions. The visualizations below confirm the findings discussed in Section 4.4.1, which stated that while the Imbalanced Synthetic Data maintains high fidelity, the Balanced Synthetic Data introduces minor statistical distortion due to rejection sampling.

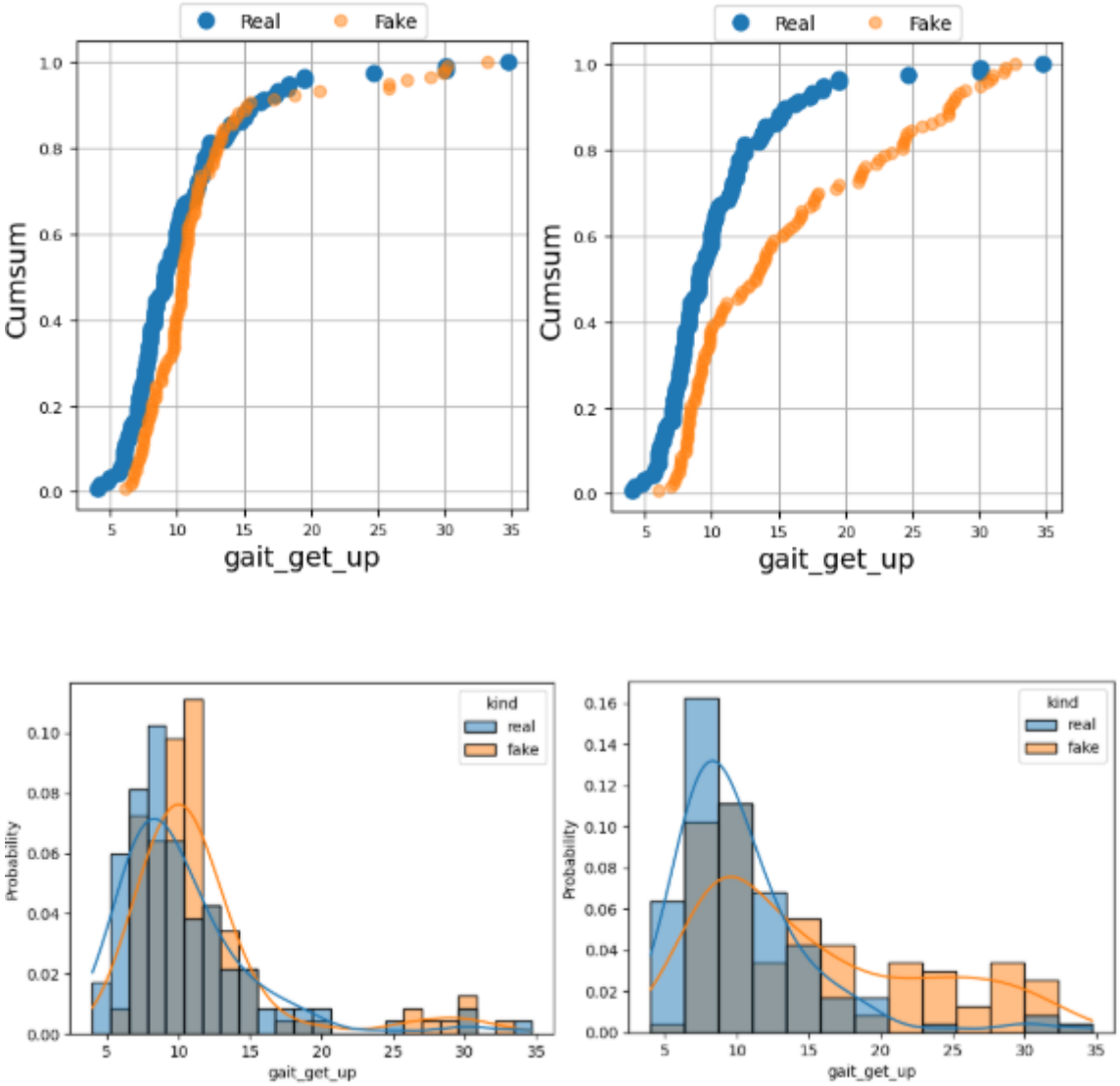
For each feature, the following 2x2 figure grid is presented to facilitate direct comparison against the Real Data baseline:

- Top-Left (Imbalanced CDF): Cumulative Distribution Function (CDF) comparing Imbalanced Synthetic Data to Real Data.
- Top-Right (Balanced CDF): Cumulative Distribution Function (CDF) comparing Balanced Synthetic Data to Real Data.
- Bottom-Left (Imbalanced PDF): Probability Density Function (PDF) comparing Imbalanced Synthetic Data to Real Data.
- Bottom-Right (Balanced PDF): Probability Density Function (PDF) comparing Balanced Synthetic Data to Real Data.

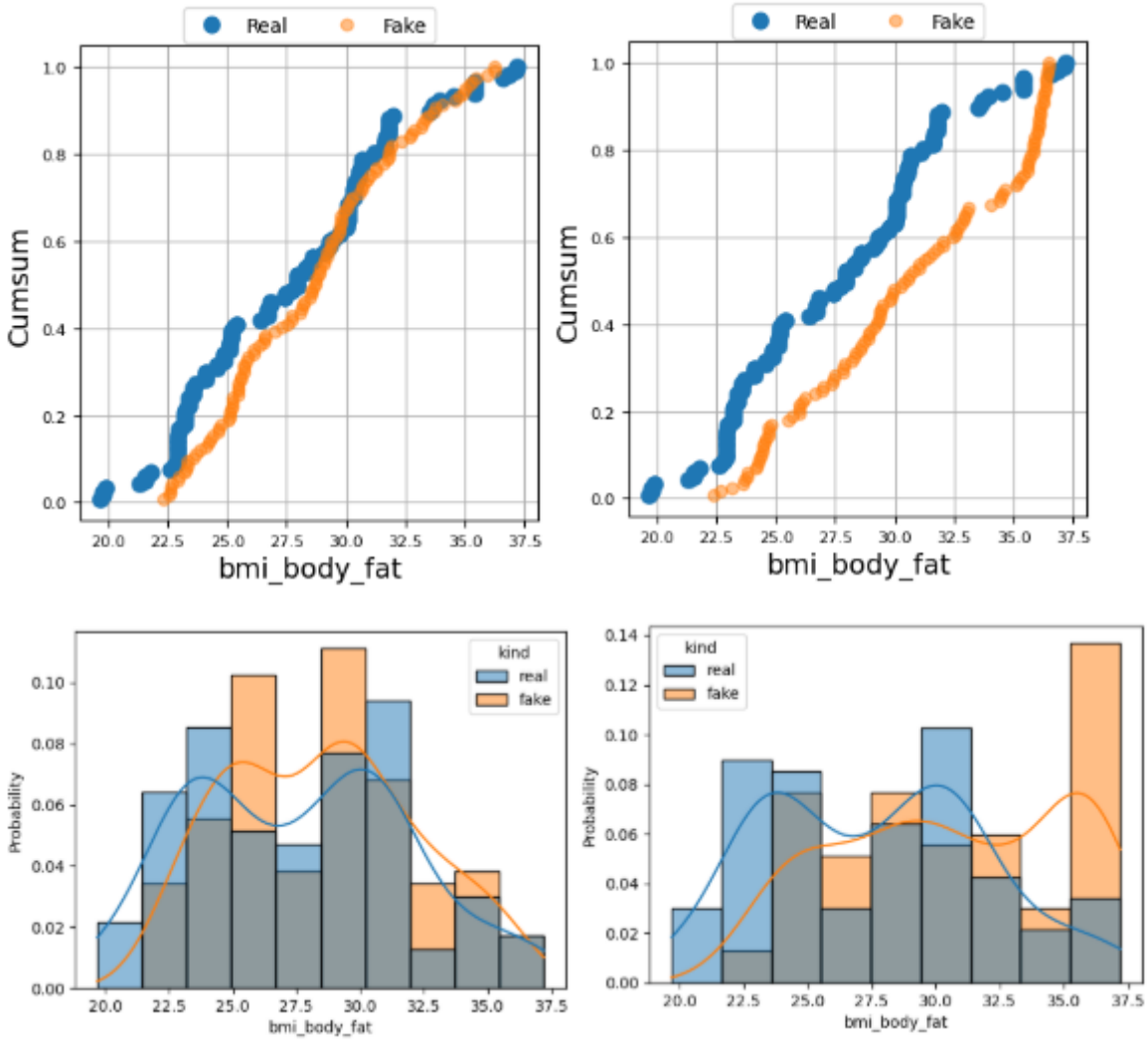
1. BMI score



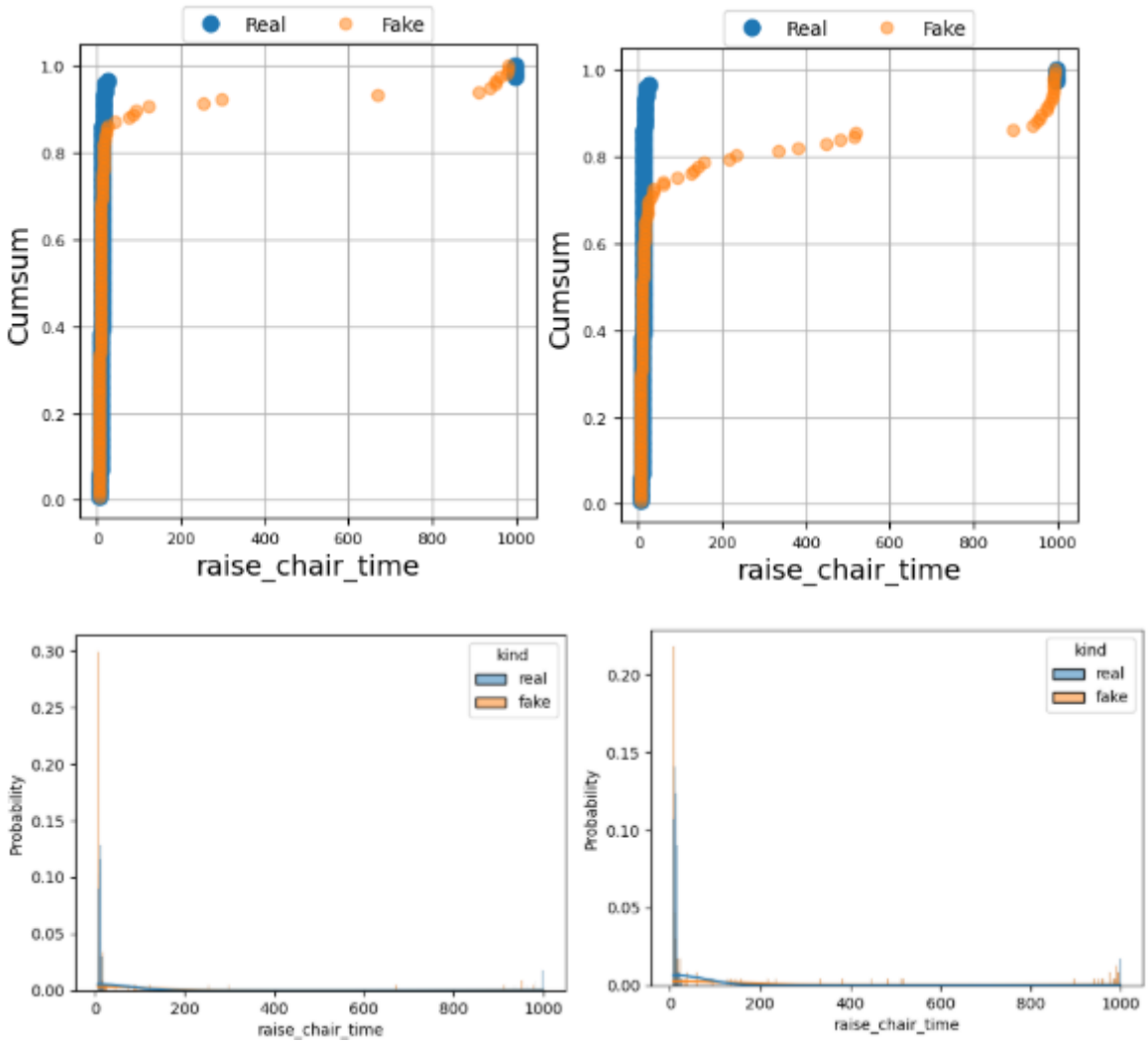
2. Gait get Up



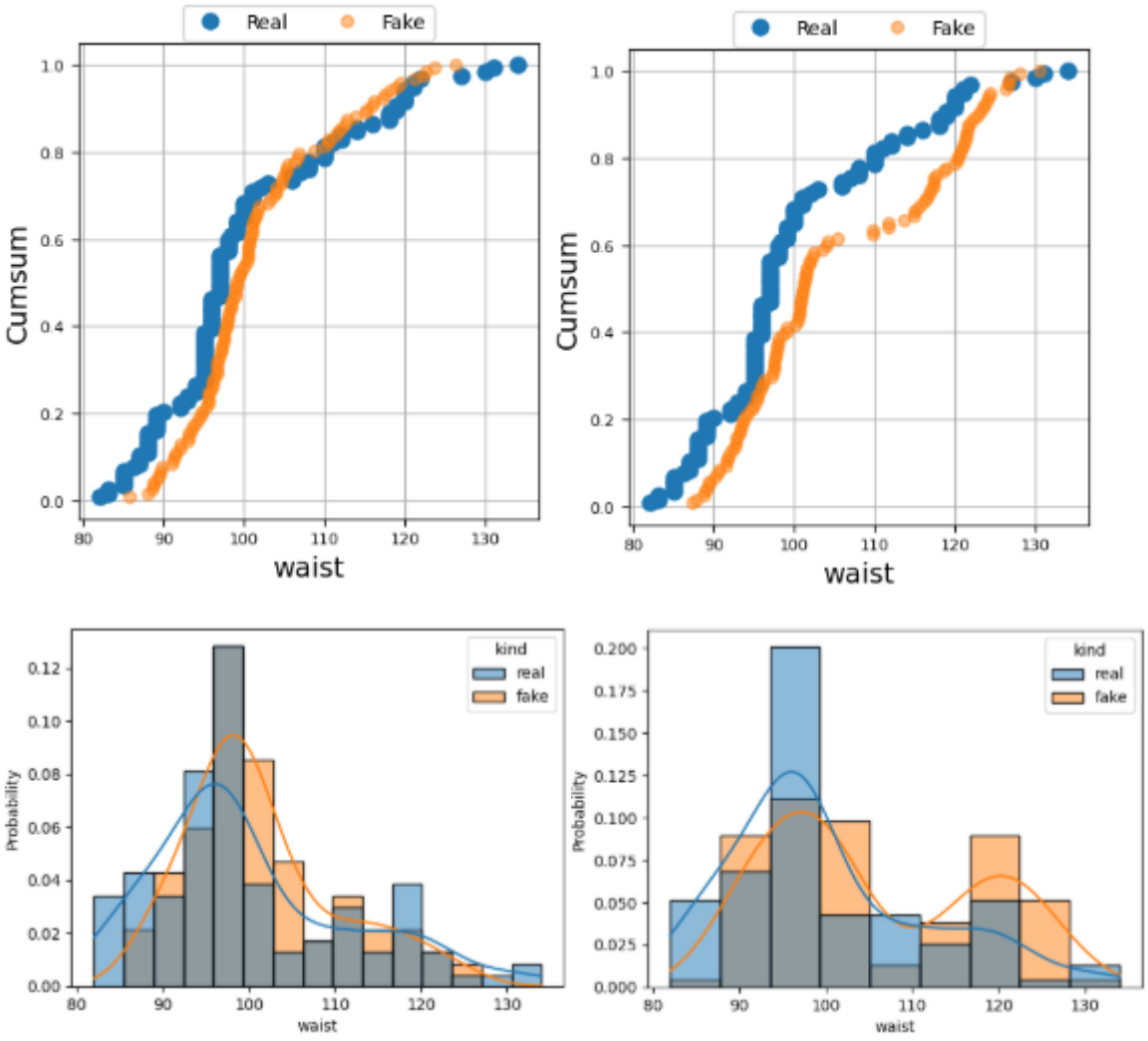
3. BMI Body Fat



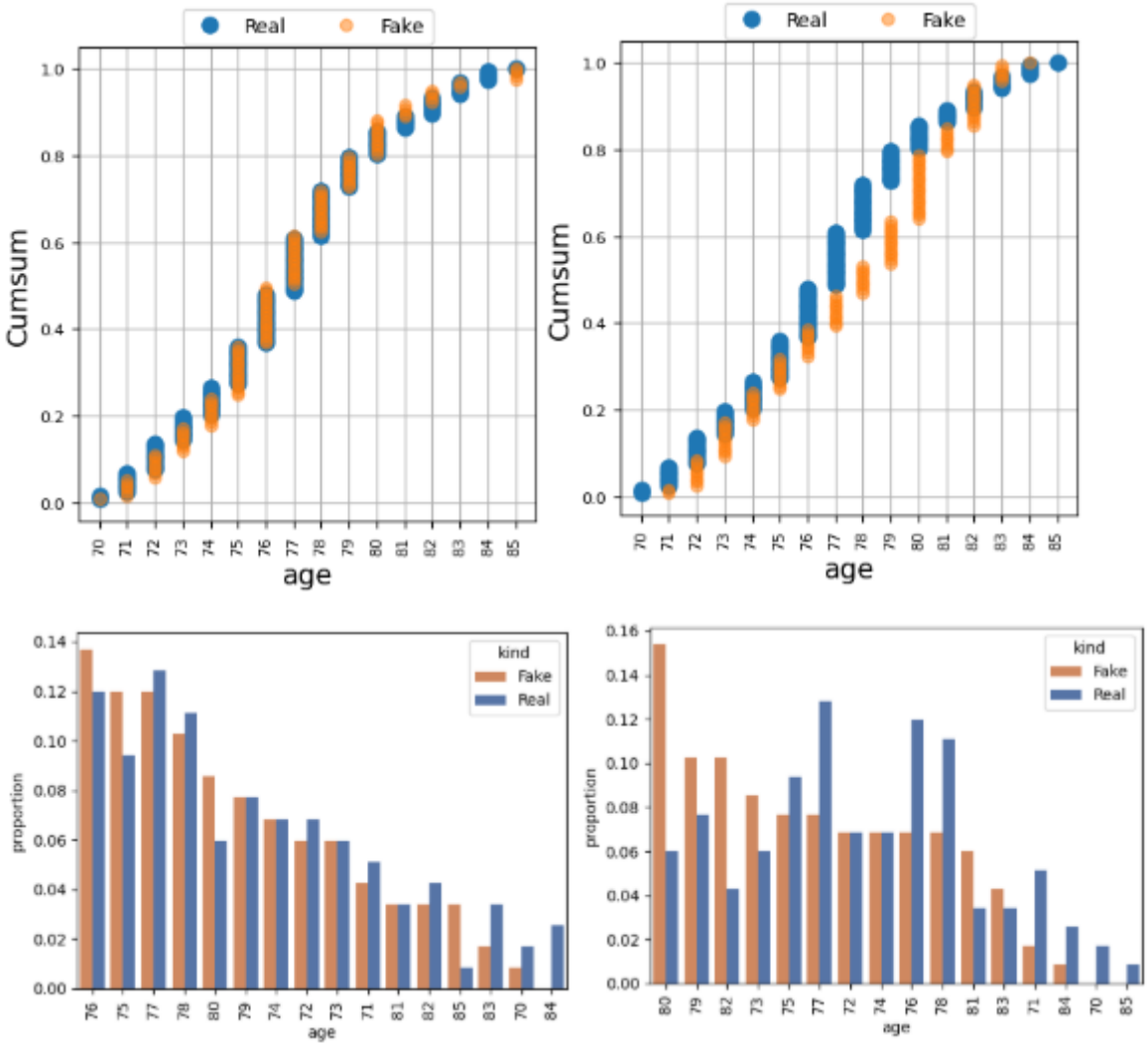
4. Raise chair time



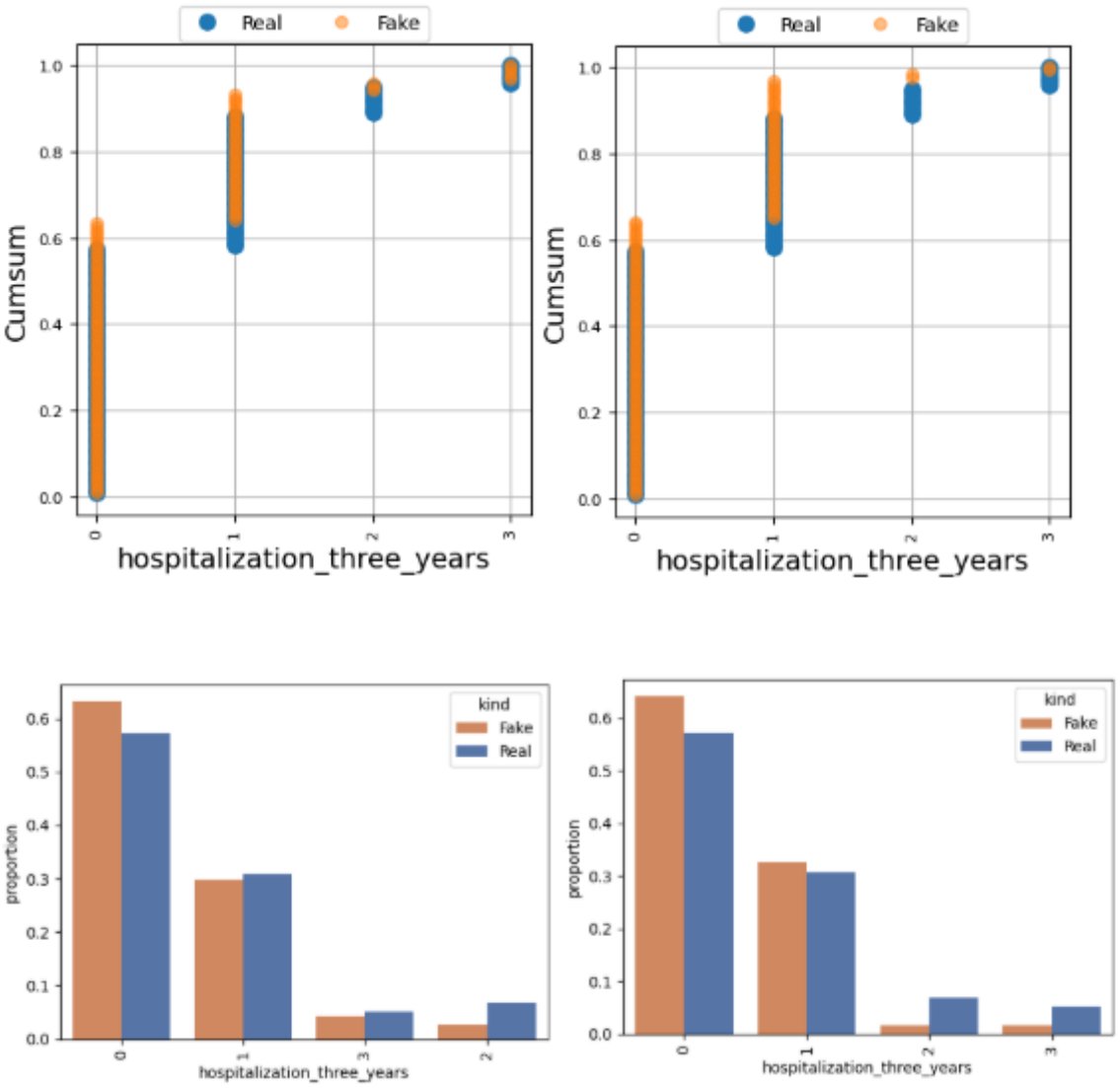
5. Waist



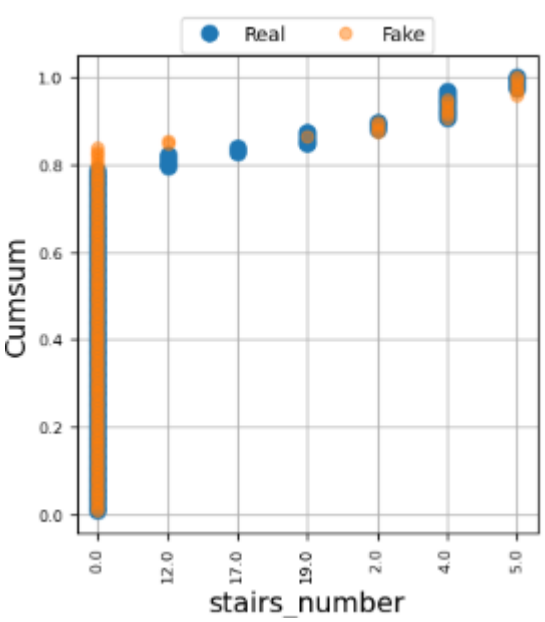
6. Age

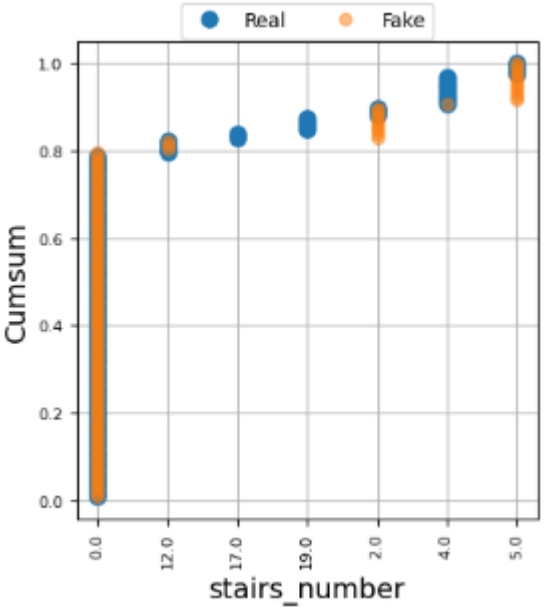


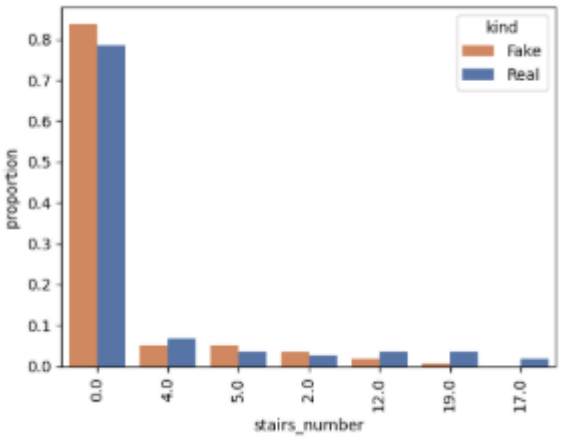
7. Hospitalization Three Years

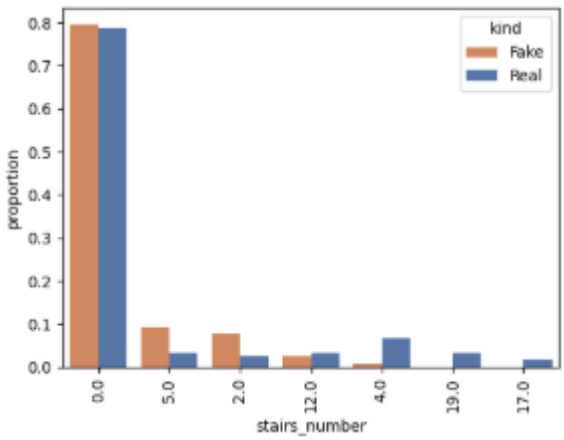


8. Stairs Number

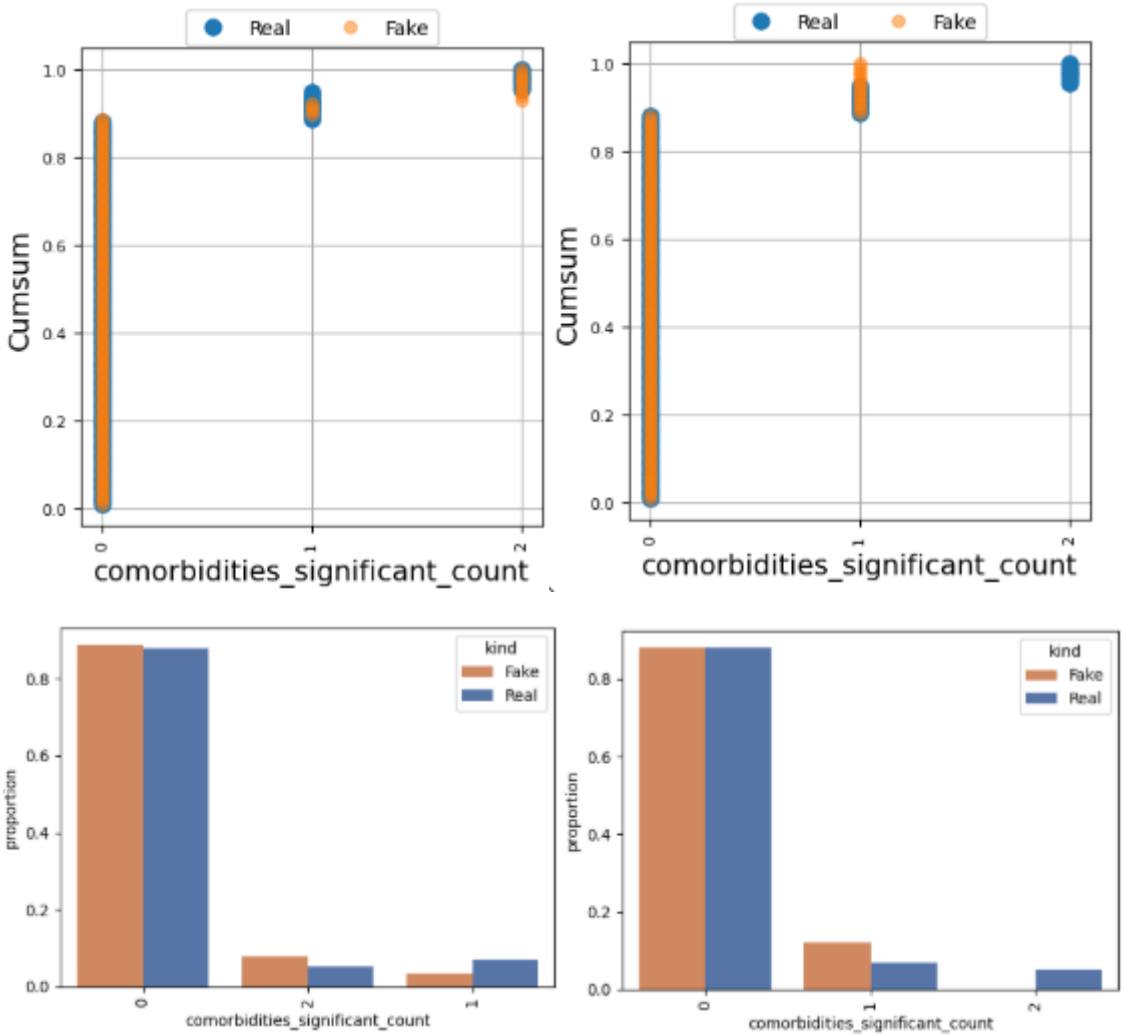




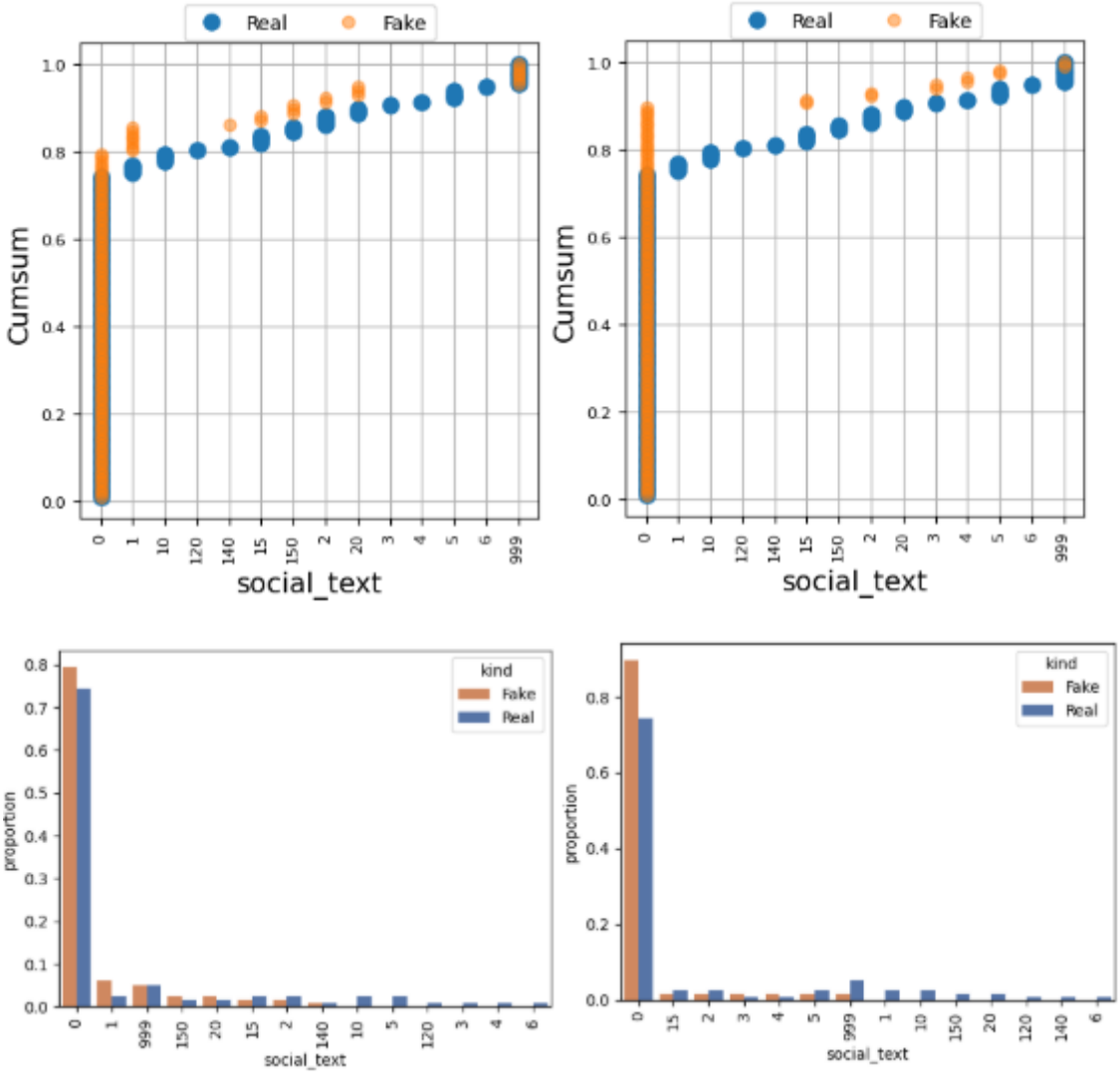




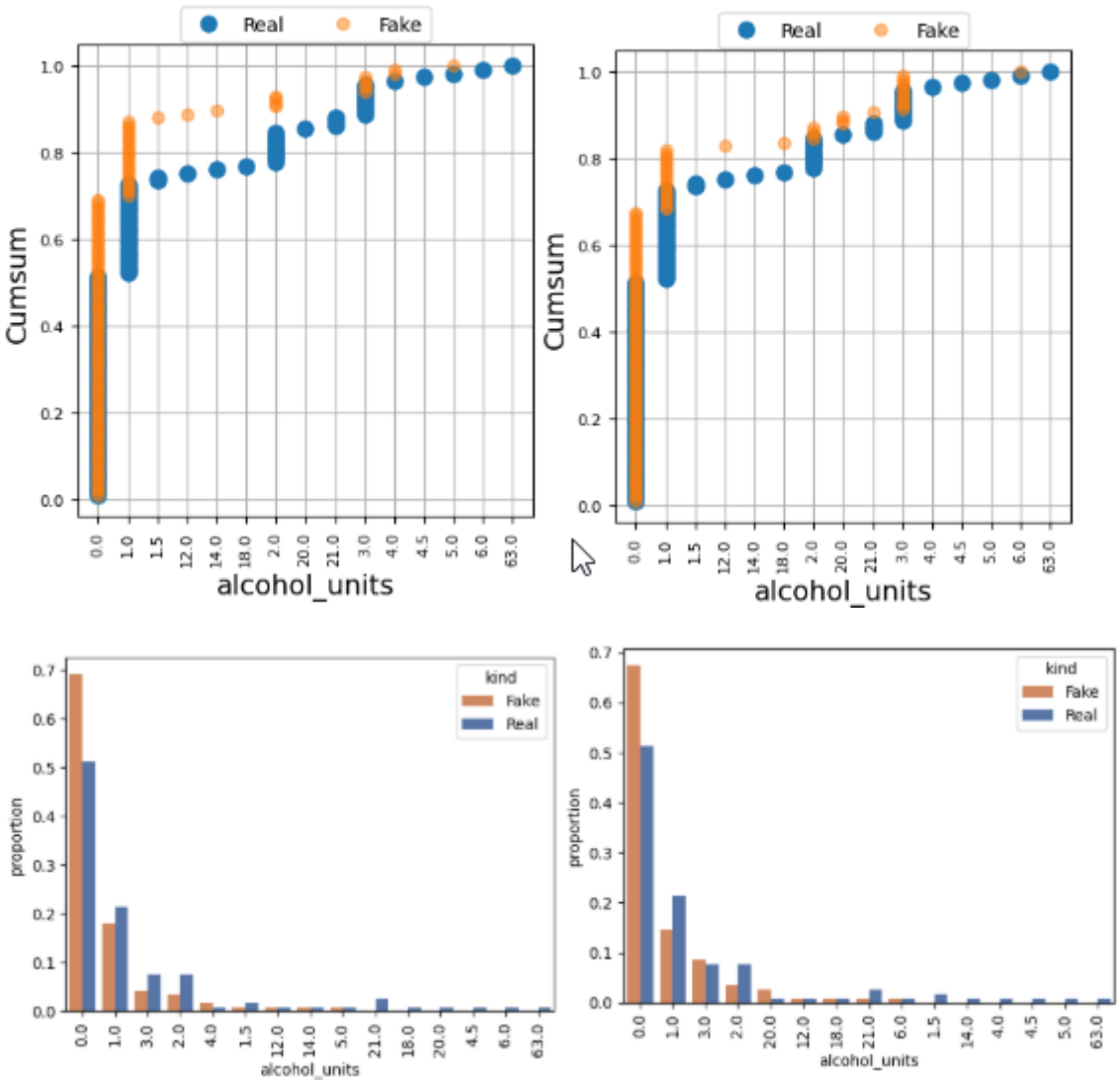
9. Comorbidities significant count



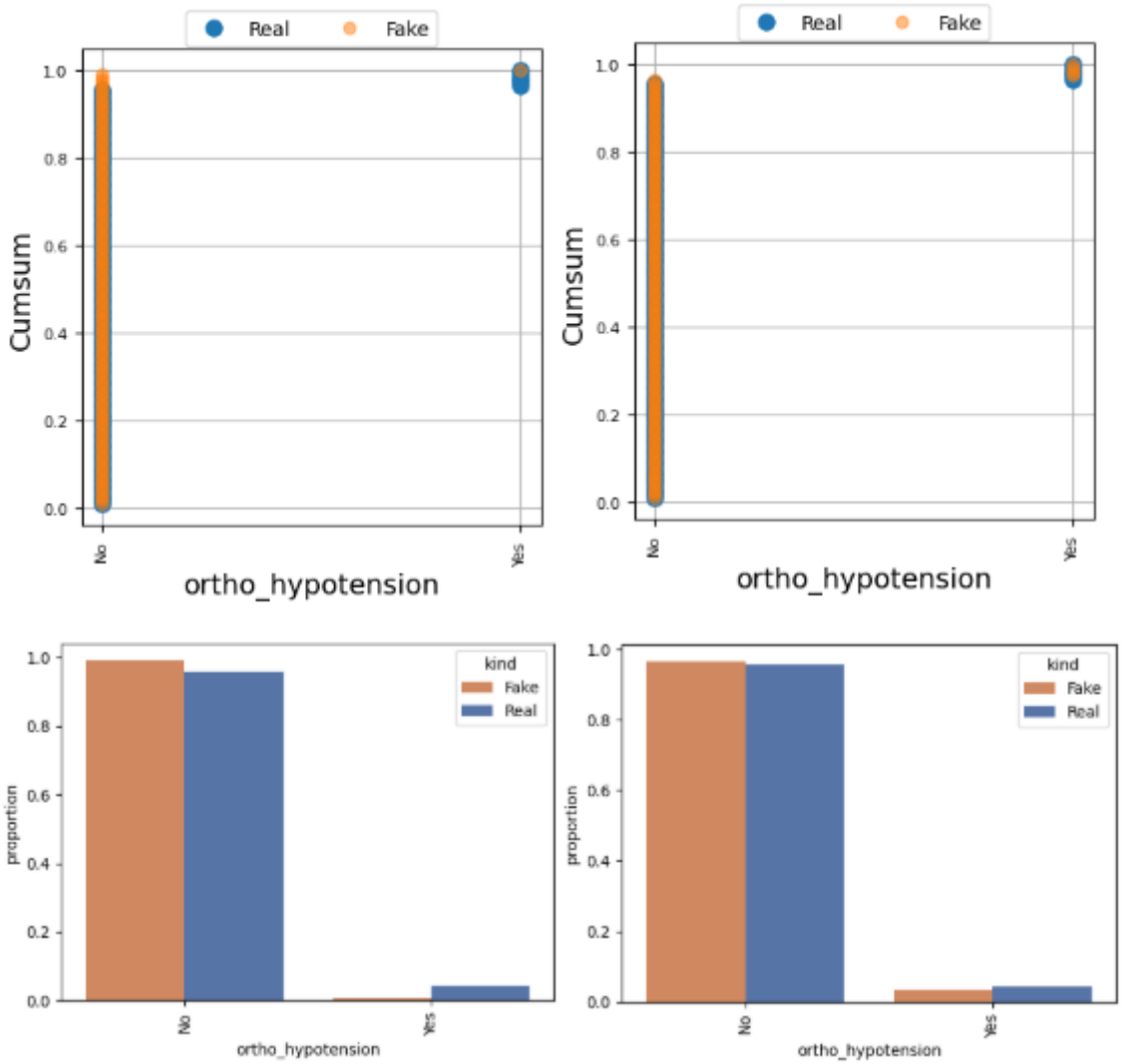
10. Social Text



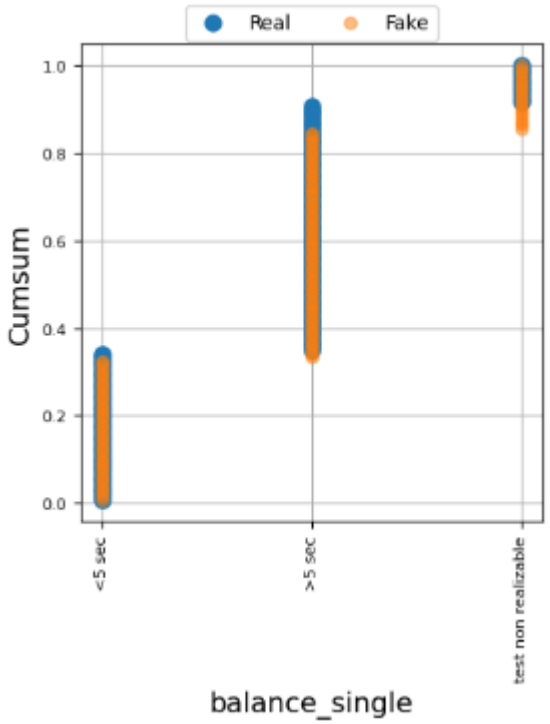
11. Alcohol units

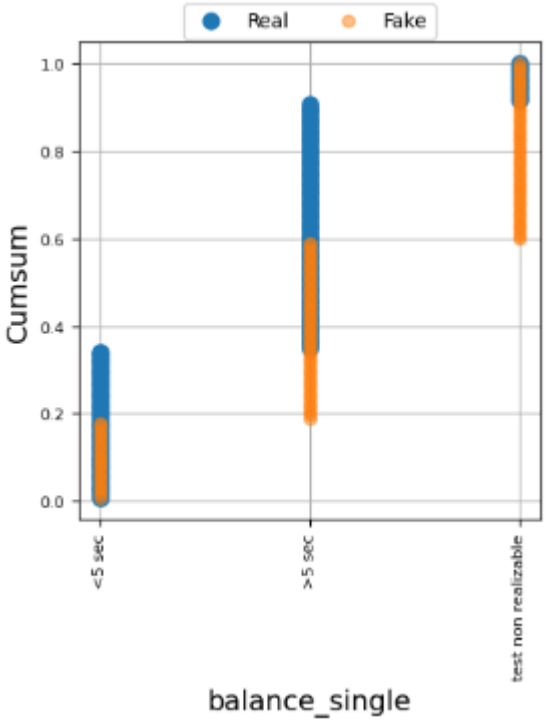


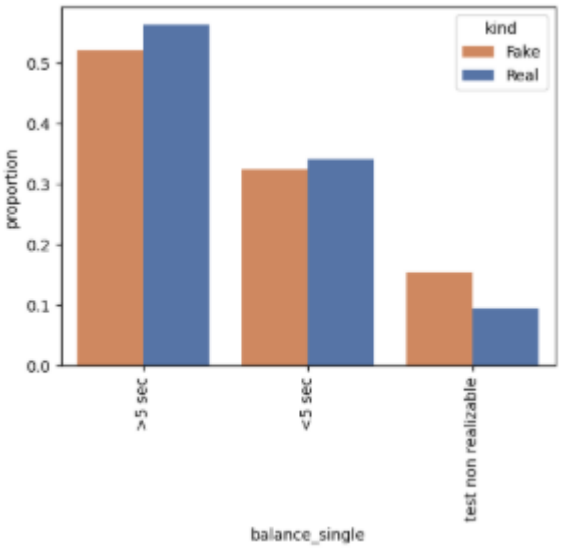
12. Ortho Hypertension

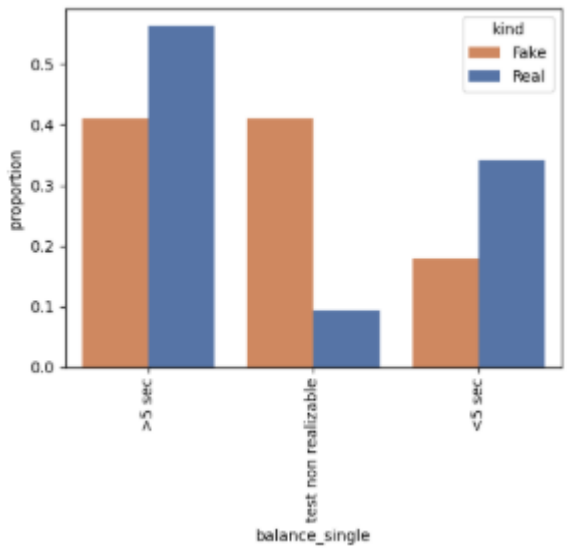


13. Balance Single

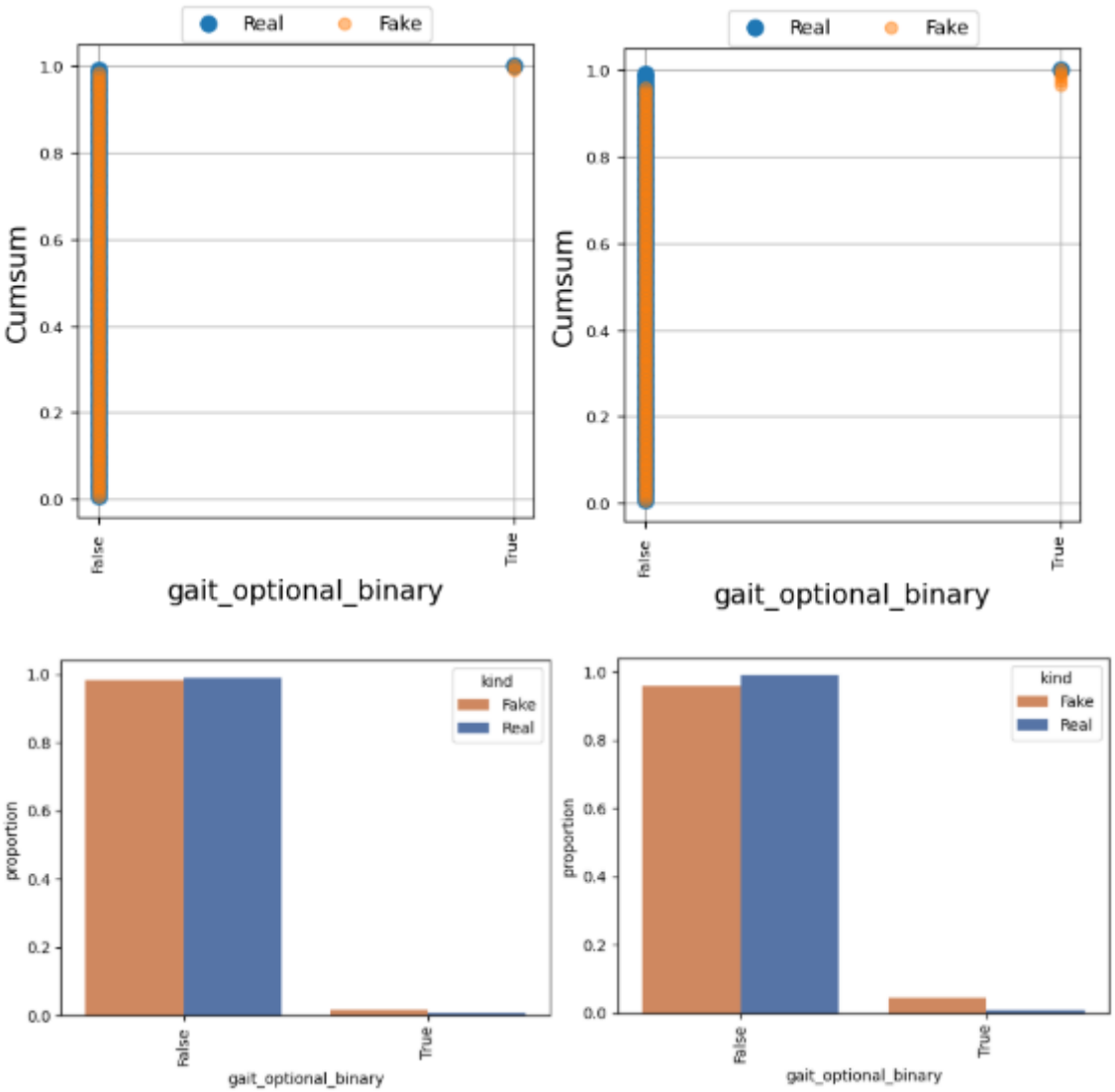




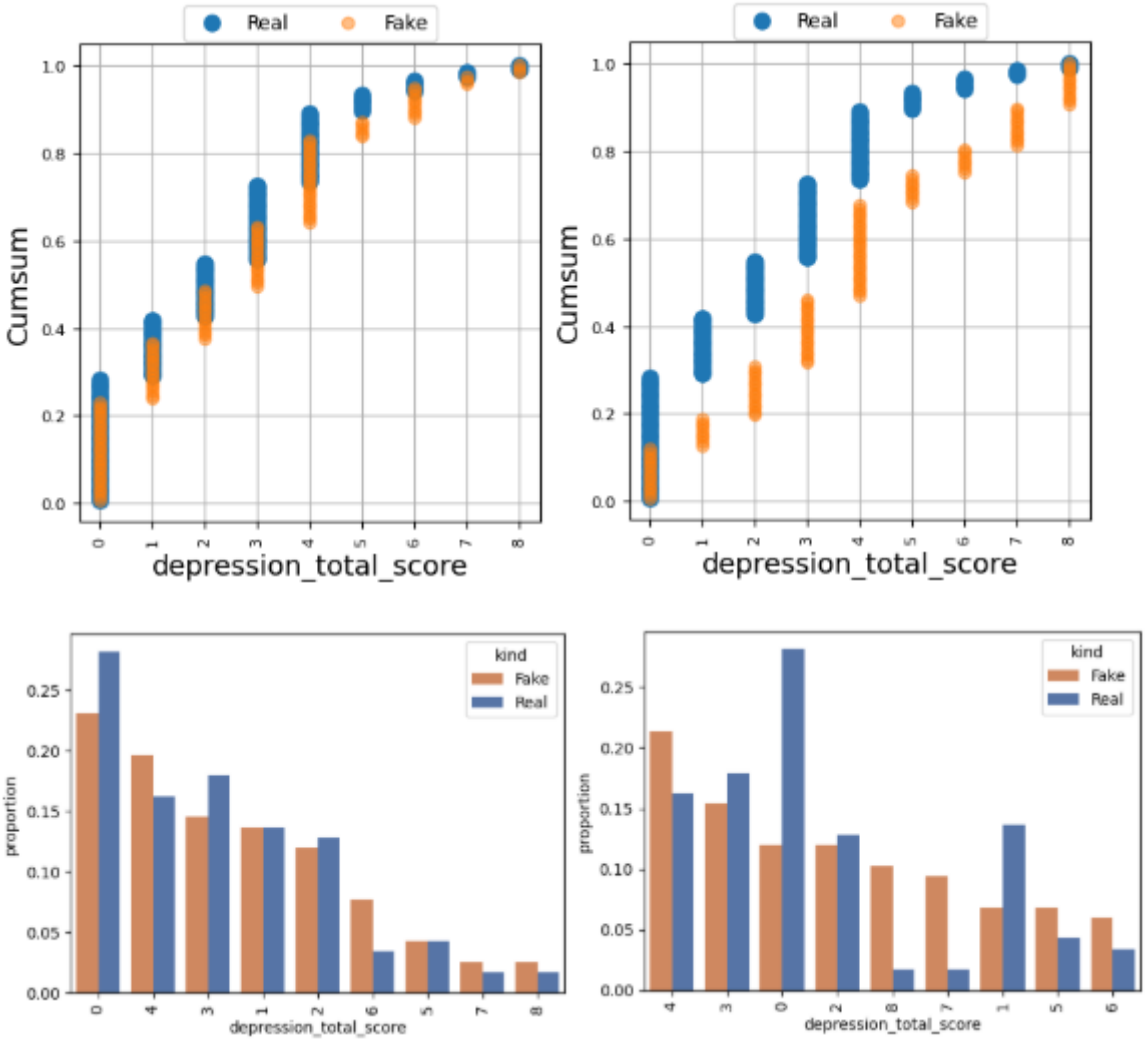




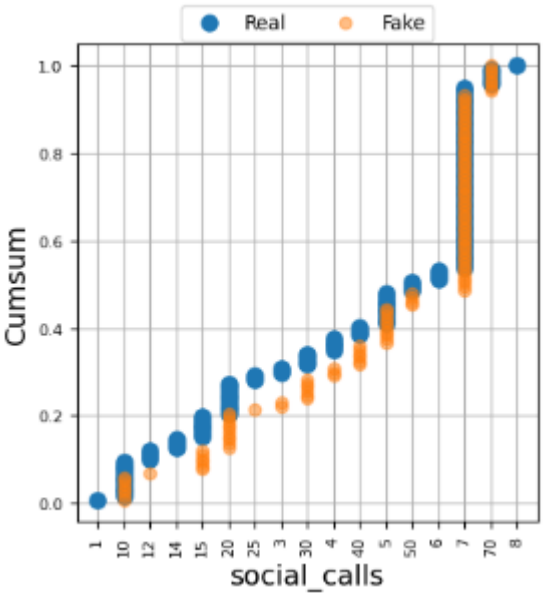
14. Gait optional binary

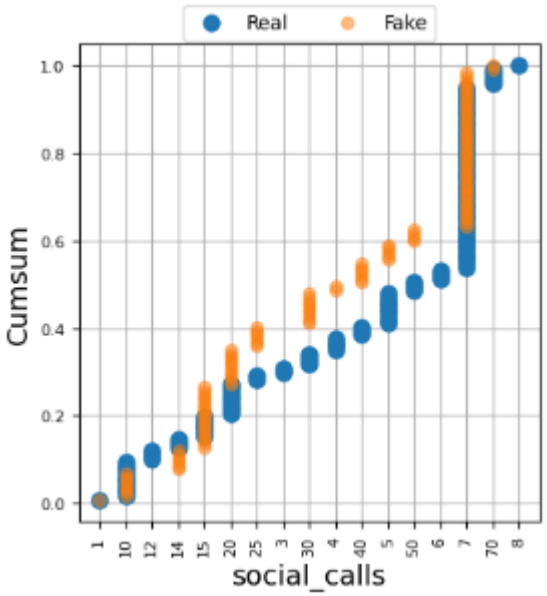


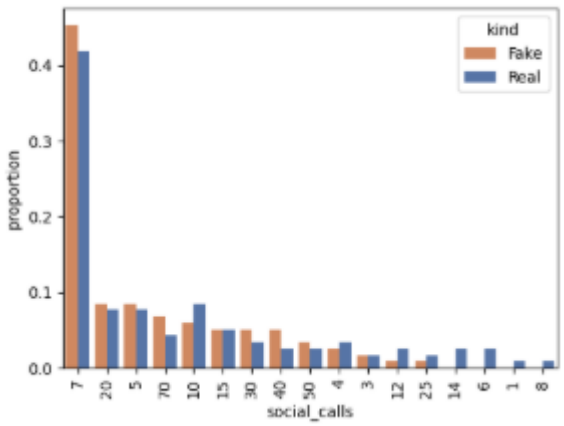
15. Depression total score

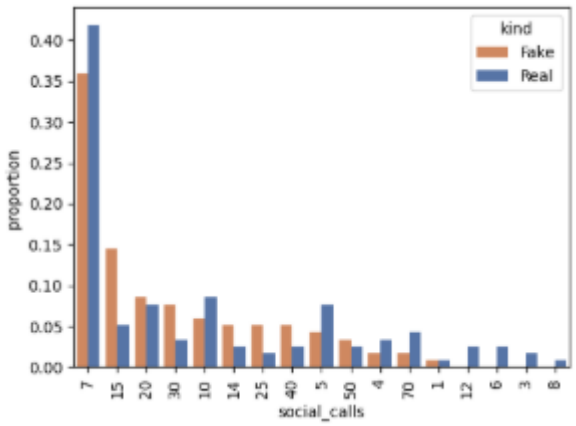


16. Social calls

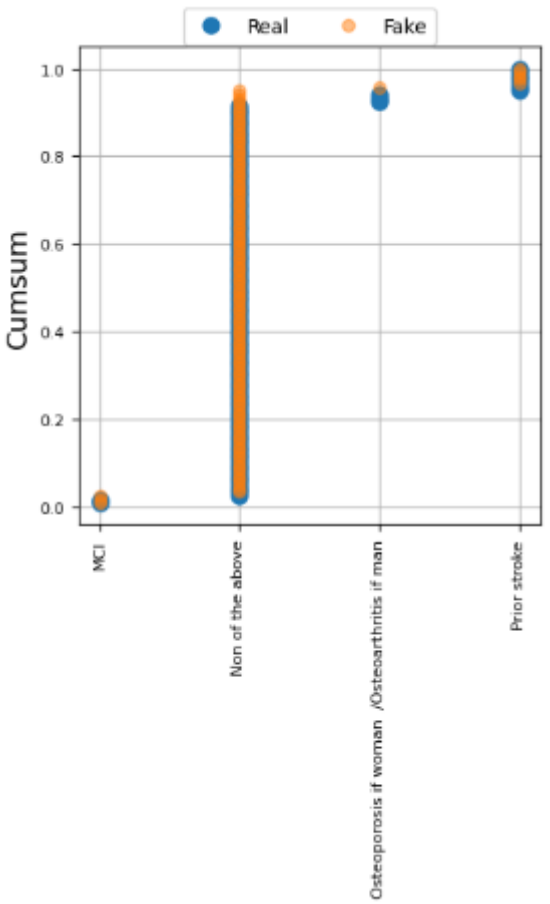


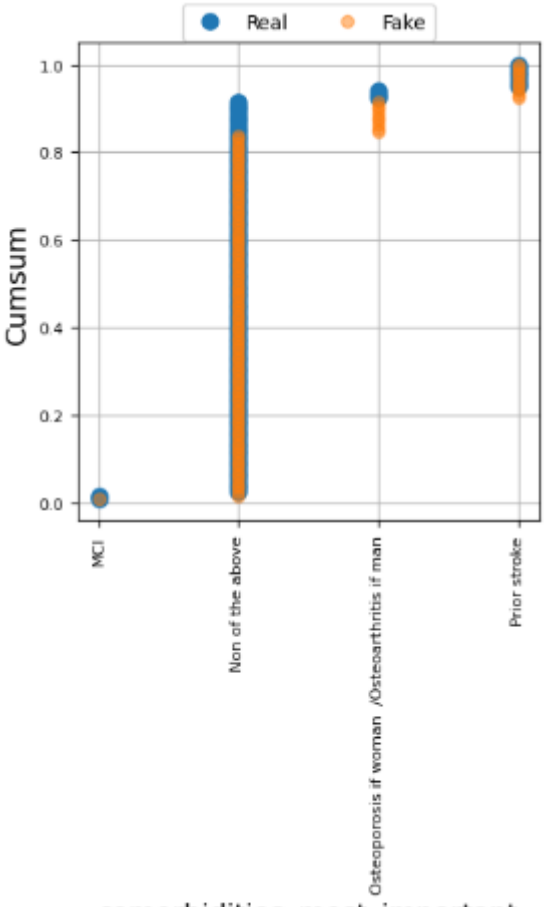


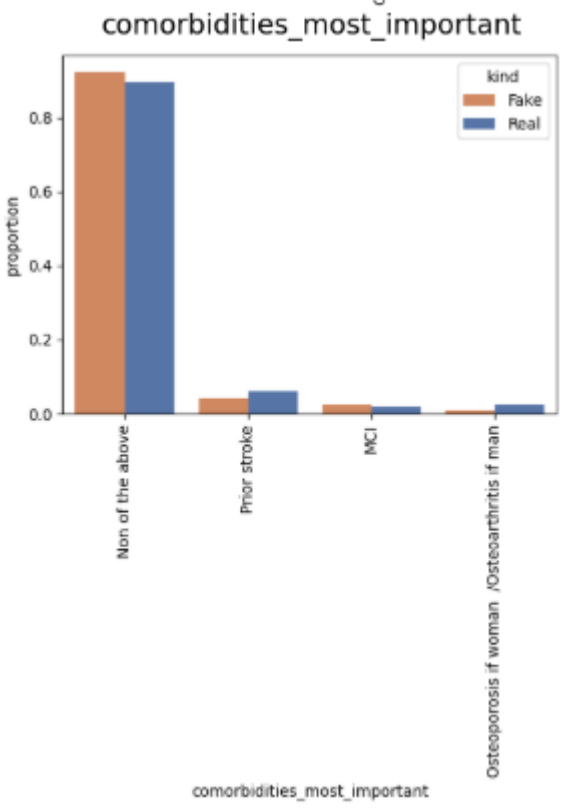


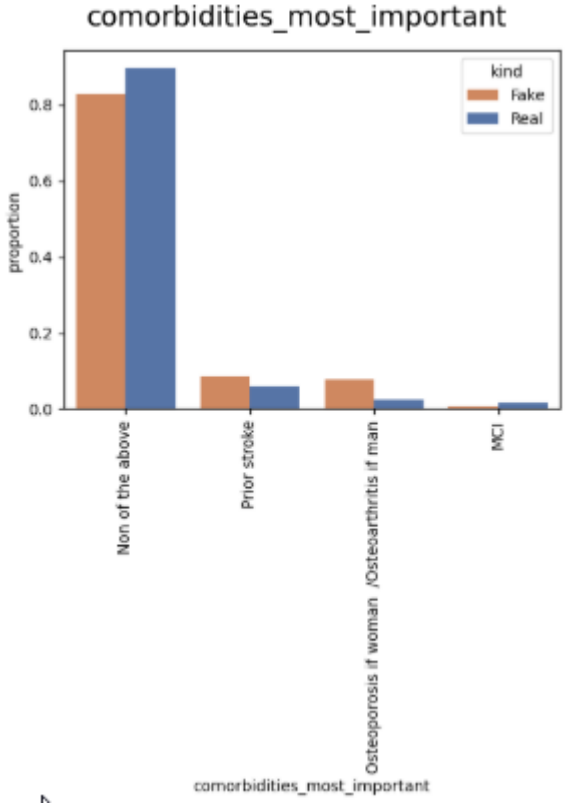


17. Comorbidities most important

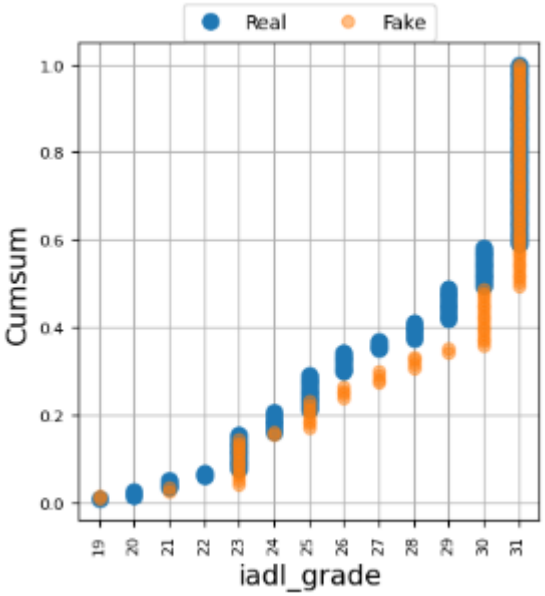


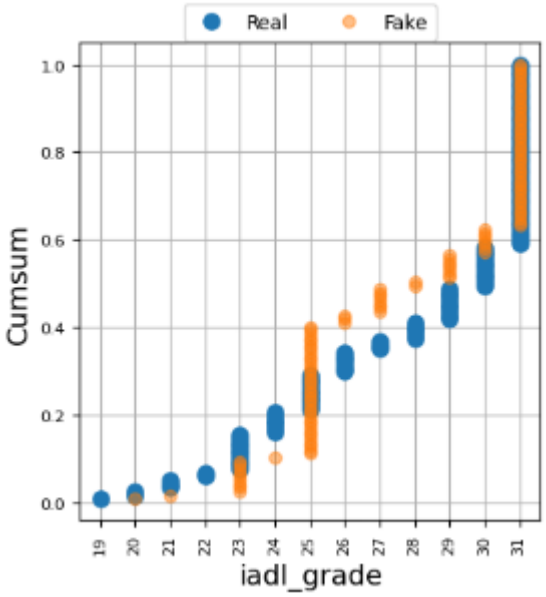


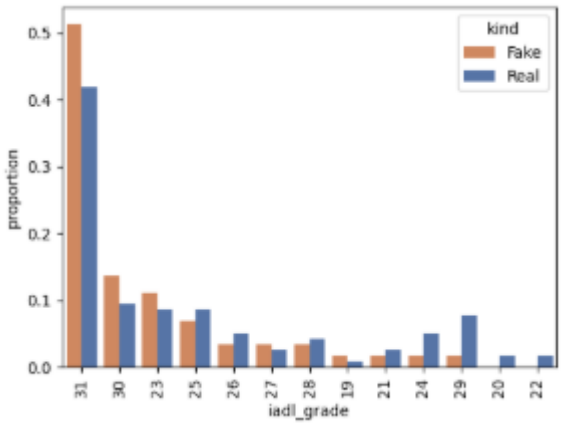


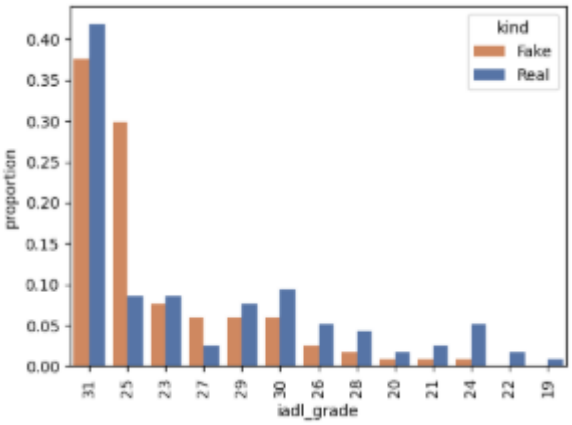


18. IADL grade

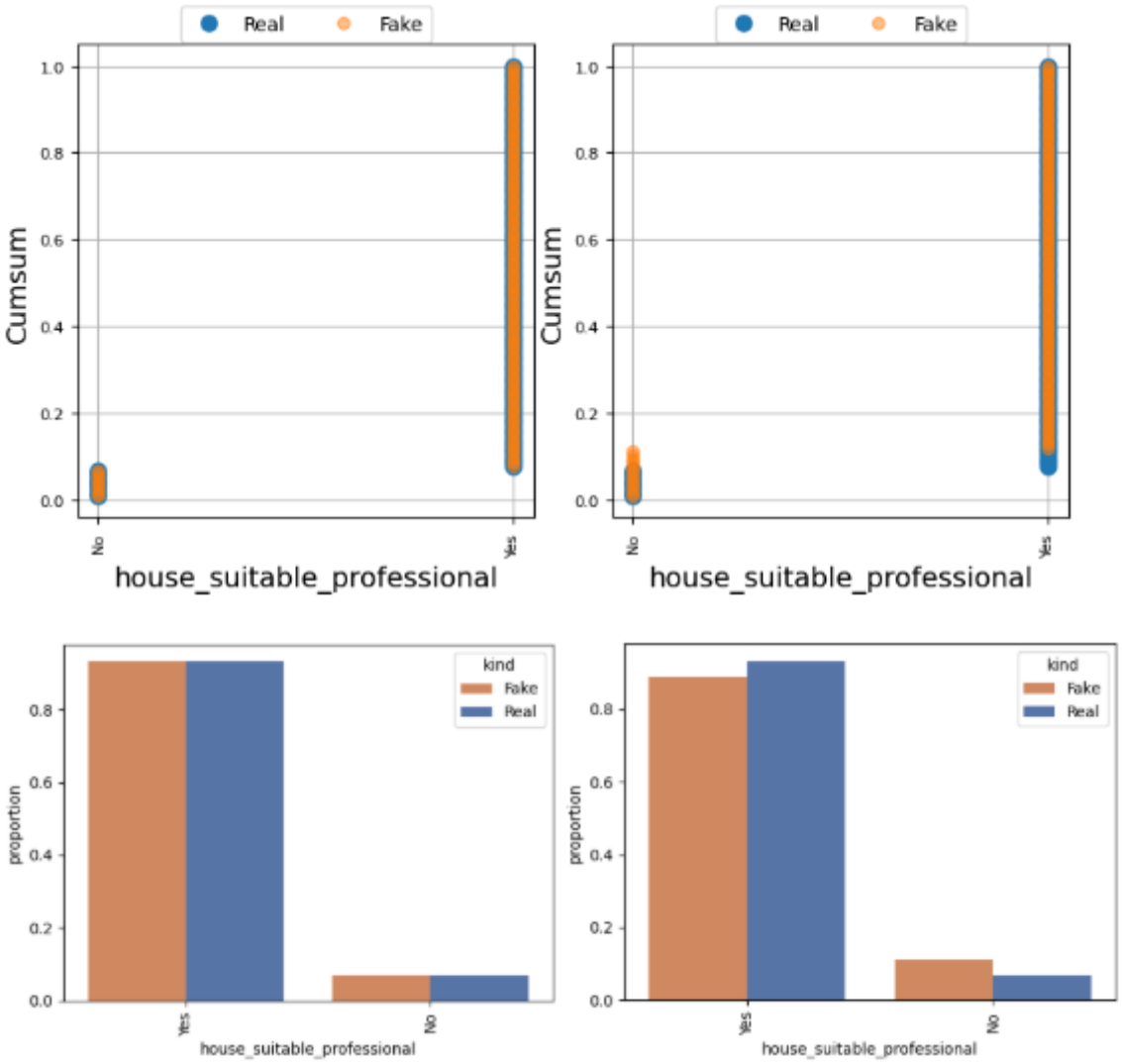




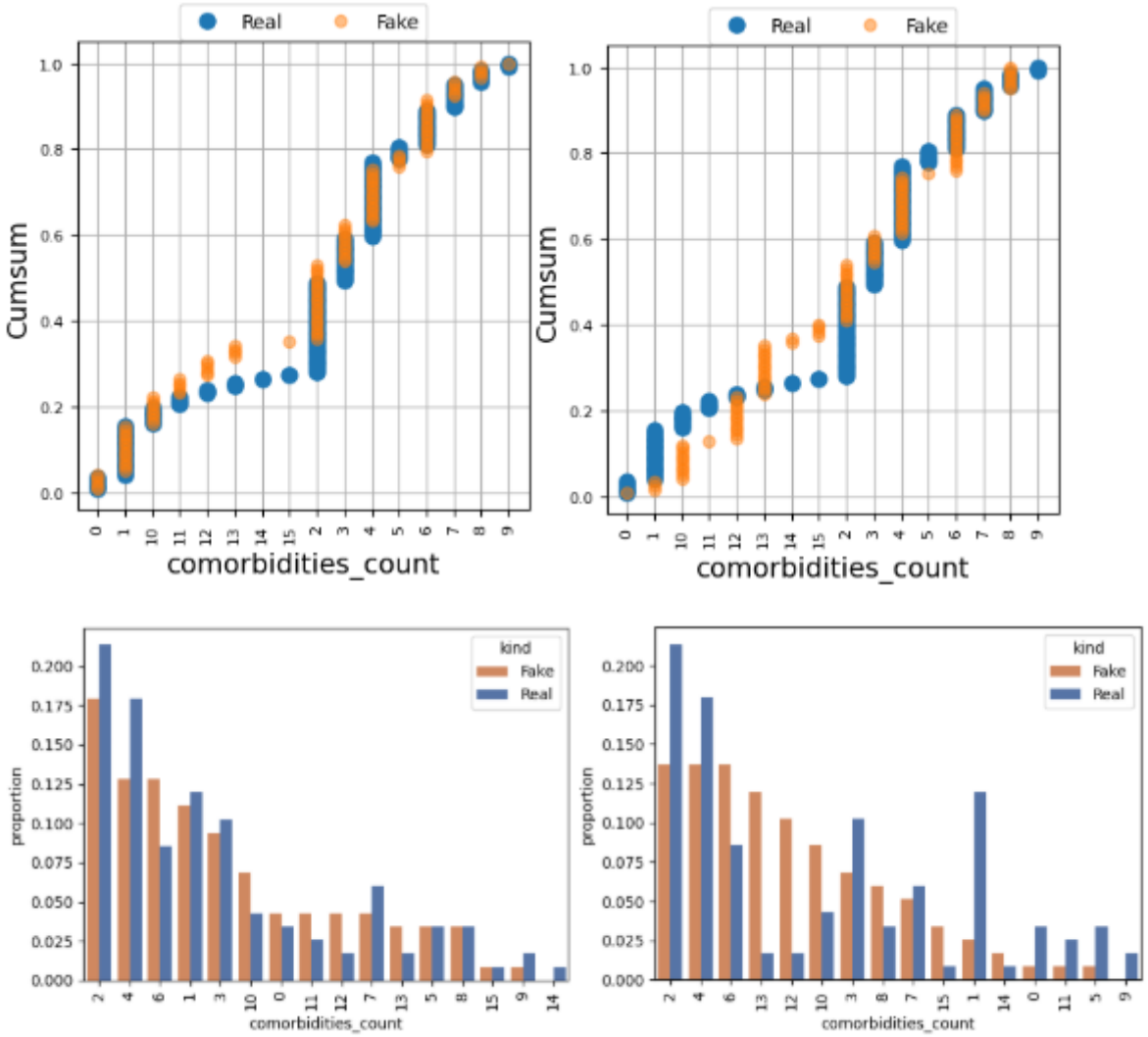




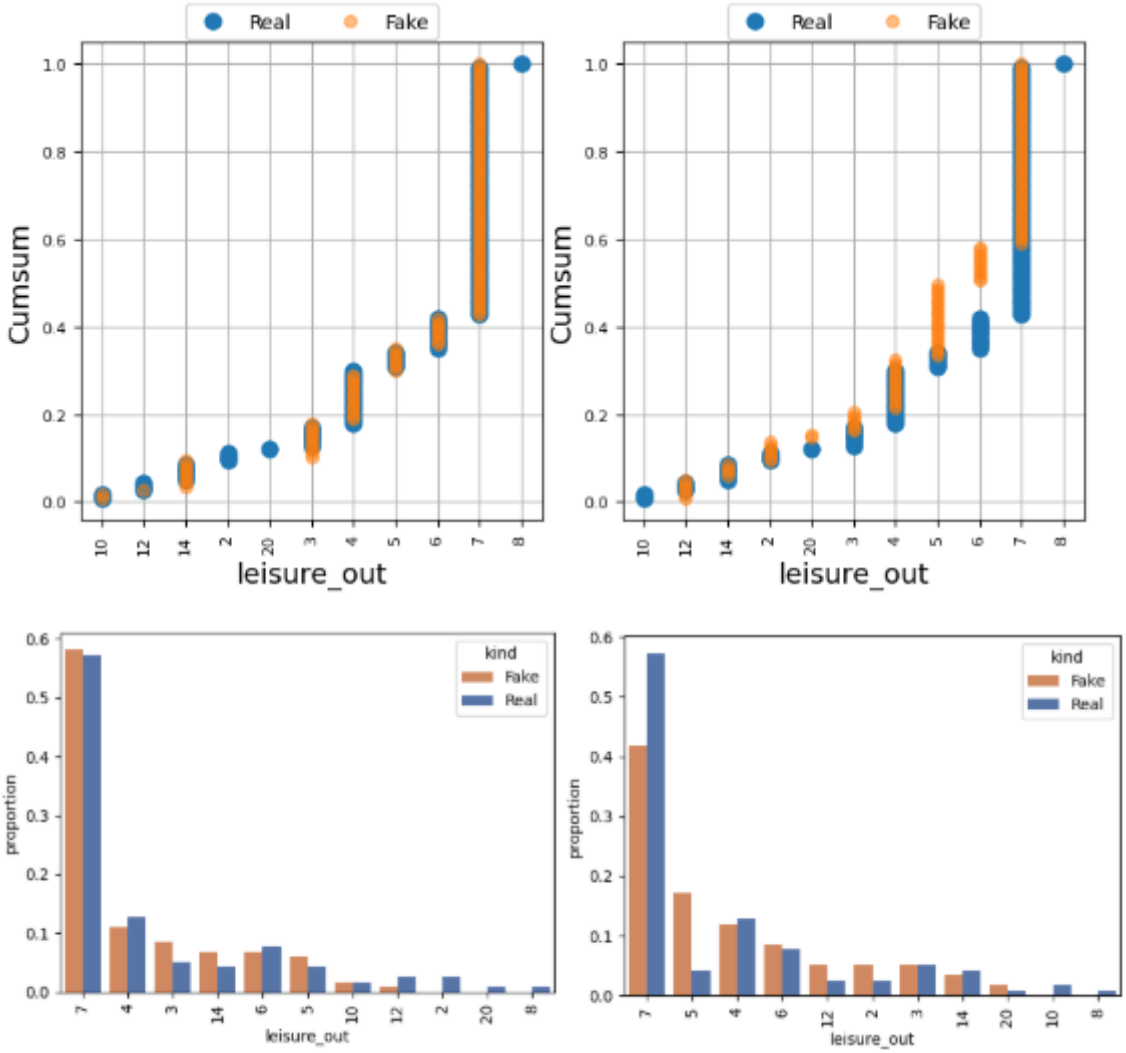
19. House suite professional



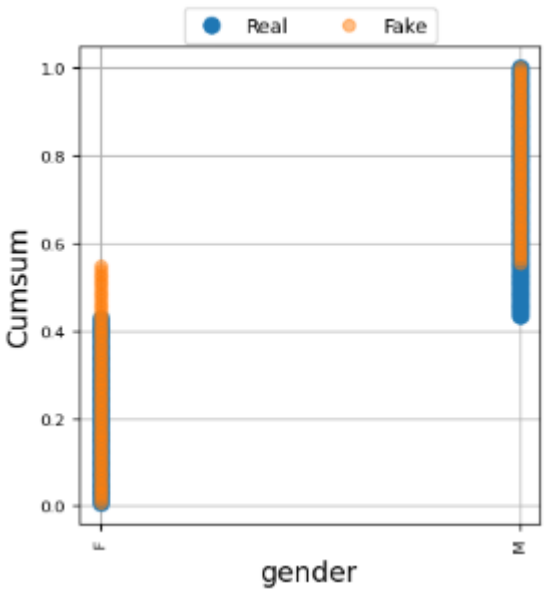
20. Comorbidities count

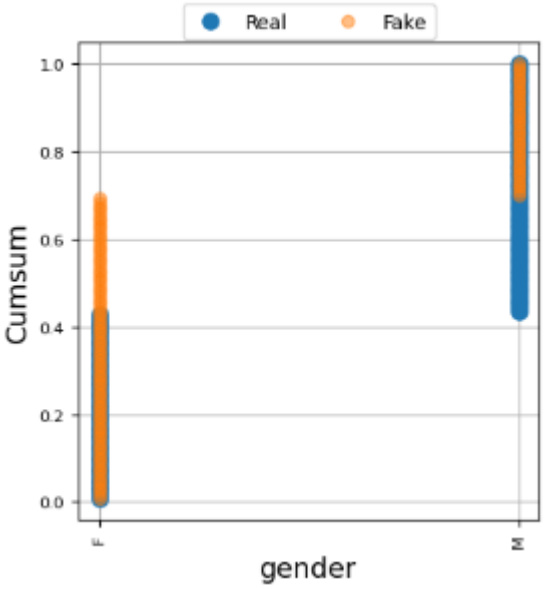


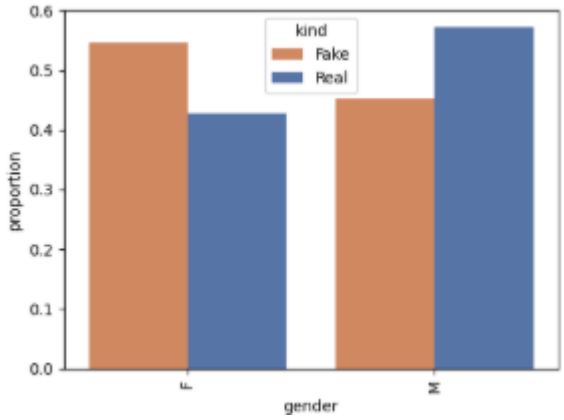
21. Leisure Out

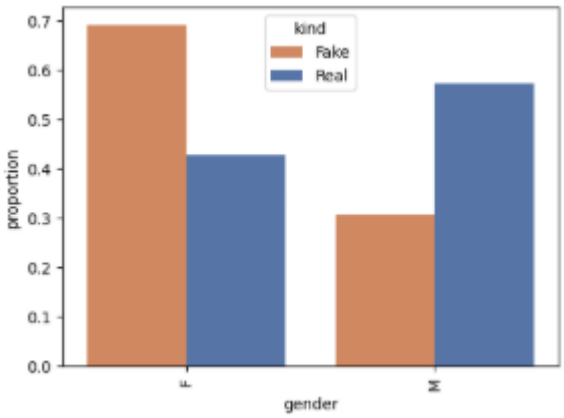


22. Gender

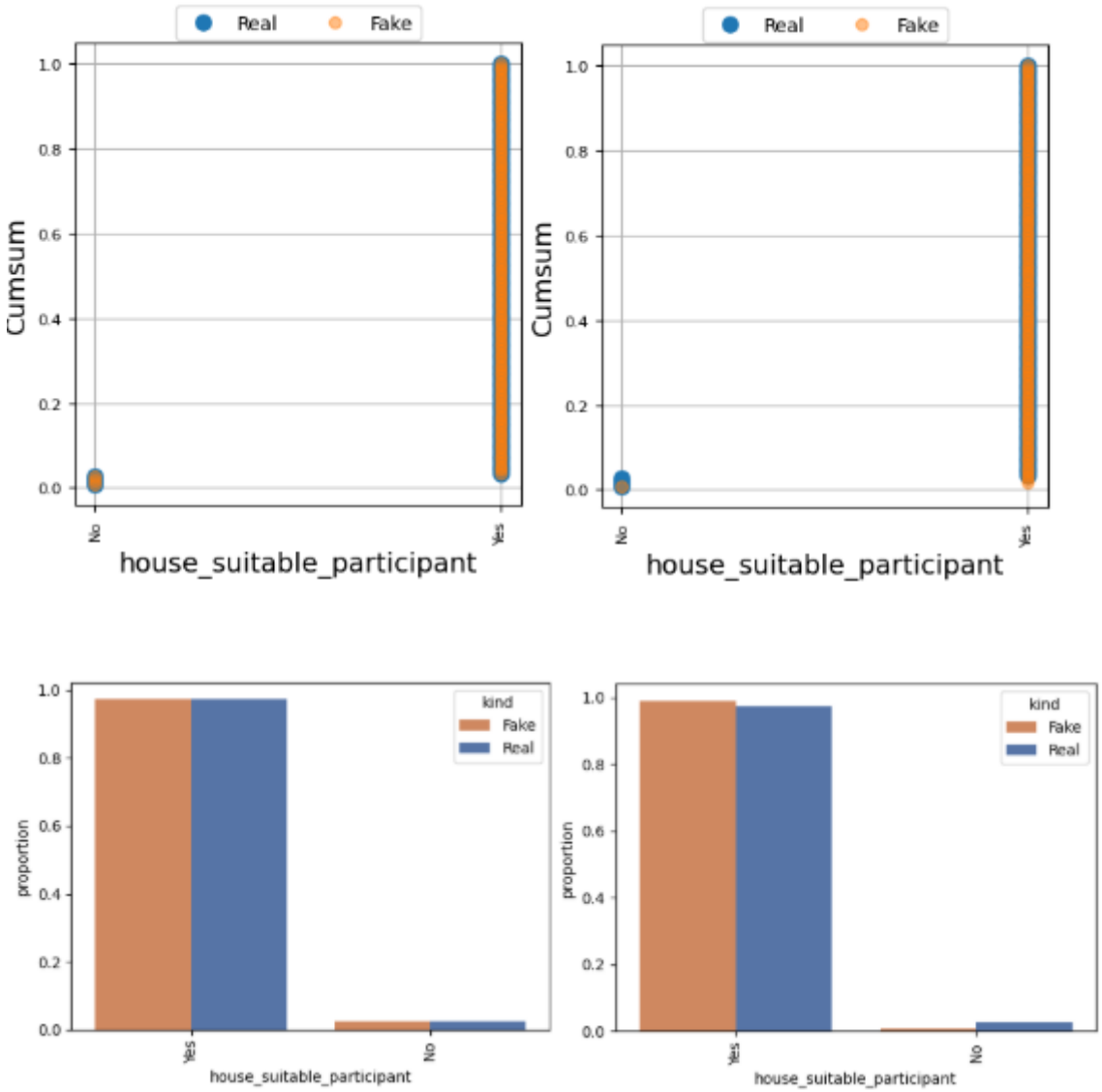

Real
Fake
Cumsum
1.0
0.8
0.6
0.4
0.2
0.0
F
M
gender


Real
Fake
Cumsum
1.0
0.8
0.6
0.4
0.2
0.0
F
M
gender

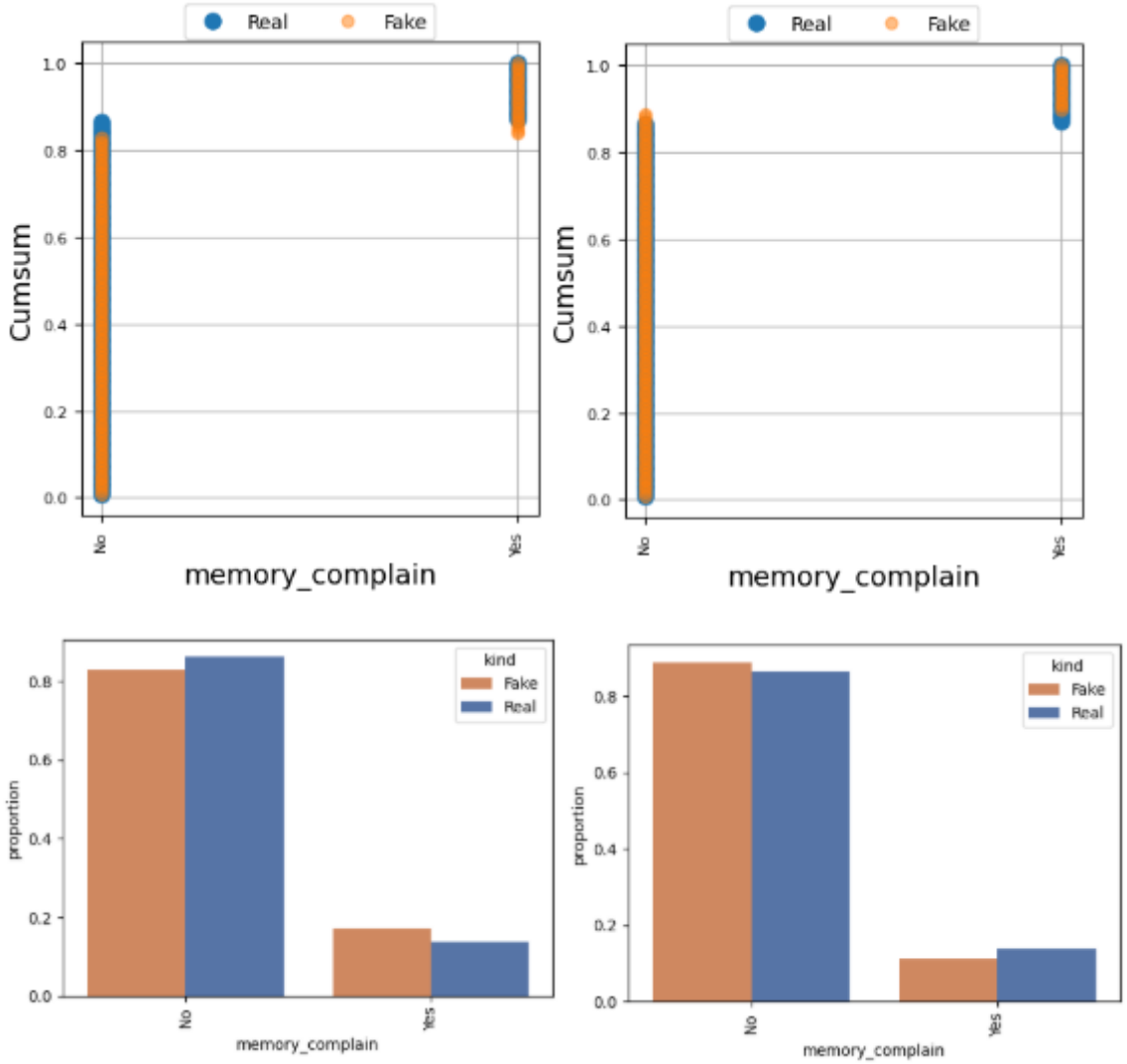

kind
Fake
Real
proportion
0.6
0.5
0.4
0.3
0.2
0.1
0.0
F
M
gender


kind
Fake
Real
proportion
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
F
M
gender

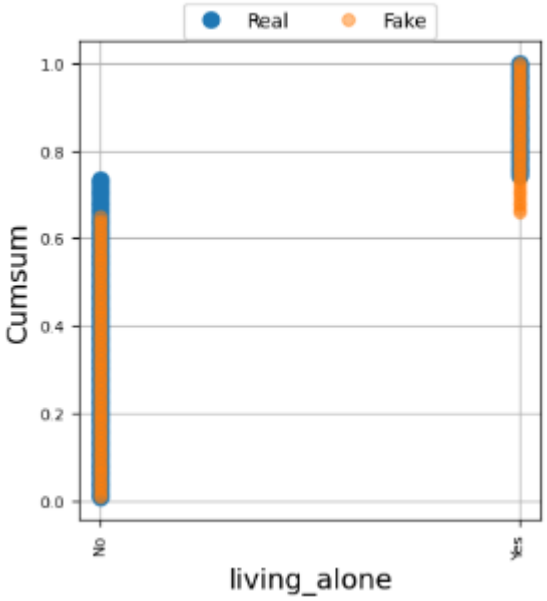
23. House suitable participant

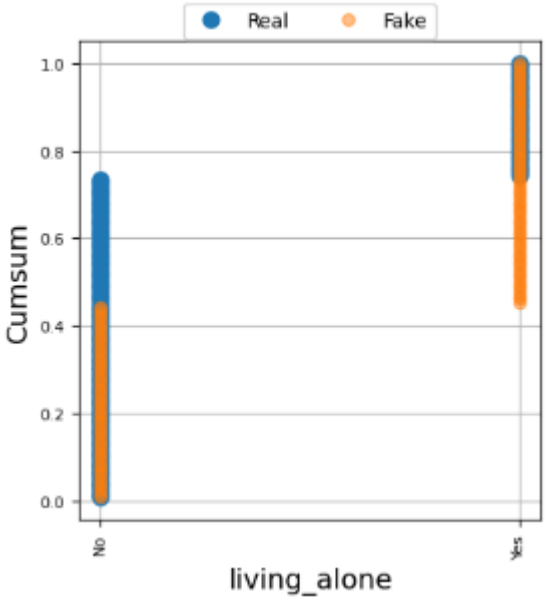


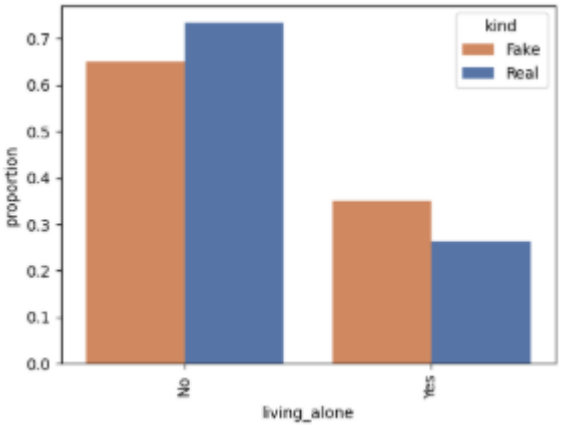
24. Memory Complain

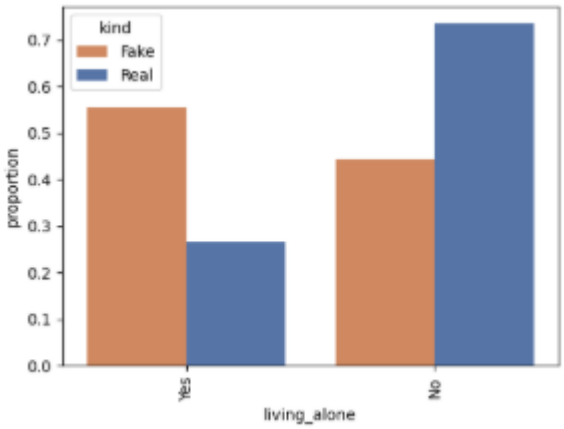


25. Living Alone

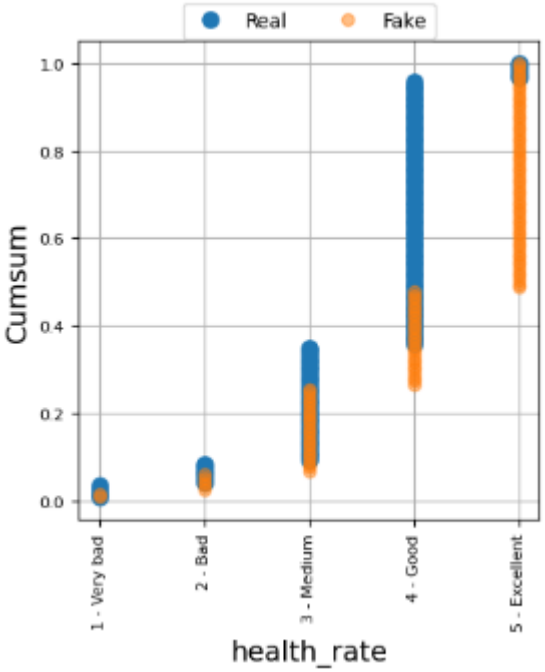


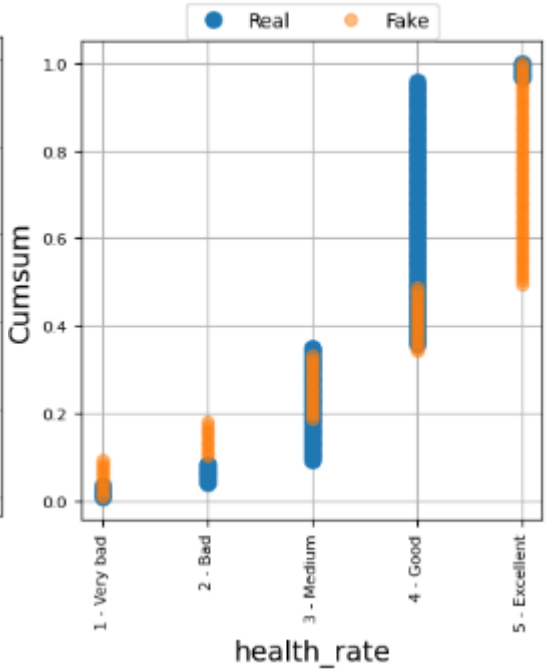


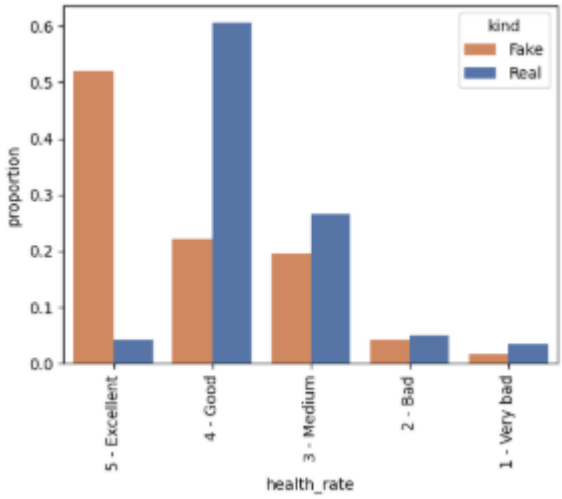


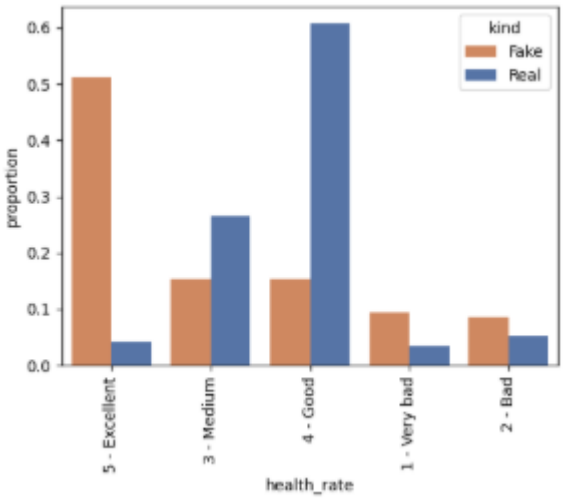


26. Health Rate

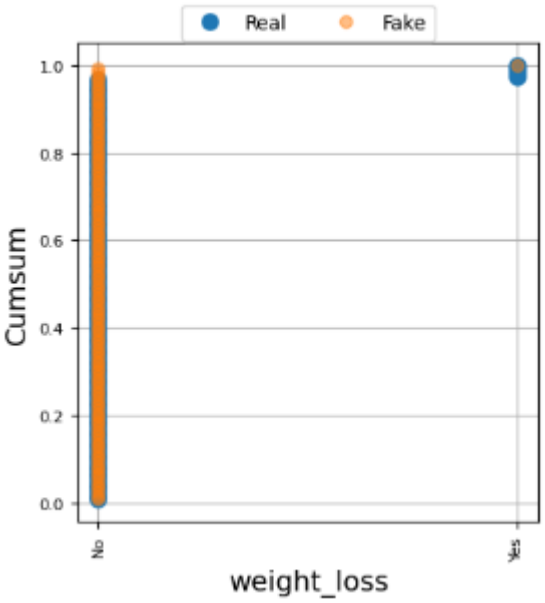


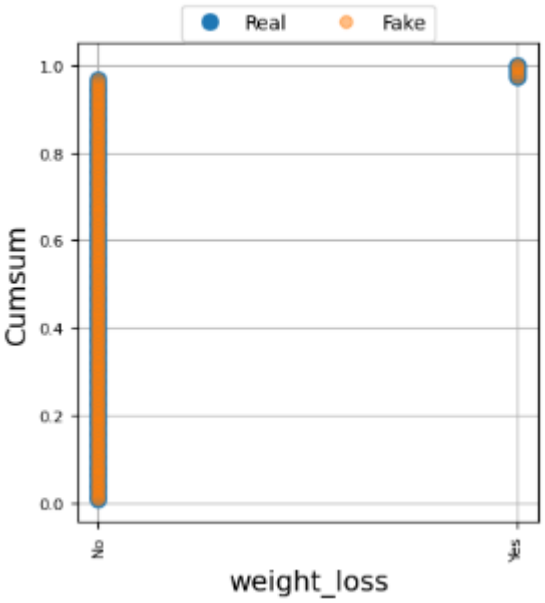


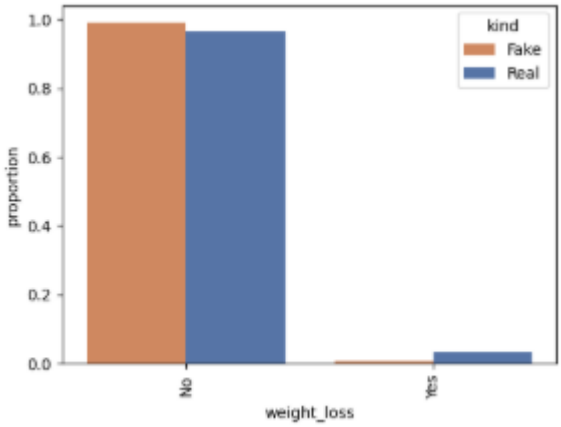


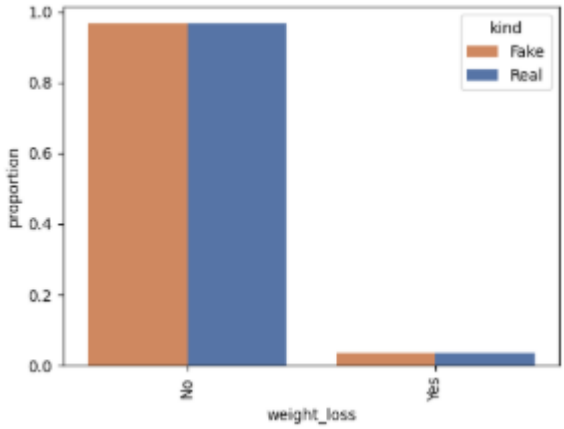


27. Weight loss

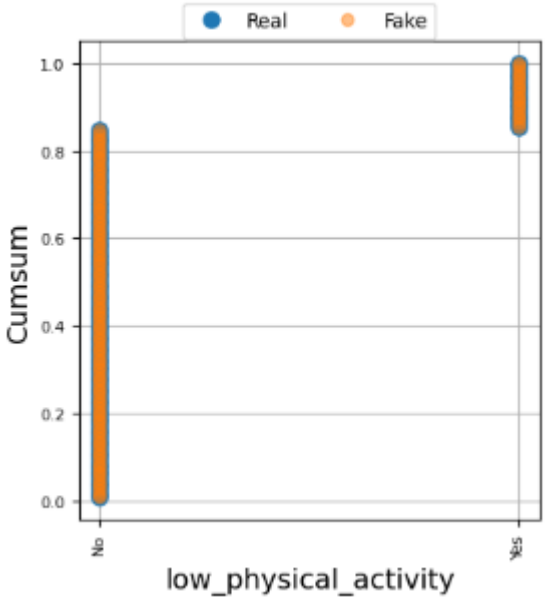


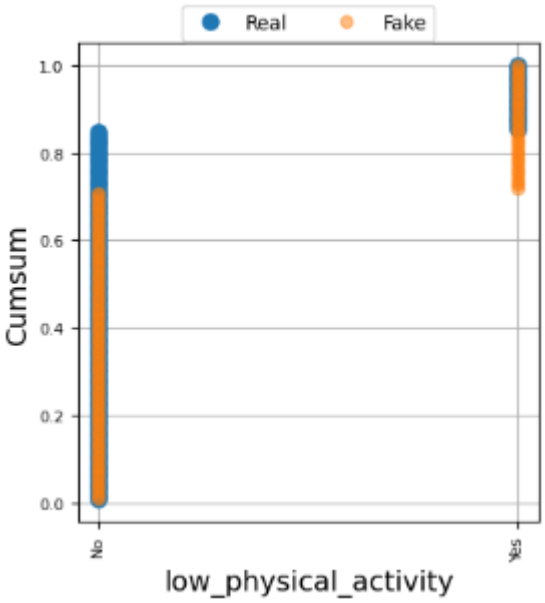


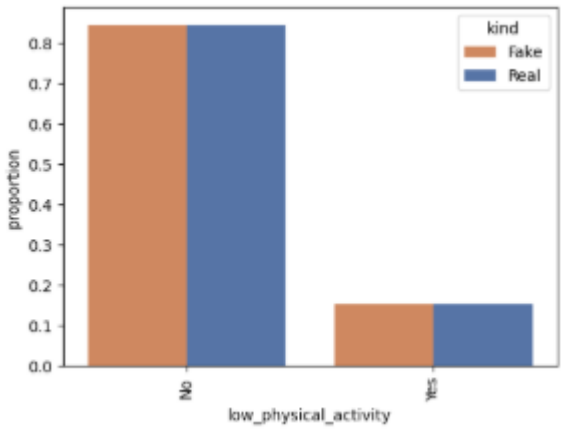


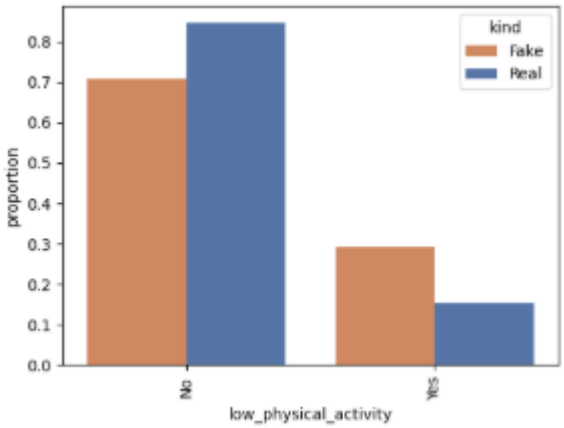


28. Low physical activity

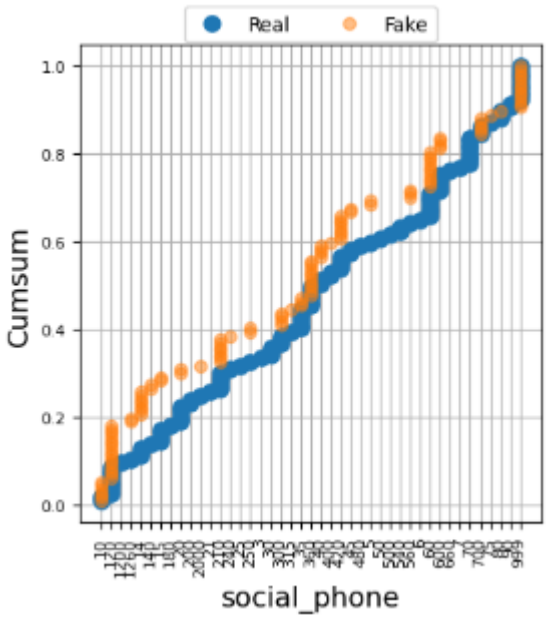


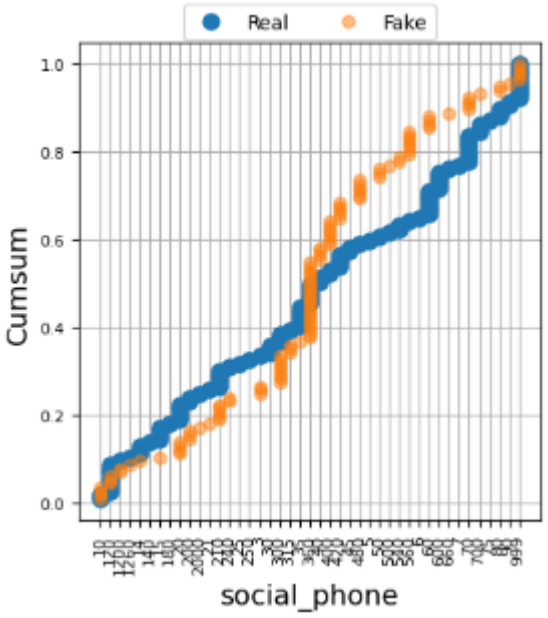


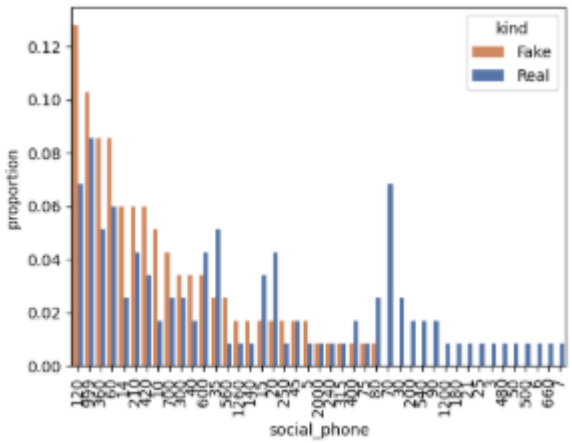


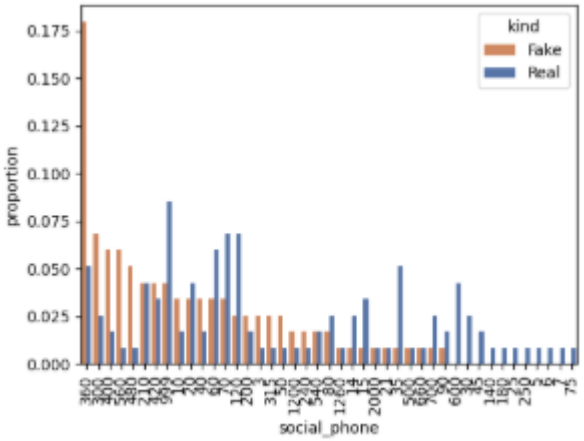


29. Social phone

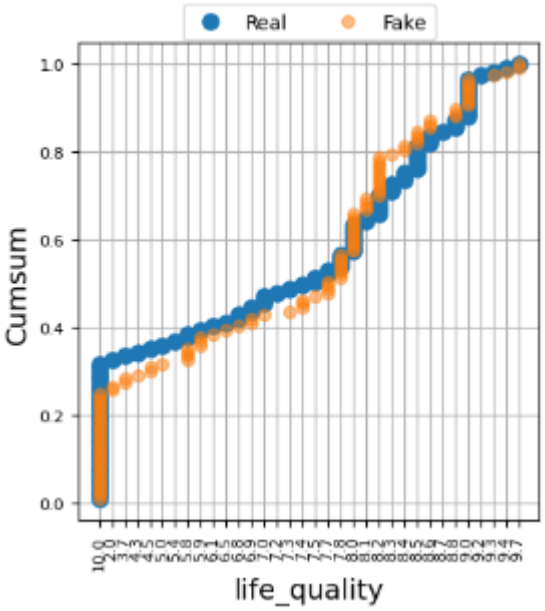


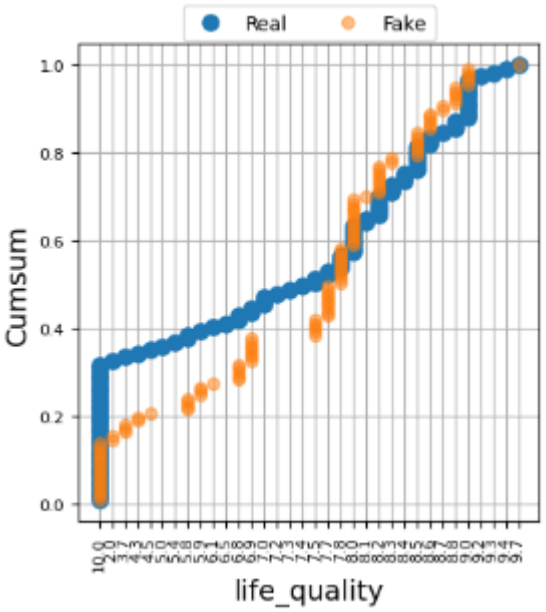


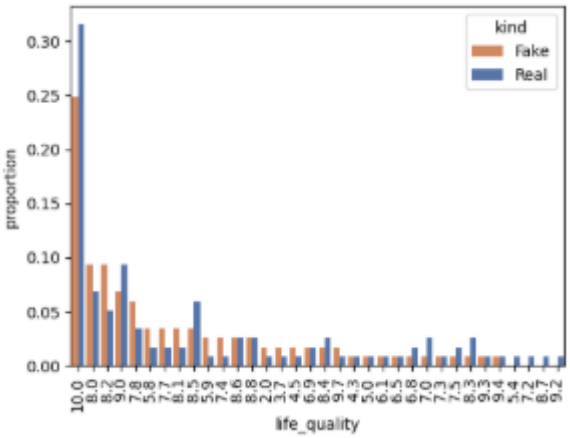


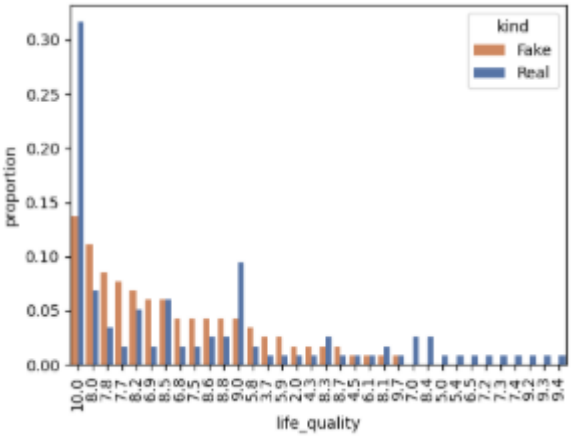


30. Life quality

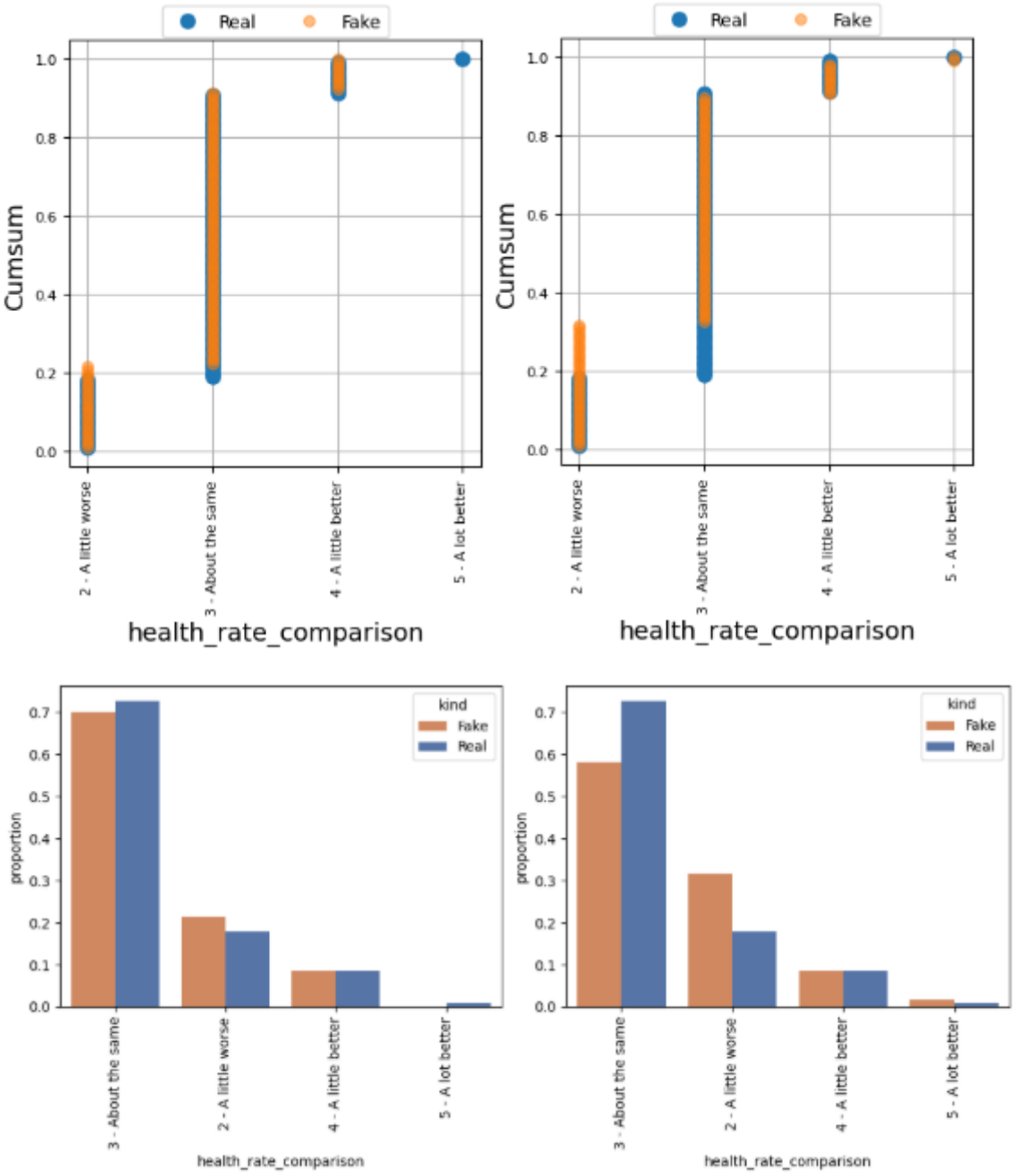




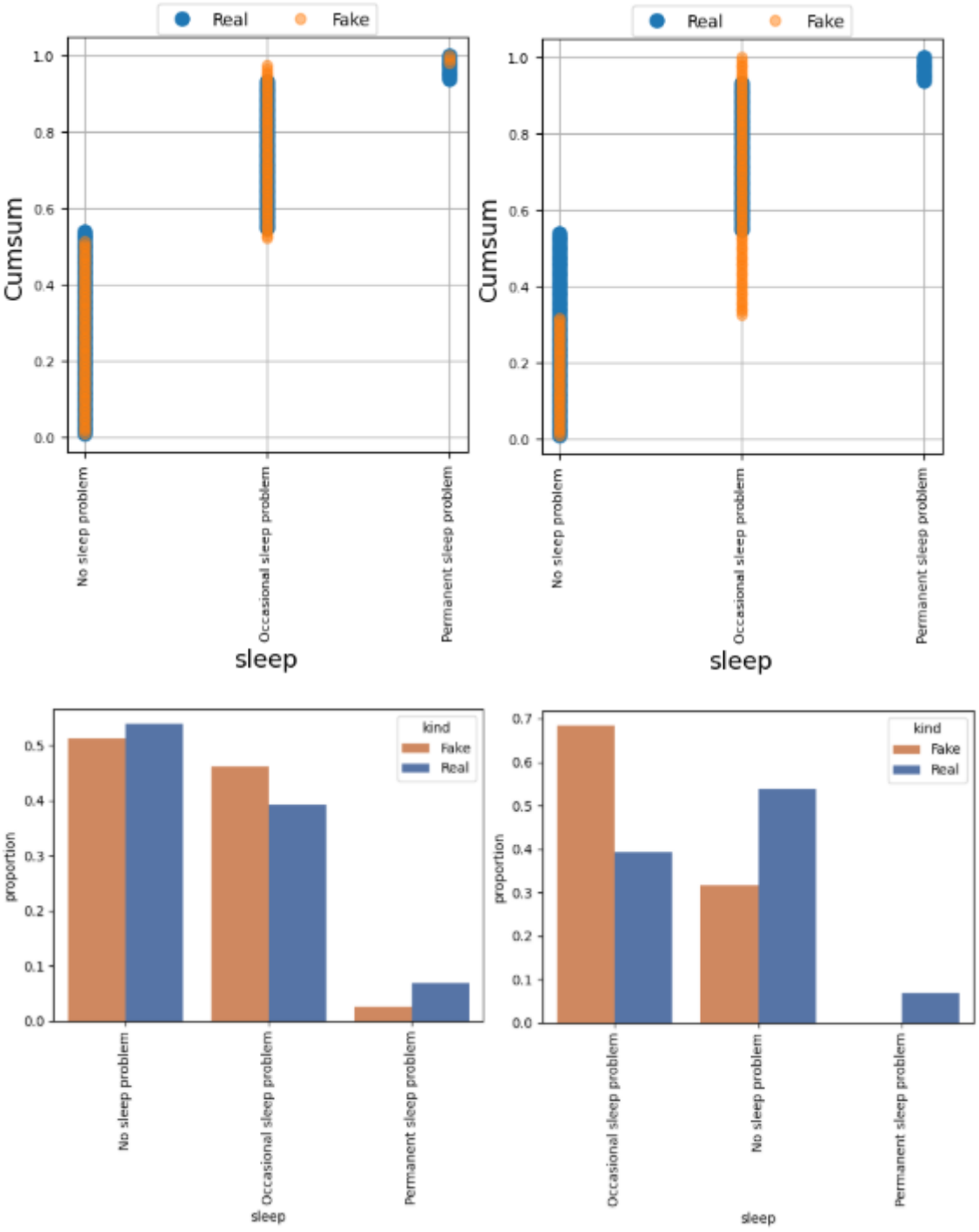




31. Health rate comparison



32. Sleep


Real
Fake
Cumsum
1.0
0.8
0.6
0.4
0.2
0.0
No sleep problem
Occasional sleep problem
Permanent sleep problem
sleep
kind
Fake
Real
proportion
0.7
0.5
0.3
0.1