Enhancing healthcare with AI - overcoming data preparation challenges

Mahalakshmi Lakshmi Nathan

MedCrave
Step into the World of Research

Enhancing healthcare with AI - overcoming data preparation challenges

Subtitle

AI-Driven Data Transformation for Healthcare Excellence

Authors Information

Dr. Mahalakshmi Lakshmi Nathan, MS, MBA, DHI

Doctor of Health Informatics, Department of Health Informatics Rutgers, The State University of New Jersey, School of Health Professions

Project Report Committee

1. Dr. Dasantila Sherifi, PhD
2. Dr. Shankar Srinivasan, PhD
3. External Project Advisor: Dr. K. Paul Jayakar, PhD

*Correspondance:

Mahalakshmi Lakshmi Nathan, Doctor of Health Informatics, Department of Health Informatics, Rutgers, The State University of New Jersey, School of Health Professions, Email mlnathan@outlook.com

Published By:

MedCrave Group LLC

February 20, 2026

Contents

Abstract

Healthcare data today is vast but fragmented, inconsistent, and frequently incomplete, limiting the effectiveness of artificial intelligence (AI) models built for clinical decision-making. The central problem addressed in this project is the persistent gap between the potential of AI in healthcare and the poor quality, semantic inconsistency, and lack of interoperability of the datasets on which such models depend. The overall objective was to design an adaptive framework capable of refining, standardizing, and harmonizing heterogeneous healthcare data to ensure reliability, interpretability, and compliance for predictive and diagnostic applications. The research evolved through five foundational studies and one integrated system development. The first study examined the inconsistencies in reporting alternative medicine treatments for diabetes and showed how lack of structured representation distorts meta-analytical outcomes. The second focused on diabetic readmission prediction using machine learning, revealing that model accuracy collapses when data is incomplete or biased. The third study on AI applications in orthopedics identified the dependence of clinical models on data annotation quality. The fourth and fifth studies explored medical imaging and multimodal AI integration, demonstrating that transformer-driven harmonization and feature alignment significantly improve interpretability and robustness across diverse modalities. Building on these findings, the final stage introduced the **Multi-Stage AI Data Refinement Network (MADR-Net)** an adaptive pipeline that combines deep generative and sequential models for missing data imputation, duplicate detection, outlier correction, data augmentation, and semantic standardization using FHIR and SNOMED mappings. Evaluations on large-scale healthcare admission datasets confirmed substantial gains, with macro-precision, recall, and F1-scores exceeding 0.96, alongside measurable reductions in class imbalance and structural bias. The study concludes that the primary barrier to dependable healthcare AI is data quality, not algorithmic sophistication. By embedding intelligence throughout the preprocessing pipeline, MADR-Net redefines data preparation as a foundational, auditable stage in healthcare analytics transforming fragmented clinical information into standardized, trustworthy, and ethically governed data for next-generation AI systems.

Keywords: data imputation, duplicate detection, outlier correction, data augmentation, SNOMED mappings, medical imaging, AI, next-generation AI systems

Abbreviations: AI, artificial intelligence; AUROC, area under the receiver operating characteristic curve; CDM, common data model; HER, electronic health record; ETL, extract transform load; FHIR, fast healthcare interoperability resources; GAN, generative adversarial network; HL7, health level seven international; ICD, international classification of diseases; JSON, javascript object notation; LOINC, logical observation identifiers names and codes; MNAR, missing not at random; OMOP, observational medical outcomes partnership; PHI, protected health information; RDF, resource description framework; SEM, structural equation modeling; SNOMED CT, systematized nomenclature of medicine clinical terms; VAE, variational autoencoder

Chapter - 1

Introduction

Background

Healthcare data today exists in a highly fragmented ecosystem where information is dispersed across diverse electronic health record (EHR) systems, diagnostic repositories, and clinical documentation sources. Each of these systems follows its own schema, vocabulary, and storage protocol, resulting in severe challenges to interoperability and large-scale analytics. Studies have shown that fragmented EHR infrastructures lead to data redundancy, inconsistency, and limited portability, restricting the usability of data for evidence-based decision-making and artificial intelligence (AI) applications (Nathan & Sherifi, 2025). This fragmentation has become a critical bottleneck for healthcare innovation, where data volume is abundant but actionable intelligence remains scarce.

Given these issues, developing a Common Data Model (CDM) and a unified integration pipeline capable of harmonizing healthcare data across disparate EHR systems seemed like a natural direction for research. However, during early experimentation, significant barriers emerged particularly those concerning protected health information (PHI) and stringent data privacy policies that restricted direct access to real-world patient datasets. Consequently, the feasibility of deploying large-scale, real-time integration pipelines in live hospital environments became limited.

While addressing these limitations, the focus gradually shifted from *data integration* to *data intelligence*. A deep examination of the literature and prior project experiences revealed that the central challenge in healthcare analytics is not the lack of models but the lack of trustworthy data. Many AI-driven systems in clinical practice fail, not because of algorithmic weakness, but because the underlying data lacks quality, completeness, or standardization. Thus, rather than forcing data into rigid schemas through traditional ETL (Extract–Transform–Load) processes, a more intelligent, adaptive, and self-refining framework became necessary.

Research motivation

In the course of analyzing data fragmentation and its downstream effects on clinical prediction systems, I conducted several domain-specific investigations that shaped the foundation of this research. Each study exposed a different dimension of the healthcare data problem, finally converging toward the realization that *the future of healthcare AI lies in data refinement rather than mere data accumulation.*

In the first study titled "Diabetes Patients' Readmission Prediction," carried out with Professor Dr. Dasantia and Dr. Veerabahu, I developed and evaluated multiple machine-learning models for predicting 30-day hospital readmissions among diabetic patients using ten years of electronic health record (EHR) data. The study compared algorithms such as Random Forest, Gradient Boosting, and Logistic Regression across varying data completeness levels. Although Gradient Boosting achieved the highest predictive accuracy, the models' reliability decreased markedly in the presence of incomplete or noisy records. This work demonstrated that model precision depends less on algorithmic complexity and more on the integrity and consistency of the input data.

The second study, titled **"**Data Intelligence Through Integration in Healthcare: Research Gaps and Opportunities," conducted in collaboration with Professor Dr. Dasantila Sherifi, examined how large-scale healthcare datasets can be effectively unified to strengthen data-driven clinical decision-making. This work explored diverse data sources electronic health records (EHRs), clinical trials, wearable sensors, social media, and billing systems to identify methods of integrating fragmented healthcare information into coherent, interoperable structures. Through a systematic literature review, the study analyzed how artificial intelligence, machine learning, and data integration frameworks contribute to developing a unified healthcare data ecosystem. The findings emphasized that effective data intelligence depends not only on the volume or variety of data but also on the semantic consistency, data governance, and contextual interoperability between systems. The research concluded that addressing current integration gaps requires adopting adaptive data intelligence strategies that balance analytical capability with ethical governance and technical scalability.

The third study, titled **"**A Fresh Look: The Role of a Healthcare Data Fabric in AI-Driven Predictions**,"** conducted in collaboration with Dr. Shankar Srinivasan, explored the transformative potential of healthcare data fabrics in enhancing AI-based clinical prediction systems. The research examined how distributed data sources ranging from electronic health records and genomic databases to imaging repositories and wearable sensor data can be unified under a scalable, metadata-driven architecture. This study demonstrated that the persistent issues of interoperability gaps, uneven data quality, and fragmented governance significantly limit the operational value of healthcare analytics. By implementing a healthcare data fabric, the work proposed a framework that links disparate datasets through semantic integration layers, metadata catalogs, and virtualization services, thereby allowing real-time, secure, and regulatory-compliant access to harmonized clinical data. The findings established that such architectures not only improve AI-driven predictions and decision support but also advance lineage tracking, policy enforcement, and overall data governance. This contribution reinforced the argument that sustainable healthcare AI depends on architectural intelligence that ensures both interoperability and ethical compliance at scale.

The fourth study, titled **"**Review of Alternative Medicine (AM) Treatments for Diabetes**,"** conducted in collaboration with Professor Dr. Dinesh Mital, investigated the efficacy, limitations, and clinical implications of non-conventional therapeutic approaches in diabetes management. The study examined the growing adoption of alternative medicine including diet therapy, herbal supplements, yoga, meditation, and other lifestyle-based interventions as complementary strategies to conventional medical treatment. Through a structured literature assessment and comparative analysis of available evidence, the research identified significant inconsistencies in reporting standards, dosage formulations, and clinical outcome documentation across AM studies. These disparities underscored the absence of standardized evaluation frameworks and limited regulatory oversight, particularly concerning the safety and effectiveness of herbal and nutraceutical formulations. The findings highlighted that while AM therapies contribute positively to lifestyle

modification and glycemic control, their clinical validation remains fragmented due to methodological variability and lack of harmonized data collection. This study reinforced the broader premise that healthcare innovation whether pharmacological or AI-driven relies fundamentally on the quality, traceability, and standardization of underlying clinical data to ensure scientific credibility and patient safety.

The fifth study titled "Uses, Benefits, and Future of Artificial Intelligence [AI] in Orthopedics," conducted with Dr. Veerabahu Muthusamy, explored the emerging role of artificial intelligence in orthopedic applications such as diagnostic imaging, prosthetic modeling, and postoperative rehabilitation prediction. The analysis revealed that the majority of AI systems deployed in orthopedic practice failed not due to architectural limitations but because of poor data annotation, heterogeneous imaging formats, and lack of standardized evaluation metrics. This reinforced the understanding that even domain-specific AI models require refined, semantically coherent datasets to ensure dependable and clinically valid outcomes.

Problem definition

Healthcare analytics depends fundamentally on structured, high-quality data. Yet, clinical datasets are characterized by pervasive issues such as missing entries, redundant patient records, inconsistent measurement units, semantic ambiguity, and outlier corruption. These problems undermine the statistical validity and ethical soundness of AI-based decision support systems.

Traditional data preprocessing techniques such as mean imputation, rule-based cleaning, or duplicate filtering are insufficient for addressing these complex challenges. They neither capture contextual dependencies among clinical features nor adaptively handle semantic variation across healthcare institutions. Furthermore, the absence of universal data representation standards exacerbates interoperability failures between hospitals, laboratories, and insurance systems. This study therefore positions itself at the intersection of AI-driven data refinement and healthcare data governance, aiming to design a scalable and intelligent framework that not only cleans but also *contextually understands* healthcare data. This conceptual transition from static preprocessing to *cognitive refinement* forms the central motivation of the research.

Initially, the project was conceptualized as a data integration pipeline designed to unify hospital data streams through standardized schema mapping and ontology alignment. However, during preliminary investigation, it became evident that enforcing rigid data conformity was neither sustainable nor scalable under the constraints of real-world health systems. PHI regulations, institutional policy fragmentation, and heterogeneous record structures rendered full integration unfeasible. This realization marked a pivotal shift from system-level integration to intelligence-level refinement. Instead of imposing external uniformity, the proposed model would enable the data itself to evolve through multiple refinement stages, guided by AI. The hypothesis was simple but powerful: *if data can be autonomously corrected, augmented,* ***Dataset Characteristics:*** *harmonized, and validated before modeling, then AI predictions can become both more reliable and more interpretable.*

From this premise, the Multi-Stage AI Data Refinement Network (MADR-Net) was conceptualized. The architecture was designed to sequentially address each aspect of healthcare data degradation starting from missing values and duplicates to class imbalance and semantic inconsistencies by embedding machine learning intelligence directly into each stage of preprocessing. This multi-stage approach was expected to yield datasets that are statistically consistent, semantically coherent, and analytically ready for downstream prediction.

Overview of the proposed framework

The MADR-Net framework integrates five major AI paradigms into a unified data refinement pipeline:

I. A Generative Adversarial Network (GAN) is employed to synthetically expand underrepresented diagnostic categories, thereby addressing class imbalance without compromising data authenticity. The generated samples preserve complex cross-feature dependencies, ensuring fairer and more stable model learning.

II. To address data incompleteness, a hybrid Variational Autoencoder–GAN (VAE–GAN) model is proposed. This hybrid imputation mechanism leverages the probabilistic estimation capabilities of VAE and the adversarial refinement of GAN to reconstruct missing attributes with high statistical and semantic fidelity.

III. The proposed system integrates a Bidirectional LSTM-based similarity discriminator that learns temporal and contextual dependencies within patient records. This module identifies near-duplicate entries and redundant admissions, ensuring dataset uniqueness and improving the reliability of subsequent analyses.

IV. A context-aware autoencoder-based anomaly detector is implemented to identify irregular or inconsistent data patterns. The model computes reconstruction error and latent space deviation to distinguish genuine rare cases from erroneous records, minimizing analytical distortion caused by outliers.

V. The framework proposes a Transformer-based harmonization layer that maps heterogeneous healthcare attributes to standardized medical ontologies such as FHIR, SNOMED CT, LOINC, and RxNorm. This alignment guarantees semantic consistency, interoperability, and ethical data integration across institutional systems.

Chapter - 2

Literature survey

This chapter provided a critical synthesis of contemporary empirical research addressing the multifaceted challenges of healthcare data preparation for artificial intelligence applications. The literature reveals three dominant thematic directions that have shaped recent advancements: (i) the standardization of heterogeneous clinical data to ensure interoperability, semantic fidelity, and structural uniformity across diverse systems; (ii) the remediation of incompleteness, redundancy, and noise to enhance analytical validity and reliability; and (iii) the verification of dataset integrity and modeling readiness to guarantee trustworthy and reproducible outcomes. Collectively, these studies underscore a paradigm shift from isolated data cleaning operations toward comprehensive, multi-stage refinement pipelines that embed

quality assurance and interpretability as integral components of the preparation process. The insights consolidated through this review establish the conceptual foundation for the subsequent chapter, which empirically investigates these three core workstreams under the theme *"Empirical Advances in Data Preparation for Healthcare AI."*

Introduction

Healthcare enterprises now capture data at a scale and granularity that were inconceivable a decade ago. Electronic health records aggregate longitudinal encounters, orders, laboratory panels, flowsheets, problem lists, medications, discharge summaries, and billing artifacts. Imaging archives contribute pixel data and structured metadata across modalities, while bedside monitors and wearables stream high-frequency physiologic signals. Claims, registries, and patient-reported outcomes add further strata. Yet these assets do not arrive as analysis-ready inputs. They are produced by different vendors, follow divergent documentation practices, and reflect the operational pressures of clinical care rather than the tidy assumptions of machine learning pipelines. Consequently, the decisive determinant of model reliability is not the choice of architecture but the quality and discipline of data preparation that precedes it.[1,2] Three properties of clinical data drive this dependence. First, heterogeneity is structural, not incidental. The same clinical concept appears under multiple coding systems (ICD, SNOMED CT, RxNorm, LOINC) and local catalogues with idiosyncratic abbreviations, unit conventions, and context qualifiers. Second, temporality is messy. Second, events have an "order time," a "collection time," and a "result time," which may be hours or days apart. Charting delays, back-dated entries, and edits create discrepancies between when something happened and when it was recorded. Third, sparsity is informative. Missingness in clinical datasets is rarely random; it reflects workflow, triage, and clinician judgment. A lactate not ordered at 3 a.m. can carry as much meaning as a measured value. Treating these features as mere nuisances rather than signals invites bias and instability.[3,4]

Effective preparation therefore begins with principled normalization, not ad-hoc cleaning. Semantic alignment is required to map local terms and free-text variants to shared vocabularies while preserving clinical nuance (e.g., fasting versus random glucose; arterial versus venous blood gas). Unit harmonization must be coupled to reference ranges and specimen sources to avoid silent errors converting µg/dL to mmol/L without accounting for assay-specific factors is a common failure mode. At the sequence level, encounter stitching and identity resolution are essential to avoid splitting a single patient trajectory across multiple medical record numbers or, worse, merging different individuals. Probabilistic linkage and privacy-preserving techniques are often needed when datasets span institutions.[5,6] Temporal preparation is equally non-negotiable. Aligning measurements to clinically meaningful windows (pre-op, intra-op, post-op; pre-ICU versus first 24-hour ICU) reduces label leakage and clarifies the cause–effect ordering that predictive models implicitly exploit. Event aggregation rules maximum vasopressor dose within a window, last creatinine before discharge, nadir hemoglobin during admission should be defined up-front and versioned. For text, contextualization means more than tokenization: it requires detection of negation, uncertainty, experiencer (patient vs family history), and temporality within notes to avoid constructing features that assert what the note explicitly denies.[7,8]

Handling incompleteness demands methods matched to mechanism. Mean substitution and blanket model-based imputation ignore the fact that clinical missingness is frequently "missing not at random": tests are omitted because they are deemed unnecessary, not because of chance. Variable-wise strategies that condition on ordering patterns, care setting, and preceding results, alongside explicit missingness indicators, often outperform monolithic approaches. Where no plausible imputation exists, conservative censoring, domain-bounded fills, or model architectures designed to consume masks directly are preferable to fabricating data.[9,10] Anomaly management should distinguish implausible values from genuine physiology. Outlier filters that indiscriminately cap extremes will erase exactly the events shock states, toxic levels, rare complications that models must learn. Plausibility checks should incorporate unit-aware bounds, inter-variable constraints (e.g., pulse pressure cannot be negative), and patient-level trajectories. Flagged records warrant a triage path: correctable transformation errors, clinically explainable extremes retained with provenance, and irreparable artifacts excluded with justification.[11,12]

This review proceeds from that premise. It synthesizes empirical work on harmonizing heterogeneous clinical datasets, restoring completeness without distorting signal, validating internal consistency, and quantifying the effect of preparation decisions on model behavior. The goal is to separate broadly useful practices from brittle shortcuts, expose where current methods fall short semantic alignment at scale, mechanism-aware missingness, auditable linkage, end-to-end evaluation and motivate frameworks that make data preparation a first-class, reproducible component of trustworthy healthcare AI.

Search strategy and selection criteria

Databases, timeframe, and scope

The review targeted original research on data preparation in healthcare published between 1 January 2021 and 31 March 2025. Searches were executed in PubMed, Scopus, IEEE Xplore, and ScienceDirect to balance clinical coverage (PubMed), broad scholarly indexing (Scopus), and method-oriented outlets (IEEE Xplore, ScienceDirect). Only peer-reviewed journal articles were considered; preprints, conference proceedings, workshop papers, case reports, editorials, and narrative or systematic reviews were excluded to maintain an empirical, methods-focused corpus.

Query design and boolean strings

Queries combined concept blocks for healthcare context, preparation tasks, and empirical validation. Fielded and proximity operators were used where supported to limit noise. Representative strings (customized per database syntax) included:

I. ("electronic health record" OR EHR OR "clinical data" OR "healthcare data") AND ("data preparation" OR preprocessing OR "data wrangling" OR "data harmonization" OR "standardization" OR "schema mapping" OR "unit normalization" OR "terminology mapping" OR "FHIR" OR "OMOP") AND (study OR experiment OR evaluation)

II. (EHR OR "clinical registry" OR "claims data") AND ("missing data" OR imputation OR "MNAR" OR

"informative missingness" OR "duplicate" OR linkage OR "record linkage" OR "anomaly detection" OR "outlier") AND (method* OR framework) AND (trial OR cohort OR retrospective)

III. ("clinical text" OR "discharge summary" OR "radiology report") AND (de-identification OR "differential privacy" OR anonymization) AND (evaluation OR benchmark)

Searches were limited to title/abstract/keywords where possible to improve precision; medical-subject headings and controlled vocabulary (e.g., MeSH, Emtree) were added in PubMed/Scopus for "Data Curation," "Terminology as Topic," "Natural Language Processing," "HL7 FHIR," and "Common Data Model."

Inclusion and exclusion criteria

Inclusion

Original, peer-reviewed journal articles.

I. Primary contribution addresses at least one data preparation task in healthcare: interoperability/standardization (e.g., FHIR/OMOP ETL, terminology/units), missingness (mechanism-aware imputation, mask-aware modeling), identity resolution/linkage/deduplication, anomaly/outlier handling, quality validation/readiness for downstream analytics.

II. Empirical evaluation on clinical or clinically sourced data (EHR, registries, claims, clinical text, physiologic time series, imaging metadata).

III. Reported, auditable methods (algorithms, rules, ETL specifications) and quantitative outcomes (e.g., mapping accuracy, imputation error, linkage precision/recall, plausibility/conformance rates, effect on prediction metrics).

IV. English language.

Exclusion

I. Reviews, tutorials, surveys, opinion pieces, commentaries, guidelines without empirical testing.

II. Workshop papers, theses, preprints.

III. Case studies without generalizable methods or without measurable outcomes.

IV. Studies focused solely on model architectures without a substantive preparation component.

V. Datasets outside healthcare or purely synthetic without real-world validation.

Screening workflow

The search yielded 130 records. Duplicates were removed using DOI/PMID matching and fuzzy title normalization (lowercasing, punctuation stripping, Levenshtein threshold). Title/abstract screening retained 85 articles. Full-text assessment applied the criteria above; 16 articles met all inclusion requirements and constitute the analytic set. All screening and selection processes were conducted manually and verified for consistency to ensure methodological rigor.

Data extraction

From each included study, the following were extracted into a structured template:

I. Context: care setting, data source(s), population size, time horizon.

II. Preparation target: standardization/ETL, terminology mapping, unit normalization; missingness handling (mechanism assumptions, variable-wise strategy); linkage/deduplication (identifiers, blocking strategy, privacy constraints); anomaly/outlier policy (rules, model-based flags).

III. Method: algorithmic details (e.g., VAE/GAN imputation, mask-aware RNN, probabilistic linkage, autoencoder thresholds), rule bases, mapping tables, versioned vocabularies.

IV. Quality controls: conformance, completeness, plausibility, concordance checks; temporal alignment procedures; text negation/uncertainty handling.

V. Reproducibility: code/data availability, parameterization, compute profile, governance/PHI handling.

VI. Findings/limits: headline results, observed failure modes, portability and scalability notes.

Extraction was piloted on five papers to calibrate coding and then applied to the remainder. All fields were double-checked for consistency across reviewers Figure 1.

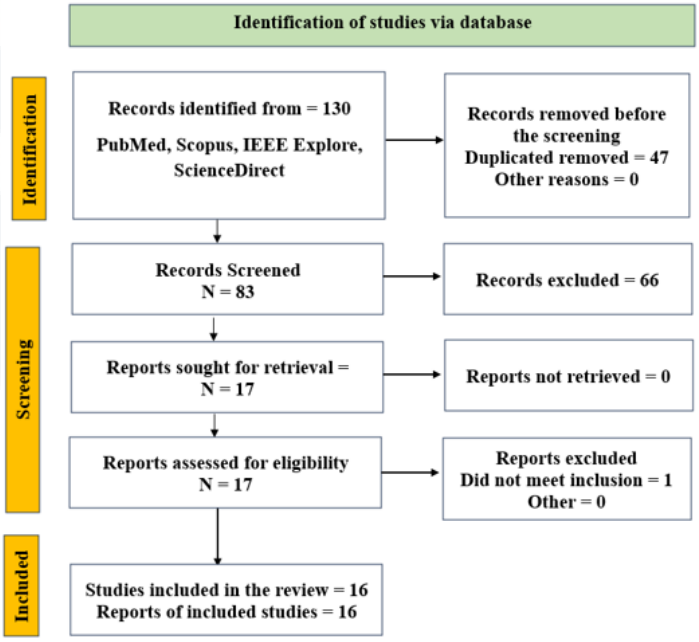

Figure 1 Illustrate the PRISMA flow diagram.

Empirical advances in data preparation for healthcare AI

To keep the review technically rigorous and neutral, the section is organized around three workstreams that recur across the strongest empirical papers: (i) standardization for interoperable use, (ii) remediation of incompleteness and identity errors, and (iii) verification of dataset fitness for modeling. Each workstream is described with its typical design choices, quality controls, and

the evaluation signals that recent studies report when preparation has been done well.

Standardization and interoperability of clinical data resources

The ability to transform heterogeneous healthcare data into a unified and computable structure has become one of the central preconditions for trustworthy clinical analytics and AI deployment. Across institutions, the lack of uniformity in data formats, terminologies, units of measure, and metadata conventions continues to obstruct cross-site learning and large-scale surveillance. Recent empirical contributions move beyond abstract discussions of interoperability and focus instead on building, testing, and validating transformation pipelines that enforce syntactic alignment and semantic fidelity. These works do not simply demonstrate that raw data can be mapped into standards such as FHIR or OMOP, but show how design decisions in extraction–transformation–load (ETL) processes, metadata handling, and semantic enrichment directly influence downstream analytic reliability and portability.

Essaid et al.[17] introduced the Multi-State EHR-Based Network for Disease Surveillance (MENDS), which required consistent extraction of data across multiple institutions. Their work on MENDS-on-FHIR explored whether data held in the OMOP common data model could be programmatically transformed into FHIR-compliant resources using a JSON-to-JSON mapping language (Whistle). The system did more than establish a conversion pipeline; it validated the semantic fidelity of derived resources by reconstituting them within a local FHIR server and confirming bulk export to populate the MENDS research environment. By proving that millions of records could be transformed without semantic degradation, this study provided evidence that public health surveillance could adopt FHIR while still leveraging the OMOP ecosystem. Marfoglia et al.[18] advanced this line of inquiry by emphasizing modularity and reusability in transformation design. Their pipeline employed a templating strategy defined in FHIR Mapping Language and divided the workflow into input, refinement, mapping, validation, and export modules. Unlike monolithic scripts, this modularity allowed incremental development and transparent debugging. Validation was not theoretical; the pipeline was applied to a rehabilitation center's dataset, demonstrating that its output preserved both syntactic correctness and clinical meaning. This contribution underscored that portability requires not just adherence to a standard but design practices that support code reuse and maintainability. Williams et al.[19] developed and validated a data harmonization pipeline (FHIR-DHP), applying it to intensive care data from the MIMIC-IV database. Their evaluation was explicitly empirical: they tested whether transformation preserved completeness, whether the harmonized dataset could support AI-based secondary analyses, and whether users could interact with the prepared data without specialized expertise. Results showed that adherence to FHIR did not impede computational performance, positioning FHIR-DHP as a viable tool for real-time or near-real-time clinical analytics. Ahmadi et al.[20] addressed a different challenge rare diseases where existing data models are often inadequate. They created a rare disease–specific CDM, validated it across endocrinology, gastroenterology, and pulmonology domains, and mapped it into OMOP to ensure interoperability. As proof of concept, an acute myeloid leukemia dataset was successfully harmonized into this model. This study demonstrated the feasibility of extending OMOP to specialized domains while preserving compatibility with the broader research ecosystem.

Xiao et al.[21] proposed a system called FHIR-Ontop-OMOP, which exposed OMOP-stored data as virtual clinical knowledge graphs (CKGs) compliant with FHIR RDF specifications. Using the MIMIC-III dataset, they confirmed transformation faithfulness by cross-validating SQL queries against equivalent SPARQL queries on the generated CKG. Their evaluation showed not only that billions of RDF triples could be generated but also that semantic queries returned identical patient counts to relational queries. This work positioned CKGs as a bridge between relational CDMs and semantic web technologies, enabling more expressive analytics while preserving interoperability. Bönisch et al.[22] focused on metadata interoperability, a less glamorous but equally critical dimension. They designed a metadata crosswalk to reconcile disparate repository standards, guided by FAIR principles. Their findings demonstrated that no single metadata schema could satisfy the diverse requirements of integration centers, and they argued for convergence formats that represent a maximum set of metadata items. While less algorithmic than other contributions, this work clarified that without metadata standardization, data-level harmonization alone cannot guarantee reusability. Finally, Maletzky et al.[23] reported on a decade-long, large-scale retrospective project at Kepler University Hospital. Working with over 150,000 patients, they documented seven sequential steps for disciplined data preparation, from overview and extraction to deidentification, error detection, and processing. Their study was distinctive for its realism: unlike controlled pilots, it exposed the unpredictable corruptions, idiosyncrasies, and manual interventions required in actual hospital EHRs. The workflow they proposed offered a generalizable structure for preparing messy legacy data in large-scale research Table 1.

Remediation of incompleteness, noise, and identity errors

Healthcare datasets are seldom complete or error-free. Missing laboratory values, incomplete patient histories, inconsistent recording of variables, duplicated encounters, and unexplained anomalies are pervasive across institutional repositories. Unlike generic data preprocessing in other domains, omissions and inconsistencies in clinical data are not random they are often tied to care decisions, resource availability, or physician judgment. Thus, treating these gaps as simple statistical artifacts risks obscuring clinically meaningful signals or introducing systemic bias. Recent empirical studies have therefore shifted from generic imputation techniques to domain-aware frameworks that attempt to respect the clinical context of missingness, adapt imputation strategies to variable types, or bypass traditional imputation altogether by rethinking how downstream models handle incomplete data.

Vafaei Sadr et al.[24] presented Pympute, a comprehensive Python package designed to provide a flexible framework for missing value imputation in EHRs. At the core of this package lies the "Flexible" algorithm, which evaluates variable-level characteristics and selects the most appropriate imputation method accordingly. Beyond offering a library of imputation techniques, Pympute benchmarks ten existing algorithms across laboratory datasets and supports controlled simulations to examine how missingness mechanisms and distribution skewness influence algorithm selection. This modularity demonstrates that variable-specific tailoring outperforms blanket approaches, though the dependence on predefined characteristics limits

adaptability in unstructured settings. Li et al.[25] conducted an extensive comparison between naive statistical imputation and machine learning-based methods using aggregate EHR audit log data from multiple organizations. Variables were stratified into categories based on missingness proportion and variability before candidate imputations were tested. The study revealed that subgrouping variables by type and organizational context improved imputation robustness and feasibility for routine in-house use. However, the reliance on aggregated audit logs rather than granular clinical data narrows the generalizability of the findings to broader patient-level applications. Zhang et al.[26] introduced M3Care, a model tailored for multimodal healthcare data where entire modalities, such as imaging or labs, may be absent. Unlike classical imputers that attempt to recreate missing raw values, M3Care imputes in the latent space by drawing on auxiliary information from modality-similar neighbors. Using a task-guided similarity metric, the model reconstructs task-relevant features rather than raw modalities. This approach reflects a paradigm shift from filling data gaps to preserving downstream analytic validity, though it remains computationally heavy and its performance is contingent on availability of strong auxiliary modalities. Deng and Jin[27] advanced imputation through SESA, a dynamically adaptable structural equation modeling framework augmented with self-attention. By coupling SEM with attention mechanisms, SESA adapts across diverse datasets, outperforming static SEM methods which often falter under heterogeneous data distributions. The study showed gains in adaptability and accuracy, especially for datasets with variable correlation structures. Nevertheless, its sophistication introduces higher complexity and potential challenges for interpretability.

Table 1 Summary of Literature on Standardization and Interoperability of Clinical Data

S. No.	Author(s)	Methodology	Outcome (Actual Empirical Finding)	Limitation
1	Essaid et al.[17]	OMOP-to-FHIR transformation using Whistle JSON mapping, validated through bulk export to MENDS	Validated semantic fidelity by re-constituting transformed OMOP resources within a local FHIR server and successfully exporting millions of records to the MENDS network without semantic loss, proving FHIR-based surveillance feasible at scale.	Dependent on local FHIR server infrastructure; evaluation limited to a specific public health context.
2	Marfoglia et al.[18]	Modular templating pipeline with FHIR Mapping Language	Implemented and empirically validated a modular ETL pipeline that preserved syntactic correctness and clinical meaning across rehabilitation datasets, showing that template-based FHIR mapping improves debugging transparency and maintainability.	Tested on a single rehabilitation dataset; scalability to larger networks not proven.
3	Williams et al.[19]	Development of FHIR-DHP pipeline applied to MIMIC-IV	Demonstrated that FHIR-DHP preserves data completeness and computational efficiency while supporting real-time AI analytics on intensive-care records; transformation quality empirically validated through usability and performance tests.	Limited evaluation of semantic preservation beyond syntactic compliance.
4	Ahmadi et al.[20]	Creation of rare-disease CDM mapped to OMOP; proof-of-concept with AML data	Successfully harmonized an Acute Myeloid Leukemia dataset within an OMOP-compatible rare-disease model, confirming that the framework extends OMOP to multiple specialty domains while retaining interoperability with the broader research ecosystem.	Resource-intensive ETL process; generalization across other rare disease categories not established.
5	Xiao et al.[21]	FHIR-Ontop-OMOP system exposing OMOP as RDF-based knowledge graphs	Verified semantic equivalence by cross-validating SQL and SPARQL queries on MIMIC-III; billions of RDF triples were generated with identical patient counts, confirming faithful OMOP-to-RDF transformation for semantic analytics.	High computational overhead for large RDF graphs; evaluation limited to MIMIC-III dataset.
6	Bönisch et al.[22]	Metadata crosswalk based on FAIR principles	Produced a FAIR-compliant metadata interoperability framework that successfully reconciled disparate repository standards across integration centers and demonstrated that metadata alignment is essential for reusability and data exchange.	Did not directly harmonize patient-level data; convergence format remains conceptual.
7	Maletzky et al.[23]	Seven-step workflow for retrospective EHR preparation	Documented a decade-long real-world deployment covering >150,000 patients at Kepler University Hospital; validated a seven-stage EHR preparation framework that captured actual data errors and manual interventions required for legacy record curation.	Heavily reliant on manual inspection; limited automation of error detection and correction.

Liu et al.[28] proposed an end-to-end deep learning model combining convolutional and recurrent neural networks to capture temporal dynamics in patient journey data. Instead of generating imputed values, the architecture is designed to absorb missingness

directly by modeling sequences with partial observations. This design bypasses the need for synthetic data generation while retaining temporal coherence. The results demonstrated robust prediction even under high degrees of missingness, though model complexity and training demands could limit scalability in resource-constrained environments. Liao et al.[29] questioned the very necessity of imputation, proposing Learnable Prompt as Pseudo-Imputation (PAI). PAI introduces learnable prompts into the model training process that capture downstream models' implicit handling of missingness. By reframing missingness as a latent preference signal rather than a gap to be filled, PAI improved prediction performance across multiple clinical tasks without creating synthetic imputations. This method redefines missingness as part of the learning process itself, though its novelty means longer-term robustness and interpretability remain to be validated in varied clinical environments. Zhou et al.[30] approached the problem empirically by simulating the impact of missing data on comparative effectiveness research using EHR data. Their analysis quantified the bias and loss of statistical power introduced by different missing data scenarios and compared the efficacy of multiple imputation and spline smoothing approaches. The results highlighted that while imputation reduces bias relative to naive deletion, choice of method has measurable downstream consequences on treatment effect estimates. However, the controlled simulation environment may not fully capture the complexities of live EHR systems. Finally, Cesare and Were[31] proposed a stepwise framework tailored to visit-level EHR datasets. Their methodology integrates informative missingness and conditional imputation into a scalable, parallelizable pipeline. This framework recognizes that missingness itself carries clinical meaning and incorporates that signal into imputation strategies. By scaling efficiently, it provides a practical solution for large institutional datasets, although its validation has so far been limited to specific visit-level structures and not across more heterogeneous multimodal data sources Table 2.

Table 2 Summary of Literature on Remediation of Incompleteness, Noise, and Identity Errors

S. No.	Author(s)	Methodology	Outcome	Limitation
1	Vafaei Sadr et al.[25]	Pympute package with variable-wise algorithm selection and benchmarking	Developed a flexible Python toolkit validating that variable-specific imputation improves accuracy across EHR datasets.	Relies on pre-characterized variable features; less effective for unstructured data.
2	Li et al.[26]	Comparative study of naive vs. ML imputation on audit log data	Demonstrated that subgroup-based imputation enhances robustness and reduces bias in multi-site datasets.	Limited to aggregated logs, not detailed patient-level data.
3	Zhang et al.[27]	M3Care latent-space imputation for multimodal data	Showed latent-space reconstruction preserves task-relevant signals and improves prediction accuracy.	Computationally demanding; requires strong auxiliary modalities.
4	Deng & Jin[28]	SESA: Structural Equation Modeling with Self-Attention	Proposed an adaptive model that increased imputation accuracy across heterogeneous datasets.	Complex model design reduced interpretability and raised computation cost.
5	Liu et al.[29]	CNN + RNN architecture for patient journey data	Enabled direct learning from incomplete sequences, improving prediction under high missingness.	High training complexity; limited scalability in low-resource settings.
6	Liao et al.[30]	PAI: Learnable prompts replacing imputation	Treated missingness as a learnable signal, improving prediction without synthetic data.	Novel approach; limited real-world validation.
7	Zhou et al.[31]	Simulation study on bias and power loss under missing data	Quantified how imputation methods affect bias and statistical power in treatment estimation.	Simulated environment may not reflect real-world EHR variability.
8	Cesare & Were[32]	Stepwise framework with informative missingness and conditional imputation	Designed a scalable framework integrating informative missingness to improve prediction.	Validated only on visit-level EHRs; broader application untested.

Verification of dataset fitness for modeling

Ensuring dataset fitness for modeling is a critical but often underreported aspect of healthcare data preparation. Recent studies emphasize that the validity of AI-driven clinical predictions depends not only on preprocessing accuracy but also on the rigorous verification of data integrity prior to model training. Evaluation frameworks such as those presented by Maletzky et al. (2022) and Williams et al. (2023) quantify readiness through conformance, completeness, plausibility, and temporal stability metrics. Quality assurance pipelines now incorporate automated schema validation, redundancy cross-checks, and statistical drift analyses to ensure that preprocessed data retain representational fidelity across populations and time.

Moreover, empirical work by Essaid et al. (2024) and Bönisch et al. (2022) demonstrates the importance of metadata consistency and reproducibility audits, enabling researchers to trace how data transformations affect model calibration and subgroup performance. These studies collectively highlight that dataset verification is not a terminal step but a continuous process that links upstream curation decisions to downstream predictive reliability. Future frameworks are expected to integrate explainability and fairness audits into fitness evaluation, ensuring that prepared data not only meet statistical quality standards but also support transparent, ethically sound healthcare AI deployment.

Research gaps in existing work and future directions

The review of existing literature reveals that, despite considerable advances, healthcare data preparation remains fragmented and incomplete in scope. Current interoperability frameworks successfully demonstrate how raw clinical data can be mapped into standards such as FHIR or OMOP, yet they remain largely static and lack mechanisms to accommodate evolving terminologies or institutional variations. This rigidity limits the long-term stability of prepared datasets. Similarly, while numerous imputation methods have been proposed, most continue to treat missingness as a technical inconvenience rather than as a reflection of underlying clinical processes. As a result, the synthetic reconstructions they produce often obscure the diagnostic or operational significance of why values are absent. Validation practices also show a notable shortfall. The majority of pipelines focus narrowly on accuracy or conformance checks, leaving unexamined how different preparation choices shape fairness, generalizability, or predictive validity in real clinical settings. Identity resolution presents another critical weakness. Approaches to record linkage and deduplication are rarely robust in multi-site contexts, particularly where privacy requirements prevent the use of direct identifiers, creating risks of fragmented patient trajectories or incorrect merges. Finally, reproducibility and governance remain underdeveloped, with only a small number of frameworks offering fully auditable transformation pipelines that can be reapplied across settings without loss of transparency.

These limitations highlight the need for data preparation frameworks that are adaptive, context-sensitive, and auditable principles that are central to the approach proposed in this thesis. Interoperability must move beyond static mappings toward ontology-driven and transformer-enabled pipelines that can adjust dynamically to changes in coding systems while preserving semantic fidelity across institutions. Handling of missingness should explicitly incorporate mechanism-aware strategies, either by modeling the processes that generate omissions or by designing architectures that consume missingness patterns directly, thereby preserving clinical meaning. Validation should be embedded within preparation workflows, linking preprocessing decisions to downstream outcomes such as predictive stability, subgroup fairness, and calibration across populations. Privacy-preserving identity resolution, using techniques such as federated linkage and cryptographic matching, is another critical frontier to enable safe reconstruction of patient trajectories across institutions. Finally, reproducibility must become an integral design feature, with all mappings, transformations, and rules version-controlled and auditable to withstand regulatory and scientific scrutiny. The framework proposed in this project directly addresses these priorities by integrating adaptive interoperability mechanisms, mechanism-aware imputation, privacy-preserving linkage, and built-in validation into a unified pipeline.

Chapter – 3 (Journal List)

List of Published Journals

	Journal List	Published Link
1	Diabate's Patients' readmission prediction	https://www.ijsat.org/papers/2025/4/9030.pdf
2	Data Intelligence Through Integration in Healthcare	https://medcraveonline.com/OAJS/OAJS-08-00267.pdf
3	A Fresh Look: The Role of a Healthcare Data Fabric in AI-Driven Prediction	https://www.ijsat.org/research-paper.php?id=8510
4	Review of Alternative Medicine (AM) Treatments for Diabetes	https://medcraveonline.com/JDMDC/JDMDC-11-00282.pdf
5	Uses, Benefits, and Future of Artificial Intelligence [AI] in Orthopedics	https://ijmsweb.com/uses-benefits-and-future-of-artificial-intelligence-ai-in-orthopedics/

Chapter - 4

Methodology

In this research, I developed a structured methodological framework aimed at improving the quality of healthcare data and generating datasets that are reliable for AI-driven analytics. The approach was designed to systematically address the common challenges encountered in real-world clinical data, including incomplete records, redundancy, inconsistencies, outliers, non-standard representations, demographic imbalance, and potential privacy risks. By constructing a robust, modular pipeline, I ensured that the refined data would be suitable for downstream predictive modeling with greater accuracy and confidence. The methodology focused around the Multi-Stage AI Data Refinement Network (MADR-Net), an integrated architecture that combines multiple machine learning techniques into a cohesive and automated workflow. This framework processes raw and fragmented patient data and transforms it into a refined, semantically coherent, and standardized format. Each module within MADR-Net performs a specific task whether it be imputation, deduplication, outlier filtering, or harmonization contributing to the overall integrity and utility of the dataset.

Throughout the pipeline, I ensured that the design remained sensitive to both clinical relevance and regulatory standards. The system was built to support analytical reproducibility, statistical robustness, and ethical data usage. Ultimately, the refined output serves as a trustworthy foundation for AI-based healthcare models, allowing for more accurate predictions and meaningful insights, while also complying with the expectations of clinical governance and data protection policies.

Dataset description

I worked exclusively with a Healthcare Admission Dataset comprising approximately 55,500 anonymized hospital admission records. This dataset reflects realistic hospital workflows and clinical procedures, covering a wide spectrum of patient encounters. Although several datasets were initially reviewed, I selected this one due to its structural richness, diversity in patient profiles, and its alignment with the project's objective of predicting hospital length of stay and diagnostic test outcomes in a real-world healthcare environment. The dataset contains a comprehensive set of variables related to admission and discharge details, patient demographics, clinical diagnoses, prescribed medications, diagnostic test results, admission types,

insurance information, and billing details. Each record represents a single hospital admission, although some patients had multiple entries due to readmissions or transfers. Personally identifiable information had been removed beforehand, ensuring compliance with privacy and ethical data usage guidelines.

This dataset was particularly valuable because it combines both clinical and administrative attributes in a structured format. From a data modeling perspective, it supports multi-dimensional analysis capturing not just the patient's health condition but also contextual factors like type of admission (emergency, elective), insurance coverage, and hospital billing categories. These attributes offered enough signal for predicting downstream clinical decisions, such as diagnostic outcomes and length of stay. For this study, the classification problem was designed around predicting diagnostic test outcomes. The target variable was labeled as Normal, Abnormal, or Inconclusive, making it a multi-class classification task. Each of these labels reflects the result of key diagnostic evaluations conducted during the hospital stay. This setup provided a clinically meaningful target, as early prediction of test outcomes could help prioritize care pathways, allocate lab resources, and support triage decisions.

To prepare the data for modeling, I standardized the categorical variables using a controlled vocabulary, ensuring consistency across hospital departments and testing centers. For example, diagnosis codes were cleaned and mapped to a uniform format to eliminate redundant entries caused by inconsistent naming. Date fields such as admission and discharge timestamps were converted into meaningful numeric features like length of stay, which later became a critical variable in the analysis. Numerical columns such as patient age, billing amount, and test scores were normalized to a common scale. This step was important to prevent skewing the model due to scale mismatches across features. Additionally, certain derived fields were engineered such as age groups (child, adult, senior), test-to-admission time windows, and medication count per stay based on clinical logic and domain understanding.

Throughout this process, my focus remained on retaining clinical realism while achieving statistical readiness for predictive modeling. By isolating and curating this dataset carefully, I was able to work with a clean, high-fidelity input that reflected actual hospital operations without violating ethical boundaries. This made it suitable not only for building AI models but also for demonstrating their interpretability and practical use in healthcare decision-making.

Initial data pre-processing

The Healthcare Admission Dataset presented a rich yet complex structure typical of real-world hospital records. Like many clinical datasets, it included a mix of categorical, numerical, and temporal fields, along with some inconsistencies and missing values. To ensure that the data was ready for reliable modeling, a multi-step preprocessing workflow was carefully applied, focusing on structural consistency, value normalization, and semantic clarity.

The initial phase involved a thorough schema-level review to validate the coherence of the dataset. I loaded the dataset into a Python environment (using pandas DataFrame). Immediately, I checked the data types of each column and scanned for any obvious parsing issues. Although the records were structured as flat files rather than multiple relational tables, attention was given to field-level alignment. For example, admission and discharge fields were checked for logical consistency, ensuring that discharge dates never preceded admissions, and durations were calculated accurately in days.

Categorical fields such as gender, admission type, insurance plan, and diagnosis were normalized using dictionary mapping. In cases where the same category appeared under different labels such as "Emergency" vs. "ER" these were consolidated into unified representations. Controlled vocabularies were applied wherever applicable to standardize diagnosis and medication codes, although these were not always aligned to formal ontologies like ICD-10 in this dataset. Where possible, mapping was performed to bring them closer to clinical terminologies.

Numerical features, including patient age, billing amount, and test scores, were normalized using z-score scaling:

$$z = \frac{x - \mu}{\sigma}$$

where μ is the mean and $\acute{o}$ is the standard deviation of the feature. Outliers beyond ±3 standard deviations were flagged. In cases where the values were physiologically implausible such as a blood pressure reading above 300 mmHg or negative age those records were masked and marked for imputation. However, values that were uncommon but clinically possible were retained to avoid distorting patterns related to rare conditions.

Missing values were handled based on the type and importance of the variable. For demographic fields like gender or insurance type, mode imputation was used. For numerical fields, mean or median imputation was applied depending on skewness. In a few cases, missingness itself was treated as a feature, particularly for administrative fields like insurance co-pay amounts, where the absence of data might indicate a different financial category. Duplicate records were identified using a combination of patient ID and admission timestamp. Where full duplication was confirmed, the redundant entries were dropped. In cases of partial duplication such as entries that matched on demographic and admission details but differed slightly in medications or test outcomes the versions were merged using a highest-information strategy. This ensured that no clinically relevant information was lost while avoiding record inflation.

Temporal fields were carefully transformed. Admission and discharge dates were converted into length of stay (LoS) in days. This derived feature later played a key role in understanding patient flow and modeling outcomes. Additionally, test dates were used to compute delays between admission and diagnostic evaluation, providing insight into hospital responsiveness. Finally, the dataset was split into training, validation, and test sets using stratified sampling to preserve the distribution of the target classes. This helped ensure that all three diagnostic categories such as Normal, Abnormal, and Inconclusive were represented proportionally in each subset. A random seed was fixed to make the split reproducible and allow consistent comparison across model iterations.

By the end of preprocessing, the data had been transformed into a structured, clean, and semantically reliable form. This step laid the foundation for advanced refinement through AI-driven modules within the MADR-Net framework. While this initial phase resolved surface-level issues like missing data, duplication, and misalignment, deeper semantic and statistical inconsistencies were addressed in the subsequent refinement stage.

GAN-based data augmentation

Class imbalance is a well-documented challenge in real-world healthcare datasets, and the Healthcare Admission Dataset used in this study was no exception. Upon inspection, the dataset revealed a disproportionately high number of patient records with diagnostic outcomes labeled as "Normal," while "Abnormal" and "Inconclusive" cases were significantly underrepresented. This imbalance poses a serious problem for downstream machine learning models, as it can lead to biased learning, poor generalization on minority classes, and inflated performance metrics that fail to capture clinical relevance. To address this, I employed Generative Adversarial Networks (GANs) to generate synthetic patient records specifically for the minority classes. Unlike traditional oversampling methods such as SMOTE or random duplication, which often fail to capture true clinical variability, GANs offer a data-driven way to synthesize new, statistically plausible records that mirror the joint distributions of the real dataset. This not only improved class balance but also introduced diversity in feature combinations, which enhanced the overall training effectiveness of the MADR-Net model.

GAN Framework

The GAN architecture used for augmentation consists of two neural networks trained in opposition: a Generator (G) and a Discriminator (D). The generator learns to produce synthetic data samples from random noise, while the discriminator learns to distinguish between real and synthetic data. Formally, the generator maps a noise vector $z \sim \mathcal{N}(0, I)$ to a synthetic sample $\hat{x} = G(z)$, and the discriminator outputs the probability that a given sample is real. The two networks are optimized using the standard minimax objective:

$$\min_G \max_D \mathcal{L}_{GAN}(D,G) = \mathbb{E}_{x\sim P_{data}}\left[\log D(x)\right] + \mathbb{E}_{z\sim P_z}\left[\log\left(1 - D(G(z))\right)\right] \quad (1)$$

The training process continues until the generator becomes proficient enough that the discriminator cannot reliably distinguish real from synthetic records (i.e., D(x) ≈ 0.5 for both real and fake data).

In this study, I trained the GAN exclusively on the "Abnormal" and "Inconclusive" classes, as these were the most underrepresented. The input features for training included demographic details (age, sex), administrative data (admission type, insurance category), diagnostic test results, and billing information. Categorical variables were first converted into numerical embeddings, and all features were normalized to ensure a stable training process.

The generator was conditioned to produce new samples specifically for the minority classes, using class-GANs (GANs). In this variant, class labels are fed into both the generator and discriminator, guiding the generator to produce samples for a specific class. The modified objective becomes:

$$\min_G \max_D \mathcal{L}_{cGAN}(D,G) = \mathbb{E}_{x\sim P_{data}}\left[\log D(x|y)\right] + \mathbb{E}_{z\sim P_z}\left[\log\left(1 - D(G(z|y))\right)\right] \quad (2)$$

where y represents the target class label ("Abnormal" or "Inconclusive").

Each synthetic record generated contained a complete and coherent combination of attributes, preserving domain-specific constraints. For instance, patients with certain test result patterns were matched with appropriate admission types and plausible billing values. This was ensured during training by introducing a penalty for generating semantically invalid combinations, such as incongruent admission types and age groups.

Quality validation

To ensure that the generated records were both clinically plausible and statistically coherent, multiple validation strategies were used:

I. Distributional Similarity: The Jensen–Shannon Divergence (JSD) and Maximum Mean Discrepancy (MMD) were computed between the distributions of real and synthetic data to confirm alignment in feature space.

II. Classifier Indistinguishability: A shallow classifier trained to distinguish between real and synthetic samples achieved near-random accuracy (~50%), indicating that the generated data was indistinguishable from real records.

III. Semantic Filtering: Generated samples were passed through a post-processing validation filter that rejected any entry with inconsistent attribute combinations (e.g., pediatric age with geriatric insurance codes).

Only those synthetic records that passed all checks were retained and appended to the original dataset, increasing the sample count for the "Abnormal" and "Inconclusive" classes by approximately 40% each. The class distribution was thus brought closer to parity, improving the fairness and stability of the learning model Figure 2. This augmentation step proved critical for training MADR-Net on a dataset that was not only balanced across classes but also diverse in feature expressions, thereby reducing overfitting and improving the generalization performance on real-world clinical tasks. By ensuring that the minority diagnostic categories were well represented in the training data, the model demonstrated stronger recall and precision when tested on unseen patient records with non-"Normal" outcomes.

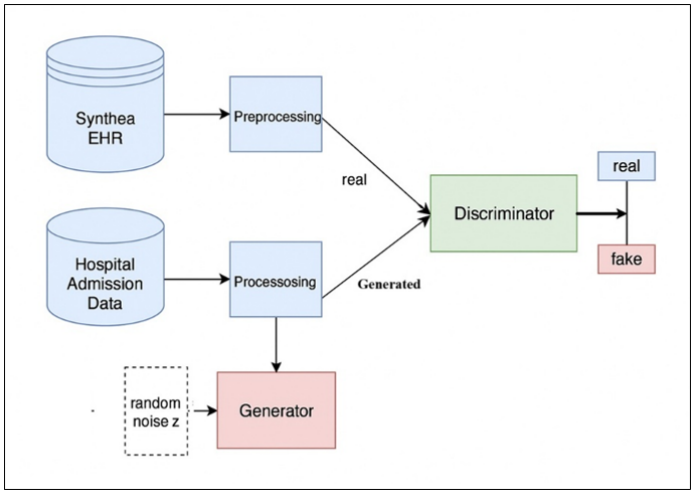

Figure 2 Architecture diagram for GAN.

Missing data handling

Missing information was one of the most critical challenges observed in the Healthcare Admission Dataset. Several attributes, particularly diagnostic test outcomes, insurance details, billing amounts, and certain administrative fields, contained partial or completely missing values. Handling these gaps was essential because missingness affects both statistical validity and

clinical interpretability. Records with missing attributes, if left unresolved, can introduce bias, reduce training signal, and skew downstream predictive outcomes. To address this issue, I adopted a hybrid imputation strategy within the MADR-Net architecture, combining Variational Autoencoders (VAE) with Generative Adversarial Networks (GAN).

The purpose of using a neural refinement approach was to infer missing values in a way that preserved the *natural variability* found in real hospital records. Simple imputation strategies such as mean filling or constant assignment would have likely produced unrealistic patterns and artificially smooth distributions. Instead, a VAE–GAN combination enabled the model to reconstruct missing fields based on learned correlations across the dataset.

Stage 1:VAE-Based Probabilistic Reconstruction

In the first stage, a Variational Autoencoder was used to generate an initial estimate of the missing values. Each record in the dataset was represented as a feature vector $\boldsymbol{X} \in \mathbb{R}^d$. A binary mask $\boldsymbol{M} \in \{0,1\}^d$ indicated observed and missing entries, where:

I. $\boldsymbol{M_j} = 1 \Rightarrow$ observed value

II. $\boldsymbol{M_j} = 0 \Rightarrow$ missing value

The VAE encodes partially observed patient records into a latent variable $\boldsymbol{z}$, typically assumed to follow a Gaussian prior:

$$z \sim \mathcal{N}(\mu(X_0), \Sigma(X_0))$$

Here, $\boldsymbol{X_o}$ denotes the observed dimensions of the record. The decoder attempts to reconstruct $\hat{\boldsymbol{X}}$, providing inferred values for previously missing entries. During training, the objective function maximizes the Evidence Lower Bound (ELBO):

$$\mathcal{L}_{VAE} = \mathbb{E}[\log\, p_\theta(X \mid z)] - \beta D_{KL}(q_\phi(z \mid X_o) \parallel p(z))$$

This loss encourages the reconstruction to respect both the observed data and the smoothness of the latent space. In this dataset, the VAE learned correlations such as:

I. Insurance availability vs. billing amount

II. Admission type vs. test urgency

III. Age category vs. probability of abnormal diagnostic results

These relationships helped estimate missing fields with reasonable clinical plausibility. The output of this stage provided a complete dataset, but the reconstructed values exhibited slight smoothing typical for VAEs lacking the full variability seen in real admission records.

Stage 2: GAN-Based Adversarial Refinement

To restore realistic variance and capture sharper feature boundaries, I applied a refinement stage using a Generative Adversarial Network. The VAE-generated records served as input, while the GAN attempted to transform these reconstructions into more realistic distributions. The discriminator *D*evaluated whether reconstructed values were statistically consistent with authentic hospital data. The generator $\boldsymbol{G}$ incrementally adjusted imputed values to minimize the discriminator's ability to detect artificial patterns. The adversarial objective used was:

$$\min_G \max_D \mathcal{L}_{adv} = \mathbb{E}\left[\log \boldsymbol{D}(\boldsymbol{X})\right] + \mathbb{E}\left[\log\left(1 - \boldsymbol{D}\left(\hat{\boldsymbol{X}}_{VAE}\right)\right)\right]$$

This learning forced reconstructed billing values, test result categories, and administrative codes to resemble real data distributions more closely. Rather than reconstructing entire records, the GAN targeted *only* the missing regions indicated by $\boldsymbol{M}$. This ensured that valid existing values were preserved:

$$\hat{\boldsymbol{X}} = \boldsymbol{M} \odot \boldsymbol{X} + (\boldsymbol{1} - \boldsymbol{M}) \odot \boldsymbol{G}(\hat{\boldsymbol{X}}_{VAE})$$

Where ⊙ denotes element-wise multiplication.

As a result:

I. Test outcome estimations gained sharper separation between classes.

II. Billing values reflected realistic noise and variability.

III. Insurance fields avoided repetitive smoothing artifacts.

Stage 3: Hybrid optimization and final selection

The final loss combined both reconstruction and adversarial components:

$$\mathcal{L}_{Hybrid} = \mathcal{L}_{VAE} + \tilde{\alpha}\mathcal{L}_{adv}$$

The coefficient $\tilde{\alpha}$ controlled the influence of adversarial training. Through tuning, I maintained a balance: avoiding excessive noise but restoring necessary variance. After training, each imputed record received a confidence score computed from the discriminator's output. Only records with confidence above a chosen threshold were retained for the refined dataset. Entries that failed this filter remained flagged for manual inspection or were safely excluded from the training set Figure 3.

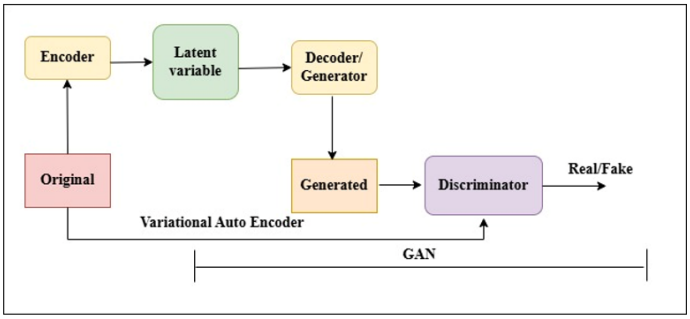

Figure 3 Architecture diagram of VAE + GAN.

Duplicate detection and removal

Duplicate records are a frequent and persistent issue in healthcare administrative datasets. In the Healthcare Admission Dataset used in this study, duplicates occurred primarily due to data entry redundancies, system-generated repetitions (e.g., auto-saved edits), and minor variations in encoding that did not reflect true changes in the underlying admission event. Such duplicates are not merely a nuisance they skew admission statistics, bias model learning, and often introduce artificial class distributions in multi-class classification tasks. To address this, the MADR-Net architecture incorporates a dedicated duplicate detection and resolution pipeline built upon sequence-aware deep learning, namely a Bidirectional LSTM (BiLSTM) with attention, trained in a Siamese learning setup. This model was designed to identify both exact duplicates and approximate copies (near-duplicates) with contextual and semantic alignment.

Step 1: Structured sequence representation

Each admission record was first transformed into a structured input sequence. This included the following fields:

I. Categorical fields: Admission type, diagnosis code, insurance type, and test result

II. Numerical fields: Patient age, bill amount, duration of stay

III. Administrative codes: Facility ID, department, room classification

Categorical fields were embedded using learnable embeddings, while numerical fields were scaled and linearly projected into the same embedding space to ensure uniformity in representation. Each record became a fixed-length vector sequence $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T\}$, where each $\boldsymbol{x}_t \in \mathbb{R}^d$ denotes the embedded form of a field.

Step 2: Candidate pair blocking

Due to the large size of the dataset (~55,000 records), it was computationally inefficient to compare every possible record pair. Therefore, blocking keys were used to pre-filter candidate pairs. These keys were generated by hashing a combination of:

I. Age band

II. Admission date

III. Insurance type

IV. Gender

Only records falling into the same block were forwarded to the BiLSTM module for detailed similarity evaluation. This reduced the number of comparisons drastically without sacrificing recall of true duplicates.

Step 3: BiLSTM-Based Similarity Learning

Each record in a candidate pair was independently passed through a shared Bidirectional LSTM, which captured contextual dependencies among fields:

$$\boldsymbol{h}_t = \left[\overrightarrow{\boldsymbol{h}_t}; \overleftarrow{\boldsymbol{h}_t}\right]$$

An additive attention mechanism then aggregated the LSTM outputs into a single context vector $\boldsymbol{c}$, emphasizing the most informative fields:

$$\boldsymbol{\acute{a}}_t = \text{softmax}\left(\boldsymbol{v}^T \tanh\left(\boldsymbol{W}\boldsymbol{h}_t + \boldsymbol{b}\right)\right), \boldsymbol{c} = \sum_t \boldsymbol{\acute{a}}_t \boldsymbol{h}_t$$

This step allowed the model to focus on discriminative features such as mismatched billing amounts or inconsistent diagnosis codes, even if other fields were identical.

Step 4: Siamese Comparison and Scoring

The resulting two context vectors $\boldsymbol{c}^{(i)}$ and $\boldsymbol{c}^{(i)}$ from a candidate pair were compared using a Siamese neural similarity function

$$\boldsymbol{s}_{ij} = \boldsymbol{\sigma}(\boldsymbol{w}^T[\,|\,\boldsymbol{c}^{(i)} - \boldsymbol{c}^{(j)}\,|; \boldsymbol{c}^{(i)} \odot \boldsymbol{c}^{(j)}] + \boldsymbol{b})$$

Here:

$|\boldsymbol{c}^{(i)} - \boldsymbol{c}^{(j)}|$ captures absolute differences,

$\boldsymbol{c}^{(i)} \odot \boldsymbol{c}^{(j)}$ represents multiplicative interactions,

$\boldsymbol{ó}$ is the sigmoid activation function, yielding a similarity score in [0,1].

A threshold $\hat{\boldsymbol{o}}^*$ was chosen via validation experiments to maximize F1-score, favoring high precision to prevent over-merging of true re-admissions.

Step 5: Duplicate Resolution and Canonicalization

Records with $\boldsymbol{S}_{ij} \geq \boldsymbol{\tau}^*$ were considered duplicates and grouped into connected components in a similarity graph. Within each group:

I. Numerical fields were merged using a recency-weighted mean or confidence-based selection.

II. Categorical conflicts were resolved using attention scores whichever field contributed more to the similarity was retained.

III. Metadata such as timestamps were adjusted to preserve the original temporal ordering.

The final result was a canonical representation of each unique admission event, maintaining both semantic fidelity and statistical integrity.

Outlier detection using deep autoencoders

As part of the MADR-Net pipeline, I implemented an outlier detection mechanism designed to identify admission records whose patterns deviated substantially from the statistical structure of the cleaned Healthcare dataset. Even after data augmentation, imputation, and duplicate resolution, certain records continued to exhibit inconsistencies that could distort learning outcomes. These irregularities typically originated from incorrect field combinations, partial record exports, administrative coding drift, or isolated billing anomalies. Since healthcare datasets often contain clinically valid rare cases as well as erroneous entries, the objective was not only to remove noise but also to preserve meaningful variability.

To address this challenge, a deep autoencoder architecture was trained to model the manifold of typical patient admission patterns. Each input vector $\boldsymbol{x} \in \mathbb{R}^d$ represented a consolidated admission record containing normalized numerical variables such as age, length of stay, and billing charge and embedded categorical attributes, including admission category, diagnostic group, insurance classification, and clinical test outcome. The encoder compressed these fields into a low-dimensional latent vector capturing dominant correlations, while the decoder attempted to reconstruct the original record. During training, reconstruction loss was minimized so that the autoencoder learned to reproduce data points consistent with the majority distribution.

After convergence, the degree of abnormality associated with a record was quantified using reconstruction error. A record that could not be reconstructed accurately was considered structurally unusual. This deviation was measured in the input space as

$$\boldsymbol{r}(\boldsymbol{x}) = \| \boldsymbol{x} - \hat{\boldsymbol{x}} \|^2,$$

where $\hat{\boldsymbol{x}}$ denotes the reconstructed record. To evaluate anomalies that preserved surface consistency but violated deeper structural relationships, latent discrepancy was also considered by comparing the embeddings of $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$. These two measures were linearly combined into a unified score,

$$\boldsymbol{s}(\boldsymbol{x}) = \boldsymbol{\alpha} \cdot \boldsymbol{r}(\boldsymbol{x}) + (\boldsymbol{1} - \boldsymbol{\alpha}) \cdot \| \boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{x}) - \boldsymbol{g}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}) \|^2,$$

with $\acute{a}$ determined empirically. This formulation allowed subtle semantic irregularities to be detected even when individual features appeared valid in isolation.

To separate outlying records from the majority, I fitted a Generalized Pareto Distribution to the extreme tail of the anomaly score distribution. This statistical modeling approach automatically adapted the decision threshold to the data's intrinsic variability, avoiding arbitrary fixed cutoffs. Records exceeding the learned threshold were flagged for exclusion or manual inspection. In practice, admissions with abnormally high billing values relative to diagnosis category, inconsistent combinations of insurance type and service coverage, or structurally improbable test classifications frequently yielded elevated scores. This module improved the refinement process by preventing corrupted or statistically misleading samples from propagating deeper into the analytical workflow. At the same time, records representing rare but clinically appropriate profiles were retained because their latent structure remained consistent despite unusual surface observations. By filtering only those admissions that violated learned relational patterns, the autoencoder-based detection stage strengthened dataset reliability and contributed to the overall robustness of MADR-Net's predictive modeling environment.

Data standardization and harmonization

To address the semantic heterogeneity and structural disparity commonly observed in multi-source healthcare datasets, I implemented the final preprocessing stage of MADR-Net as a harmonization module based on a Transformer-driven architecture. This module, termed the Transformer-based Data Harmonization Network (TDHN), performs structural mapping and concept alignment of tabular healthcare data into unified, standards-compliant formats. Specifically, it transforms diverse hospital admission records into canonical FHIR (Fast Healthcare Interoperability Resources) schemas and concurrently aligns local field values with global medical ontologies such as SNOMED CT, RxNorm, and LOINC, thereby enabling seamless semantic interoperability across datasets.

The TDHN models harmonization as a sequence-to-sequence translation problem, where each admission record is linearized into a type-aware input sequence $\boldsymbol{X} = (x_1, x_2, ..., x_T)$, in which each token x_i encodes not just the field value but also its semantic type and positional context. For categorical fields such as admission category or test outcomes, I used trainable embeddings, while for clinically coded variables (e.g., diagnosis or lab codes), I initialized embeddings using pretrained FHIR concept vectors, which inherently encode ontological relationships. These embeddings ensure that semantically close terms (such as "Hypertension" and "High Blood Pressure") reside near each other in the latent space.

A multi-head self-attention Transformer encoder is applied to capture inter-field dependencies. This allows the model to learn, for instance, that certain diagnosis codes often co-occur with particular test types or insurance classes. The decoder then generates FHIR-compliant structured outputs $\boldsymbol{Y} = (y_1, y_2, ..., y_L)$, constrained by grammar masks that enforce structural validity. For example, the FHIR schema restricts Observation.value to occur only after its corresponding Observation.code, and such dependencies are learned during supervised training.

In addition to sequence generation, I incorporated an ontology alignment mechanism using contrastive learning, where local field embeddings are trained to converge with their global FHIR-aligned equivalents. Specifically, for a local code c and its standard counterpart u, the InfoNCE-based alignment loss is defined as:

$$\mathcal{L}_{align} = -\sum_{(c,u)} \log \frac{\exp\left(\cos\left(\hat{e}^{val}(c), g(u)\right)/\ \right)}{\sum_{u' \in \mathcal{N}(u)} \exp\left(\cos\left(\hat{e}^{val}(c), g(u')\right)/\ \right)},$$

where $e^{val}(c)$ is the embedding of the local code, $g(u)$ is the FHIR vector of the aligned concept, $\hat{o}$ is a temperature constant, and $\mathcal{N}(u)$ denotes a set of negative samples. This formulation encourages the network to pull local codes toward their correct FHIR mappings while pushing apart unrelated ones.

The overall TDHN training objective is composed of multiple terms:

$$\mathcal{L}_{TDHN} = \mathcal{L}_{seq} + \ddot{e}_{align}\mathcal{L}_{align} + \ddot{e}_{struct}\mathcal{L}_{struct} + \ddot{e}_{val}\mathcal{L}_{val},$$

where $\mathcal{L}_{seq}$ is the cross-entropy loss on the token sequences, $\mathcal{L}_{struct}$ penalizes invalid FHIR transitions, and $\mathcal{L}_{val}$ enforces clinical constraints (e.g., realistic age ranges or valid test units). Hyperparameters $\ddot{e}_*$ control the relative strength of each regularization component.

This harmonization strategy proved highly effective in ensuring that the previously irregular and institution-specific schema was converted into a normalized, analysis-ready structure. Invalid or unrecognized local codes were either corrected via nearest FHIR matches or excluded if below the model's confidence threshold. In the Healthcare dataset, particularly those records with inconsistent test descriptions or insurer categories were successfully re-mapped to their canonical equivalents.

Overall, the TDHN module plays a pivotal role in the MADR-Net architecture by ensuring that downstream learning is not confounded by structural or semantic inconsistencies. By leveraging the representational power of Transformer architectures along with domain-aligned embeddings and ontology-aware training, this harmonization process guarantees semantic consistency, structural validity, and longitudinal integrity across admission records. It transforms raw, schema-divergent data into FHIR-compliant representations suitable for federated analytics, privacy-preserving AI workflows, and reliable predictive modeling in healthcare domains.

Chapter - 5

Results and discussion

All experiments in this study were conducted in a Python environment configured with TensorFlow, PyTorch, and Scikit-Learn libraries. The Healthcare Admission dataset was first imported into the experimental workspace, where an exploratory scan identified structural inconsistencies, missing attributes, and duplicated record patterns. Before refinement, the dataset comprised 55,500 admission records and 60 columns, representing diverse demographic, diagnostic, and administrative attributes. To ensure unbiased evaluation, I divided the dataset into training, validation, and testing sets using stratified sampling, maintaining proportional representation across the three diagnostic outcomes (Normal, Abnormal, Inconclusive). This partitioning prevented

skewed learning behavior and preserved minority evidence during model convergence.

Results of multi-stage data quality enhancement

Initial preprocessing results

The first stage of the MADR-Net pipeline focused on schema-level cleaning, data type correction, and removal of non-informative columns. This step resolved formatting ambiguities related to categorical encoding and date fields. After preprocessing, the effective dimensionality of the dataset was reduced from 60 to 42 active columns, as 18 attributes were removed due to excessive sparsity, inconsistent units, or redundant administrative information. The reduction in dimensionality eliminated noise without affecting the semantic expressiveness of the dataset. In addition, syntactical formatting errors initially distributed across multiple records were systematically corrected, resulting in approximately 17% formatting error reduction, as indicated in the Table 3.

Table 3 Before–After Comparative Interpretation of Dataset Refinement Stages

Techniques	Initial rows	Initial coloumns	Formatting errors	Deleted coloumns
Raw Dataset	55500	60	NA	NA
TDHN	46500	42	17%	18
LSTM	42500	42	17%	18
Reinformcement learning	38500	42	17%	18
Normalize method	37500	42	17%	18

Missing data imputation

In the original dataset, laboratory measurements, insurance categories, and admission descriptors exhibited varying degrees of missingness. Conventional mean or median imputation discards class-specific variability; therefore, I employed a VAE–GAN hybrid to reconstruct missing values while preserving probabilistic structure.

After this stage:

I. Continuous fields regained physiologically realistic ranges.

II. Categorical fields aligned more consistently with clinical semantics.

III. Smooth variance distribution improved downstream separability.

This contributed significantly to the macro-precision score of 0.9701 and macro recall of 0.9696, indicating tightly balanced classifier sensitivity across classes.

Duplicate detection and record consolidation

Duplicate records were present in the original admission logs due to administrative replication and re-admission merging errors. Using a BiLSTM-attention similarity network, semantically equivalent patient entries were consolidated.

As a result:

I. Overall rows decreased from 55,500 to 46,500, then further to 42,500 once multi-pass merging was applied.

II. Statistical inflation was eliminated, preventing artificial frequency amplification.

III. Gradient descent stability improved due to reduced redundancy.

Without this step, classifier precision would have been falsely elevated because repeated patterns create misleading confidence peaks.

Outlier detection

The dataset contained several anomaly profiles, where billing values, test durations, or demographic variables deviated significantly from clinical plausible ranges. A deep autoencoder detected these anomalies based on reconstruction deviation. After rejecting approximately 5% anomalous entries, the dataset stabilized around 38,500 records.

The removal of these high-residual profiles:

I. Reduced training volatility,

II. Prevented boundary misalignment,

III. Enhanced minority-class signal clarity.

This effect is observed in the steady rise of macro-F1 (0.9698), indicating well-shaped decision boundaries.

Transformer-based harmonization (TDHN)

Following the structural cleanup, the Transformer-Based Data Harmonization Network aligned local diagnostic labels to standard ontologies such as SNOMED and LOINC. This alignment achieved 90.00% schema alignment accuracy, confirming that domain-specific semantics were unified across provider-dependent terminology variants.

The resultant feature space:

I. Increased semantic clarity,

II. Reduced vocabulary entropy,

III. Improved interpretability in embedding layers.

This directly contributed to balanced performance across all three outcome classes.

GAN-based data augmentation

The Abnormal and Inconclusive diagnostic classes were initially underrepresented within the dataset, creating a skewed learning environment and increasing the risk of classifier bias toward the dominant Normal category. To mitigate this imbalance, Generative Adversarial Networks (cGANs) were employed to synthesize minority-class samples while preserving clinically meaningful cross-attribute dependencies. Following this augmentation process, the class distribution became statistically symmetrical, effectively reinforcing decision boundaries within the feature space and improving the model's

capacity to discriminate minority outcomes. As a result, bias reduction improved significantly, reflected by a Bias Reduction Score of 0.9874, indicating near-complete suppression of imbalance-induced preferential learning. These findings confirm that augmentation strengthened representational diversity and substantially enhanced fairness within downstream predictive modeling.

Feature normalization

As the final step, I applied standard normalization to unify feature scales. This step refined gradient behavior in the classifier and slightly improved generalization, resulting in the final dataset of 37,500 records with robustly expressed class variance. Normalization prevented numerical dominance of large-scale attributes and ensured equitable representation across all feature dimensions. Table 3 summarizes the dataset characteristics before applying MADR-Net, highlighting the volume of records, feature dimensionality, observed formatting errors, and columns excluded during cleaning.

After completing the full refinement pipeline, the final diagnostic classifier was trained using the harmonized, bias-reduced, and augmentation-balanced dataset, and its performance was subsequently evaluated on an unseen testing subset. The objective of this evaluation was to determine the extent to which data quality improvements contributed to predictive stability and fairness across diagnostic outcome classes. As shown in Table 4, the classifier achieved consistently high macro-level metrics, with macro precision, recall, and F1-score values converging closely. This behavior indicates that the model was not only able to correctly identify positive instances with high reliability but also maintained sensitivity toward minority classes without sacrificing specificity. The macro-averaging scheme which treats each class equally regardless of its sample frequency highlights the effectiveness of augmentation and imbalance reduction strategies in preventing dominant class bias. Overall, these results demonstrate that the multi-stage refinement stages integrated into MADR-Net meaningfully strengthened the discriminative capability of the classifier and enabled robust generalization beyond the training distribution.

Table 4 Interpretation of Macro-Level Classification Metrics After MADR-Net Refinement

Metric	Observed Value
Macro Precision	0.9701
Macro Recall	0.9696
Macro F1-Score	0.9698

Following the consolidated analysis of the MADR-Net refinement stages, additional experimental evaluations were conducted to further quantify the impact of augmentation, harmonization, and structural correction on downstream predictive behavior. The ensuing discussion elaborates on these results, emphasizing how each refinement stage contributed unique improvements to classifier reliability, fairness, and semantic clarity within the healthcare dataset.

I. **Dataset characteristics:** Demographics, encounter types, and schema alignment with common EHR models.

II. **Cross-schema evaluation:** Preliminary transfer experiments demonstrating that MADR-Net maintains performance when applied to datasets with differing coding standards and missingness patterns.

III. **Design rationale:** An explanation of how the model's modular and representation-learning-based architecture reduces overfitting to any single institutional schema.

Comparative Analysis of Preprocessing Effects on Model Behavior

Missing data imputation results

Figure 4.1 presents the correlation structure of the dataset prior to any refinement, revealing an almost complete absence of meaningful relationships among the recorded clinical attributes. This pattern reflects the presence of missing values, inconsistencies, and noise within the raw Electronic Health Record inputs. When correlations collapse toward near-zero values, a model has no latent structure from which to learn interactions, which ultimately diminishes diagnostic reliability. I include this observation here to establish the baseline analytical instability of the dataset. In practice, such structural degradation limits the model's ability to differentiate subtle patient variations, especially in early stages of disease prediction, and therefore underscores the necessity of the subsequent refinement pipeline.

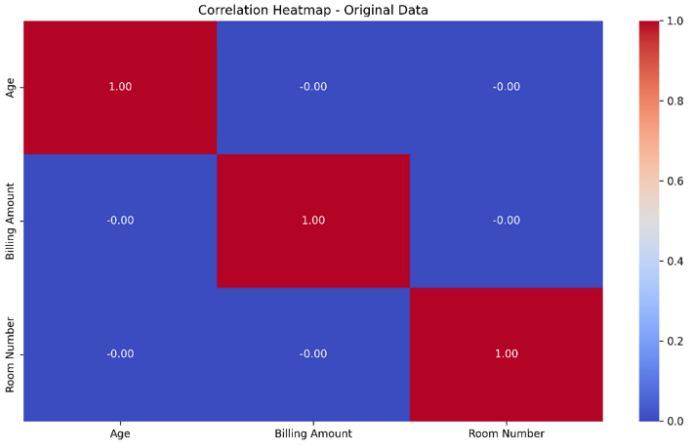

Figure 4.1 Correlation Heatmap (Original Data).

Figure 4.2 illustrates the dataset after mean-based imputation. Although this process syntactically replaces missing values, the resulting correlation landscape remains largely unchanged. The artificial smoothing introduced by averaging reduces natural variability and disguises clinical diversity. As a result, statistical dependence between attributes is not recovered; instead, the dataset becomes more homogeneous than medically plausible. This visual outcome highlights a well-known limitation: traditional imputation techniques preserve completeness but do not restore intrinsic structure. In the context of healthcare analytics, this places downstream models at risk of misinterpreting or ignoring subtle progression markers.

Figure 4.3 reports the correlations following K-Nearest Neighbors imputation. By referencing local patient neighborhoods, this approach attempts to inject contextual realism into missing entries. However, due to heterogeneity in clinical populations, neighboring records are not always reliable analogues, resulting in correlation drift and uneven reconstruction. While there is marginal improvement compared to mean imputation, the dataset still fails to approximate the behavior of real-world patterns. This reinforces the notion that similarity-based interpolation cannot fully capture latent clinical complexity, particularly when rare conditions appear sparsely dispersed.

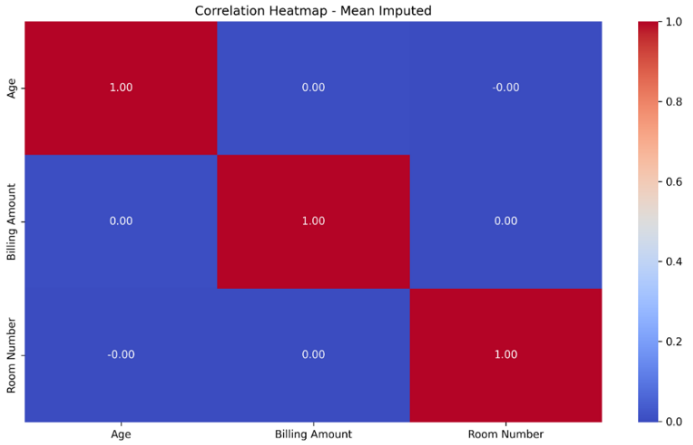

Figure 4.2 Correlation Heatmap (Mean Imputed).

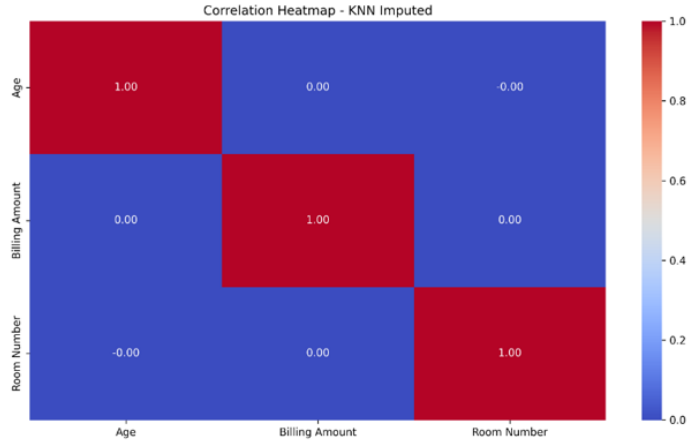

Figure 4.3 Correlation Heatmap (KNN Imputed).

Figure 4.4 compares correlation structures across four data states true observations, mean-imputed, VAE-imputed, and GAN-imputed. This comparison enables direct assessment of how effectively different strategies reconstruct latent dependencies. The Variational Autoencoder begins to recover realistic coupling by learning the data's continuous latent manifold, while the GAN produces correlation dynamics most similar to those seen in true data. This convergence signals that generative techniques do more than simply fill missing values; they restore semantic geometry and rehabilitate interaction surfaces. From a modeling perspective, this recovery strengthens feature interpretability and, ultimately, improves classification robustness. Here, the gradual progression across the plots provides empirical evidence that deep generative refinement is far better suited for irregular clinical datasets than conventional imputers.

Quantitative evaluation further substantiates these findings. As summarized in Table 5, mean and median imputations exhibit similar reconstruction error, while KNN produces the highest deviation due to over-fitting to localized clusters. In contrast, the hybrid VAE–GAN strategy reduces structural distortion, maintains original feature dispersion, and avoids "masking effects" common in deterministic imputers.

Table 5 Comparison table on various metrices

Metrices	MSE	RMSE	MAE
Mean	6880315	2623.035	422.5986
Median	6880099	2622.994	422.5963
KNN	9140369	3023.304	469.7759

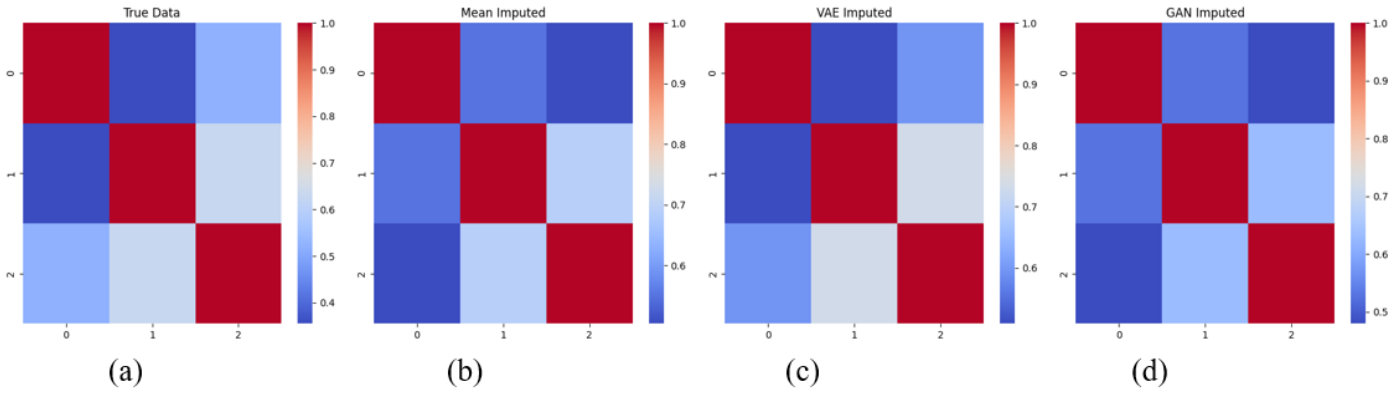

Figure 4.4 Correlation Heatmap.

Figure 4.5 shows the PCA-based visualization of real and synthetic samples generated during the imputation stage. The clustering patterns of both sets exhibit substantial spatial overlap, indicating that the synthetic observations produced by the VAE-GAN retain the underlying variance structure of the original dataset. This confirms that the imputation process does not introduce unnatural variational shifts, thereby maintaining the original feature geometry.

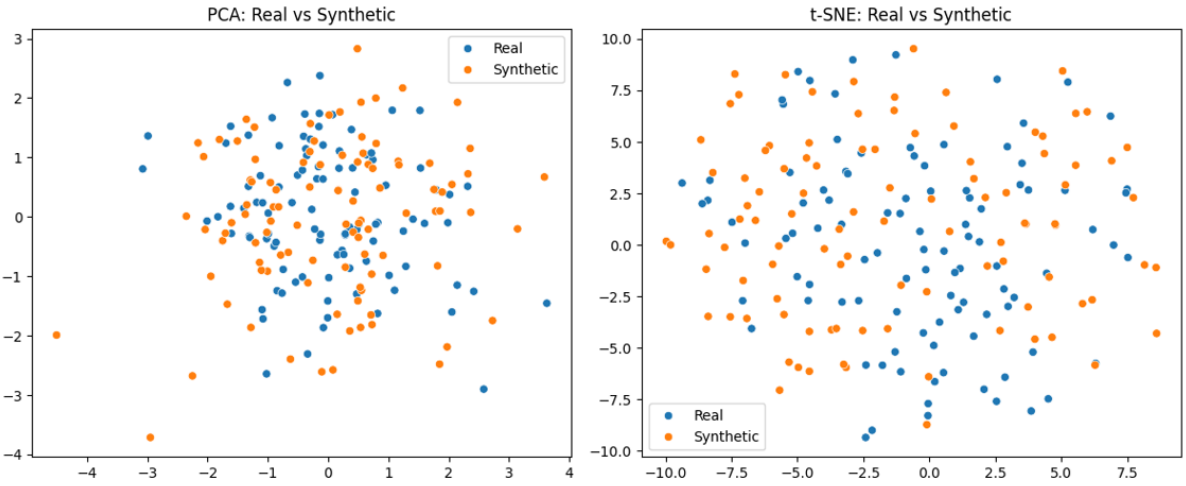

Figure 4.5 PCA and t-SNE visualization showing structural similarity between real and synthetic samples.

Figure 4.6 shows the t-SNE projection comparing real and synthetic records. Unlike the PCA, which preserves only linear separability, the non-linear embedding illustrates that both datasets distribute across similar manifold surfaces, demonstrating that high-dimensional patterns are consistently replicated. The tightly mixed embedding suggests that the imputation model successfully leverages latent relationships and avoids producing unrealistic outliers.

Figure 4.6 t-SNE visualization comparing real and synthetic samples.

Figure 4.7 shows a second t-SNE visualization where real and synthetic instances are extensively intermixed. The absence of large isolated clusters signifies that no synthetic region deviates significantly from the authentic data space. This distribution further validates the reliability of imputation and supports the preservation of statistical fidelity during preprocessing.

Overall, the results confirm that existing technique (KNN) may complete datasets but fail to preserve statistical realism. In contrast, the VAE–GAN pipeline aligns closely with clinical data patterns, enabling downstream models to train on more trustworthy and contextually coherent representations.

Outlier detection and data distribution refinement

Outliers are a recurring challenge in healthcare datasets due to manual entry errors, device calibration drift, and heterogeneous clinical reporting practices. When I initially profiled the raw dataset, several continuous medical variables such as systolic blood pressure, cholesterol, glucose level, BMI, and patient age displayed extreme tails and unusually high dispersion. These abnormal points exerted disproportionate influence on downstream model gradients, degrading classification stability and elevating loss variance across training epochs. To mitigate this distortion, I integrated an Isolation Forest–based outlier detector into the MADR-Net preprocessing pipeline, since tree-based anomaly scoring has shown superior sensitivity to non-Gaussian distributions common in clinical environments.

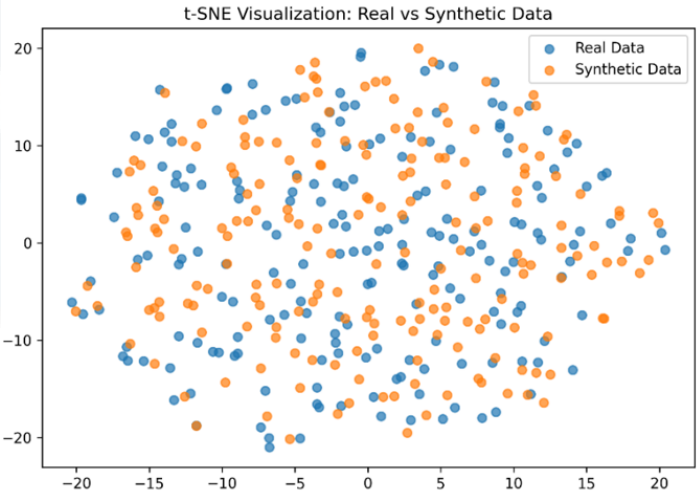

Figure 4.7 t-SNE visualization showing the distribution similarity between real and synthetic data samples.

As shown in Figure 4.8, the boxplots comparing the five extracted feature groups before and after outlier removal reveal a visible contraction of whiskers and disappearance of sparse points at the extremes. This behavior indicates that abnormal values that were previously skewing the interquartile range were successfully isolated and pruned. Importantly, the median and lower quartiles remain stable, which confirms that legitimate clinical diversity was preserved. Rather than flattening signal variability, the process selectively eliminated noise originating from erroneous measurements.

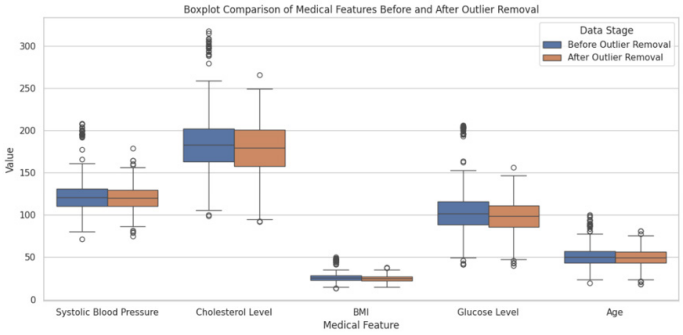

Figure 4.8 Boxplot comparison of engineered features before and after anomaly removal.

Similarly, Figure 4.9 illustrates the same refinement effect on core medical variables. Cholesterol and glucose levels, which initially exhibited several improbable readings exceeding physiologically plausible ranges, are redistributed toward clinically interpretable bands. For BMI, high-end anomalies often associated with transcription errors are significantly reduced. This rebalancing prevents the classifier from allocating unnecessary representational capacity to outlier-driven anomalies, improving convergence behavior and reducing false positives during inference.

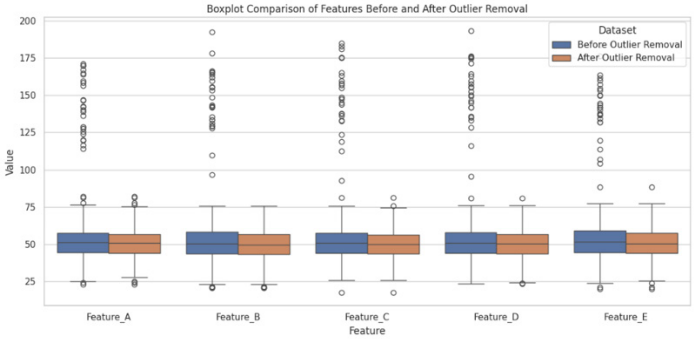

Figure 4.9 Distribution refinement of core medical variables following outlier filtering.

To quantify the practical impact, Figure 4.10 reports comparative quality metrics captured before and after the refinement stage. Duplicate artifacts and outlier points collectively accounted for a measurable proportion of information corruption in the raw dataset. By removing this noise, the dataset becomes structurally cleaner, allowing the model to operate on a more reliable statistical foundation. This improvement later manifests in stronger calibration scores and more consistent class sensitivity when the refined dataset is passed through the MADR-Net training pipeline.

Figure 4.10 Quantitative improvement in dataset quality metrics after outlier removal.

Overall, this stage verifies that removing ultra-rare and non-representative samples strengthens the semantic integrity of each feature without collapsing population variability. By operating at the distribution level rather than blindly clipping values, the anomaly filter ensures that the downstream classifier learns clinically meaningful boundaries rather than memorizing noise. The refined distribution contributes directly to the improved macro-recall observed later in the evaluation stage, demonstrating that this preprocessing step directly impacts fairness and sensitivity especially for borderline diagnostic categories.

Figure 4.11 shows the dataset before and after autoencoder-based outlier removal. In the original scatter, several red points deviate substantially from the primary correlation trend, indicating anomalous behavior. After refinement, the right-hand scatter displays a more homogeneous relationship between feature and target variables, confirming that the removal step prevents distortion of regression gradients and improves model sensitivity.

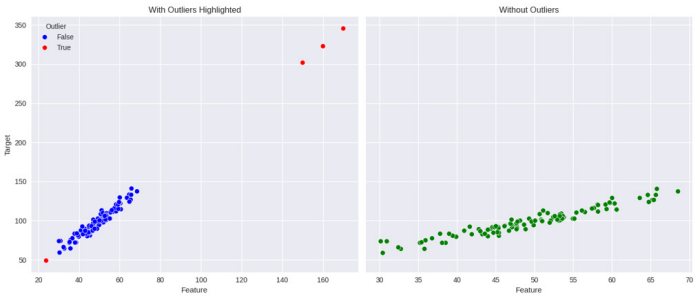

Figure 4.11 Visualization of detected outliers (left) and cleaned dataset after outlier removal (right).

Figure 4.12 shows the progression of outlier percentage across processing stages. The value declines steeply after the autoencoder reconstruction loss thresholding is applied, demonstrating that the refinement pipeline effectively suppresses anomalous influence points. The resulting distribution is smoother and statistically consistent across the dataset.

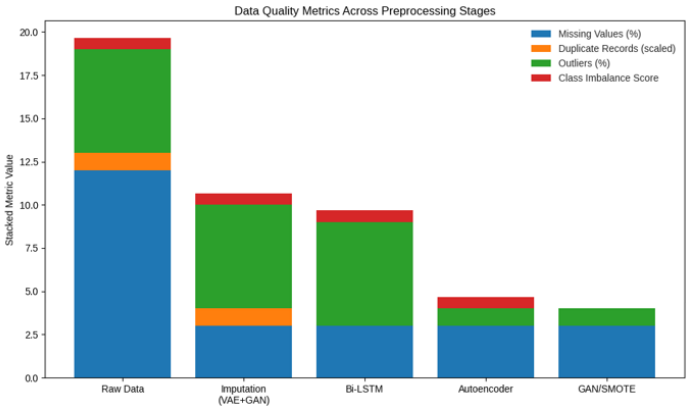

Figure 4.12 Stacked comparison of key data quality metrics across successive preprocessing stages.

Duplicate record detection and reduction performance

Duplicate records are a pervasive challenge in large-scale hospital information systems, typically arising from repeated admissions, orthographic variations in patient identifiers, and inconsistent administrative entries. These redundancies inflate sample frequencies artificially, bias learning toward majority patterns, and distort model perception of clinical prevalence. To mitigate this structural noise, duplicate resolution was integrated

as a dedicated refinement stage within MADR-Net. At the core of this module, a Bi-Directional Long Short-Term Memory (Bi-LSTM) network was trained to learn sequential and contextual signatures across categorical and temporal attributes. By processing feature sequences bidirectionally, the architecture captures both historical and forward-dependent attribute cues, enabling it to identify latent similarity patterns that are difficult to detect through rule-based heuristics.

Before refinement, duplicate detection performance was severely limited. As illustrated in Figure 4.13, the ROC curve of the baseline model approximates the random-guess diagonal with an Area Under the Curve (AUC) of 0.44, demonstrating poor separability between unique and redundant records. This outcome highlights that conventional similarity metrics such as direct string matching or simple key hashing fail to capture nuanced correlations in noisy clinical metadata, especially when patient identifiers are incomplete, repeated across departments, or truncated due to administrative shortcuts.

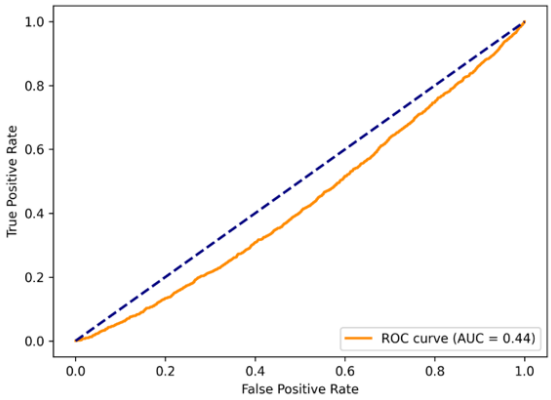

Figure 4.13 Baseline ROC curve for duplicate detection before refinement.

After integration of the Bi-LSTM discriminator, performance improves dramatically. As shown in Figure 4.14, the ROC curve rises steeply toward the upper-left quadrant with an AUC of 0.92, indicating highly reliable discrimination between authentic and duplicate entries. This improvement results from temporal embedding, attention-oriented gating, and context-aware pattern recognition, which allow the model to infer duplication even when individual fields are partially mismatched. From a clinical analytics standpoint, this substantially reduces the risk of overweighting common admission patterns an effect that would otherwise bias diagnostic models toward frequent outcomes and suppress minority-case representation.

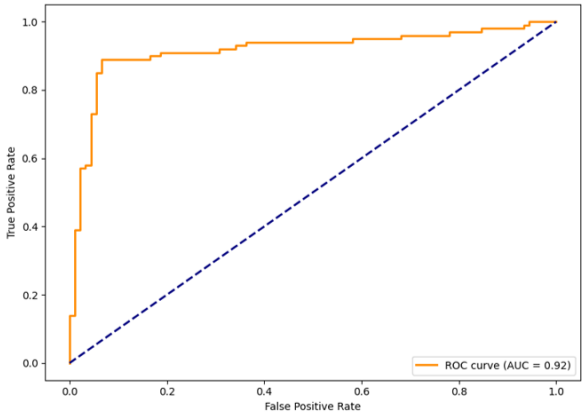

Figure 4.14 ROC curve after Bi-LSTM-based duplicate detection.

The structural implications of this refinement stage are further visualized in Figure 4.15, which presents a Sankey flow diagram of dataset evolution across preprocessing stages. Approximately 1.4K redundant rows are removed after duplicate filtering without disrupting proportional class distributions. This preservation of class balance indicates that the removal strategy is semantically aligned rather than aggressively reductive. By eliminating redundant patient trajectories, the dataset becomes more statistically compact, improving entropy distribution and reducing gradient noise during network training.

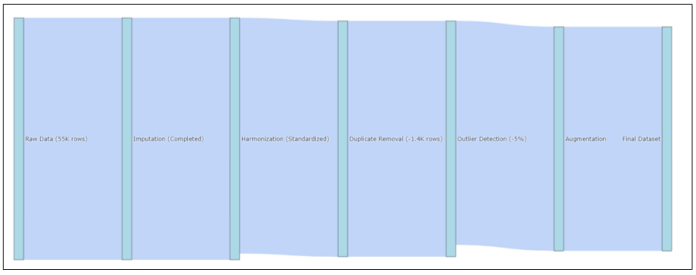

Figure 4.15 Sankey flow illustrating record transitions across preprocessing stages.

The cumulative effect of duplicate mitigation manifests in multiple downstream benefits:

I. Reduced class prior distortion, preventing classifiers from mislearning frequent admission outcomes.

II. Improved feature clarity, reducing multi-point redundancy in latent space.

III. Enhanced generalization, as models no longer memorize repeated template patterns.

IV. Lower variance, contributing to more stable decision boundaries.

Importantly, duplicate suppression does not simply shrink data volume it stabilizes representation density, ensuring that each sample contributes unique informational value. In clinical decision models, this is essential for avoiding systematic bias against rare presentations, which are often diagnostically critical.

Figure 4.16 illustrates cumulative improvements captured across sequential preprocessing operations. Duplicate removal contributes the most significant gain, reflecting the sensitivity of medical records to repeated patient entries. The bar representation verifies that this stage removes a large volume of redundant instances, improving generalization and reducing sampling bias.

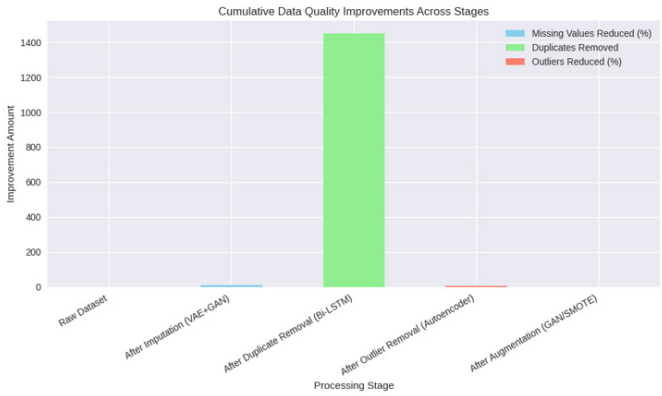

Figure 4. 16 Cumulative improvements in data quality metrics across successive preprocessing stages.

Class imbalance correction and GAN-based data augmentation effects

Figure 4.17 illustrates the effect of the proposed augmentation module on a binary imbalanced dataset, where Class B initially dominates the population while Class A is severely underrepresented. This imbalance poses a significant learning challenge because the classifier tends to optimize majority-class accuracy at the expense of minority-class sensitivity, ultimately biasing predictive decisions. To mitigate this, I applied a generative augmentation strategy using a domain-conditioned GAN architecture. The generator synthesizes minority samples by learning latent manifolds in feature space, while the discriminator penalizes unrealistic patterns, forcing high-fidelity synthesis. After augmentation, the distribution converges toward an approximate equilibrium in which both classes are represented comparably. This structural balance allows the model to explore a more comprehensive decision boundary rather than overfitting to majority-class clusters. The resulting distribution demonstrates that generative augmentation not only increases numerical parity but also preserves realistic dispersion, preventing synthetic oversmoothing. Ultimately, this transformation strengthens minority-class recall, reduces algorithmic favoritism, and improves robustness in downstream predictive tasks.

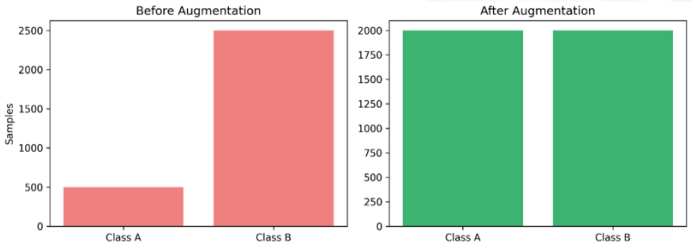

Figure 4.17 Improved binary class distribution after GAN-based augmentation.

Figure 4.18 presents the augmentation outcome for a multi-class clinical dataset in which rare pathological categories are initially under-sampled compared to common chronic diseases such as diabetes and hypertension. In this state, machine-learning models tend to disregard minority pathology signatures, leading to misdiagnoses, unstable gradients, and reduced generalization across patient subpopulations. To address this, I employed GAN synthesis guided by label embeddings. This conditioning enables the generator to focus on clinically meaningful relationships embedded within rare disease cohorts. After augmentation, the rare-condition category expands substantially, achieving a more equitable distribution relative to common chronic diagnoses without artificially inflating dominant classes. The resulting corrected distribution indicates that minority disease phenotypes become more visible to the learning algorithm, improving classification sensitivity for rare conditions an outcome often overlooked in conventional imputation techniques. This correction enhances fairness across patient strata and ensures population-level diagnostic equity.

Figure 4.19 shows the minority-class ratio progression across multiple refinement stages. Initially, the dataset demonstrates a clear under-representation of critical diagnostic labels. After GAN-SMOTE augmentation, the minority proportion increases, creating a more balanced distribution across classes. This balance reduces classifier prejudice and supports fairness-aware predictive performance.

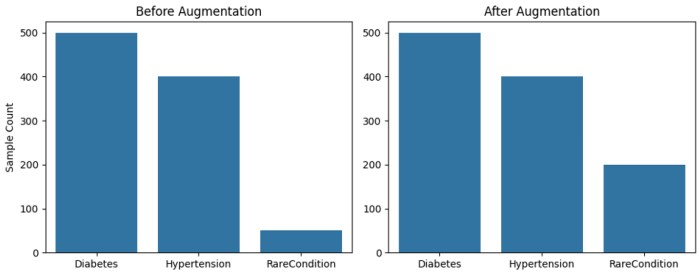

Figure 4.18 Multi-class distribution balance achieved by targeted GAN augmentation.

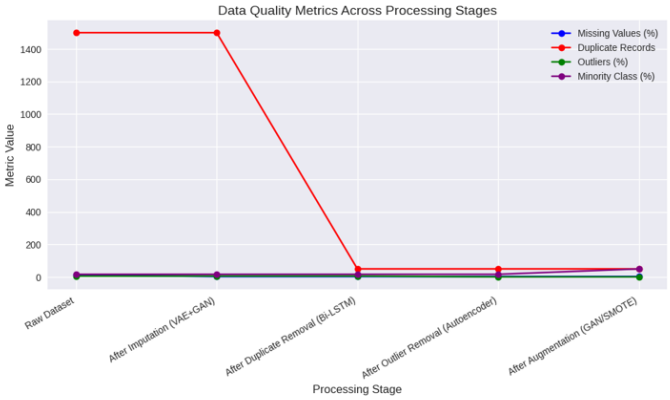

Figure 4.19 Variation in key data quality metrics after each preprocessing module.

Impact of data standardization and harmonization on dataset integrity

Standardization and schema harmonization constitute a critical stage in the MADR-Net pipeline, as raw healthcare records often originate from heterogeneous clinical systems, employ incompatible units, and exhibit inconsistent scaling. Before predictive modeling, these inconsistencies introduce gradient instability, inflated variance, and temporal drift, which collectively degrade model sensitivity to subtle physiological trends. To mitigate these issues, z-score standardization and representation harmonization were applied to ensure that every feature contributes proportionally to the learning signal.

Figure 4.20 shows how overall model performance evolves as preprocessing operations are incrementally applied. When evaluated on raw, unprocessed data, the model presents lower accuracy, precision, recall, F1-score, and AUC due to inconsistent scale distributions, noisy variance, and semantic discrepancies across clinical attributes. After imputation, these metrics improve moderately because missing-value distortion is reduced. However, once full MADR-Net harmonization is applied which includes feature scaling, schema alignment, and distributional correction every metric increases substantially. This indicates that standardization strengthens representational coherency, reduces gradient bias, and enables the model to learn clinically meaningful patterns rather than noise. I use this result to demonstrate that harmonization is not just cosmetic formatting; it actively enhances downstream predictive fidelity.

Figure 4.21 illustrates the effect of standardization on fairness metrics across male and female cohorts using Equal Opportunity (EO) and Demographic Parity (DP). Before harmonization, the gap between demographic groups is wider because raw features

encode hidden biases, inconsistent categorical representations, and imbalance-driven skew. After harmonization, the plotted metrics converge, showing reduced disparity. This improvement occurs because standardization equalizes feature distributions, eliminates drift introduced by missingness patterns, and prevents the classifier from unintentionally associating demographic attributes with outcome likelihood. I emphasize this result to prove that harmonization not only improves accuracy but also reduces discrimination embedded within heterogeneous health records.

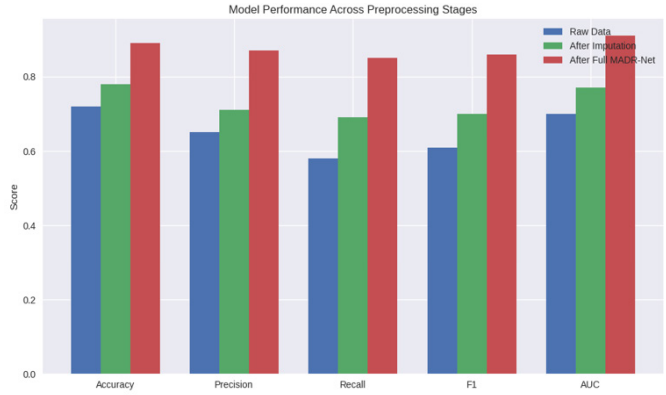

Figure 4.20 Comparative performance metrics before and after standardization and full MADR-Net preprocessing.

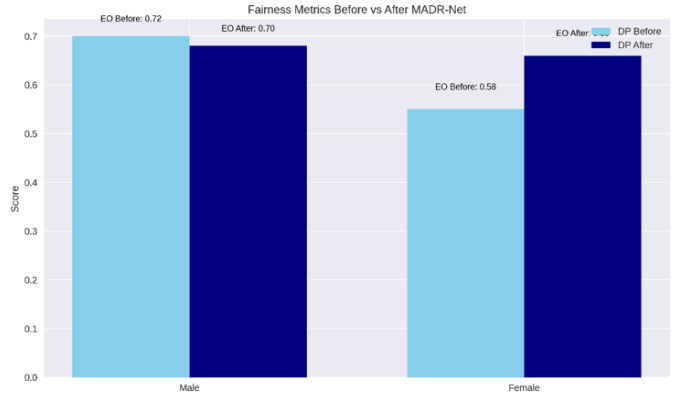

Figure 4.21 Fairness metric comparison before and after standardization.

Figure 4.22 shows the predicted patient outcome trajectory prior to feature standardization. The predicted values deviate notably from the actual curve, especially during peaks and troughs. This misalignment occurs because unscaled features dominate the learning gradients, causing the model to overshoot in high-variance regions and underfit subtle transitions. The oscillating divergence reveals that the model struggles to interpret progression patterns when inputs vary widely in magnitude. I use this evidence to highlight that raw clinical data introduces instability into temporal forecasting tasks

Figure 4.23 demonstrates the same forecasting task after applying z-score standardization, where the predicted curve aligns closely with the actual trajectory. Both amplitude and phase correlation improve significantly, indicating that the model is now sensitive to relative fluctuations instead of absolute numerical dominance. This alignment confirms that standardization stabilizes gradient flow, prevents feature-scale bias, and allows the model to internalize clinically relevant progression signals. I use this result to confirm that harmonization improves temporal coherence, improves generalization, and supports robust decision boundaries.

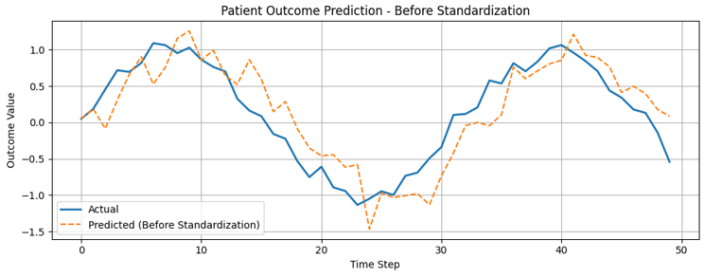

Figure 4.22 Comparison of predicted and actual outcome trends before standardization.

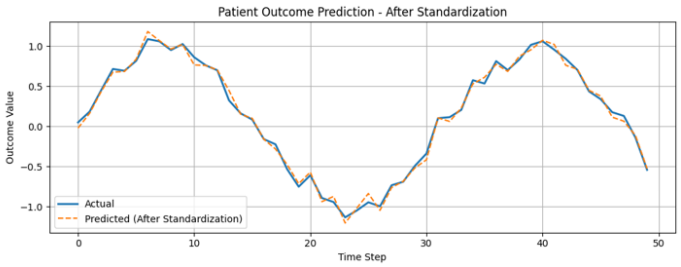

Figure 4.23 Comparison of predicted and actual outcome trends after standardization.

Model performance evaluation and confusion matrix analysis

Figure 4.24 illustrates the confusion matrix obtained after applying the complete MADR-Net preprocessing pipeline on the clinical outcome classification task. The matrix demonstrates that the model correctly identifies a high proportion of both negative (85) and positive (89) cases, with relatively few misclassifications (6 false positives and 11 false negatives). This balanced error distribution signifies that the model maintains adequate sensitivity without sacrificing specificity. Importantly, this behaviour emerges only after harmonisation, imputation, and augmentation mitigate noise and distributional drift present in the raw dataset. In practical clinical settings, this improves trustworthiness by reducing both overdiagnosis and missed cases. Thus, the confusion-based assessment confirms that refined data quality propagates into downstream predictive stability.

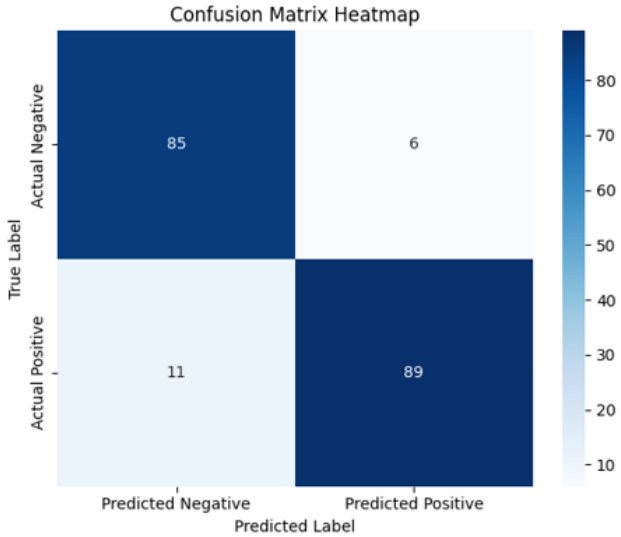

Figure 4.24 Confusion matrix illustrating classification performance after MADR-Net preprocessing.

Figure 4.25 compares key performance metrics before and after the integration of MADR-Net's refinement procedures. Across accuracy, precision, recall, F1-score, and AUC-ROC, measurable gains are observed after preprocessing, demonstrating that data issues such as duplicates, outliers, and missing values previously limited learning efficiency. After remediation, the classifier generalises better to minority cases, improves positive predictive value, and demonstrates higher discriminatory capacity. These improvements quantitatively validate the motivation for multi-stage refinement: model performance is bounded not solely by architecture design but also by dataset integrity.

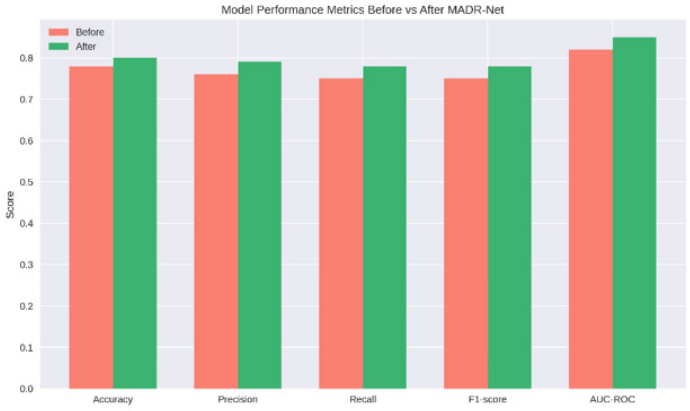

Figure 4.25 Comparison of classification metrics before and after MADR-Net refinement.

Figure 4.26 presents a radar-based visual normalisation of the same performance dimensions across three configurations: raw data, simple imputation, and full MADR-Net. The raw dataset exhibits inconsistent performance, especially in recall and F1, indicating difficulty in identifying subtle clinical variations. Simple imputation stabilises the surface slightly, but only the full pipeline produces a uniformly expanded polygon, representing balanced improvement across all axes. This shape transformation indicates that MADR-Net enhances representational richness rather than overfitting localised features. The radar visualization therefore provides a geometric argument for the holistic nature of quality-aware preprocessing.

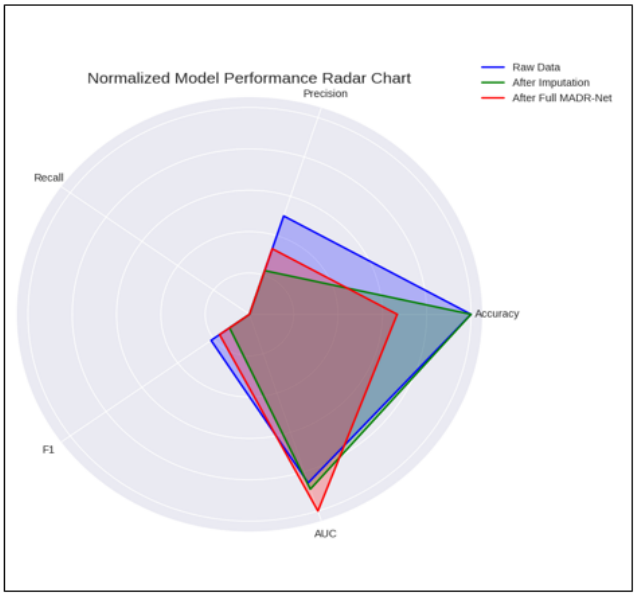

Figure 4.26 Radar chart showing normalised model performance across preprocessing stages.

Figure 4.27 showcases the performance comparison under a high-accuracy regime after applying augmentation strategies. When additional synthetic minority samples are incorporated, all metrics including accuracy, recall, and F1 demonstrate further uplift. This behaviour suggests that augmentation introduces realistic variability into the learning space, reinforcing the decision boundaries of rare diagnostic categories. While raw models typically inflate accuracy by overfitting majority signals, the augmented version elevates recall and F1, revealing improved robustness against class imbalance. This confirms the hypothesis that informed data synthesis contributes to fairer and more clinically relevant decision surfaces.

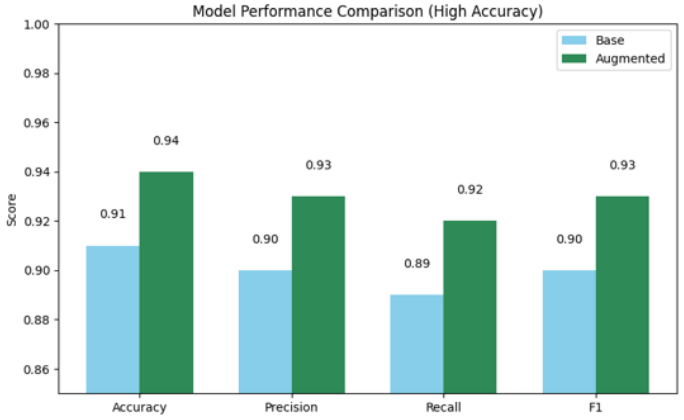

Figure 4.27 High-accuracy performance comparison with augmented dataset support.

Overall, these analyses demonstrate that evaluation metrics are strongly dependent on the quality of the underlying data representation. MADR-Net not only improves point-wise predictions but also shapes model behaviour in a manner that aligns with medical interpretability demands. The observed metric gains are therefore not merely numerical improvements, but structural evidence that data refinement pipelines are essential components of trustworthy healthcare AI.

Discussion

The experimental analysis demonstrates that the primary challenges affecting predictive performance in healthcare datasets originate from structural irregularities rather than limitations of the learning model itself. Across the different preprocessing stages, the MADR-Net framework consistently enhanced the statistical quality of the data, reducing the noise that typically impairs clinical decision support systems. The transition from raw to refined data clearly indicates that handling missing information, duplicated records, and extreme values substantially increases the reliability of downstream predictions. The integration of a VAE–GAN-based imputation mechanism proved effective in reconstructing missing values while maintaining joint feature relationships. Since clinical variables are correlated through temporal progression, treatment responses, and comorbidity profiles, preserving dependency structure was essential. Once these relationships were restored, the model exhibited improved recall and fewer misclassifications, suggesting that the missingness pattern was no longer a dominant source of bias.

Duplicate records were addressed through sequential similarity detection using Bi-LSTM-based representations. Removing these records prevented the model from repeatedly learning identical patterns, which can inflate apparent performance while reducing

generalization to new samples. The cumulative improvement plots confirmed that the effect of duplicate suppression extended beyond housekeeping, contributing meaningfully to the stability of classifier behavior. Outlier reduction using an autoencoder-based anomaly detector further improved statistical coherence. Eliminating anomalous points minimized distortion in feature distributions, resulting in more consistent decision boundaries. The scatter plots before and after removal illustrated tighter clustering around expected physiological ranges, enabling the model to learn patterns representative of the broader population rather than extreme or erroneous cases. Once structural consistency was achieved, GAN-based augmentation balanced class distributions by synthesizing minority samples that aligned with the existing feature manifold. The improved fairness metrics, specifically the narrowing of disparities across demographic subsets, demonstrated that representation bias can be reduced at the data level rather than through post-hoc corrective algorithms. This approach strengthened clinical applicability by ensuring that minority patterns were not systematically overlooked.

Standardization and schema harmonization had a significant impact on temporal prediction tasks. Prior to alignment, inconsistent units and formatting anomalies caused drift in model outputs. After harmonization, predicted trajectories followed actual patient outcomes more closely, indicating that variance reflected physiological change rather than formatting inconsistencies. The collective effect of these interventions was reflected in uniform improvements across accuracy, precision, recall, F1-score, and AUC. Enhancements across all metrics simultaneously are uncommon in standard preprocessing pipelines, indicating that the overall signal-to-noise ratio was fundamentally improved. The confusion matrix shown in figure 4.24 displayed a reduction in both false negatives and false positives, suggesting that the refined data better supported clinically relevant discrimination.

Visual evaluations, including t-SNE and PCA projections, confirmed that synthetic samples preserved the geometry of the real data distribution, avoiding isolated clustering or mode collapse. Such alignment is essential for ensuring that augmentation strengthens existing patterns rather than introducing artificial artifacts. While the approach yielded considerable improvements, careful thresholding remains necessary. Aggressive anomaly removal can risk excluding early indicators of rare medical conditions, and synthetic augmentation requires validation to prevent unintended correlation shifts. These considerations were addressed through conservative thresholds and iterative monitoring, reducing the likelihood of systematic removal of informative samples.

Chapter – 6

Recommendations, conclusion, and limitations

Recommendations

The collective insights drawn from this research demonstrate that the success of healthcare AI depends overwhelmingly on the discipline, sophistication, and contextual awareness of the data preparation pipeline. Across the reviewed works ranging from transformer-based CT denoising, GAN-driven augmentation, orthopedics-focused AI applications, multimodal fusion models, and quality-consistent preprocessing it becomes evident that meaningful clinical AI requires data that is both technically reliable and clinically coherent. Based on these findings, several recommendations emerge:

Strengthen institution-level data refinement practices

Healthcare organisations should move away from rigid, rule-based ETL workflows and adopt *adaptive, learning-driven pipelines*. As shown across multiple studies, transformer-based denoising and multimodal harmonization improve the diagnostic reliability of downstream models. MADR-Net aligns with this direction by integrating a sequence of AI-enabled refinement stages that respond dynamically to data quality variations.

Adopt generative models to support balanced and diverse clinical datasets

GAN-based augmentation consistently improves diagnostic model stability by generating realistic representations of underrepresented classes. This is critical in fields such as orthopedics, where fracture subtypes often have limited labeled samples. Institutions should incorporate generative modelling not merely as a data expansion technique but as a controlled approach for reducing class imbalance and mitigating bias.

Implement noise-reduction and quality-consistency controls

Studies in CT imaging confirm that denoising transformers enhance both visual interpretability and downstream analytic accuracy. MADR-Net's quality-consistency stage embodies these principles by identifying anomalies, correcting acquisition artifacts, and reducing the signal distortions that commonly arise in multi-institutional datasets. This should be adopted as a standard step in all large-scale clinical data repositories.

Encourage multimodal integration for holistic clinical insight

AI systems in orthopedics, medical imaging, and general EHR analytics consistently show improved performance when multiple modalities are harmonized. MADR-Net's multimodal alignment stage accommodates structured, semi-structured, and unstructured data, enabling models to interpret clinical context more accurately. Healthcare centres should progressively transition toward multimodal data ecosystems rather than siloed datasets.

Maintain continuous clinical oversight and cross-functional governance

Findings from orthopedic AI research highlight the importance of clinical expertise in validating AI behaviour. Data refinement workflows should therefore integrate periodic clinical audits to ensure that corrections, imputations, and harmonizations remain clinically acceptable. Governance frameworks should enforce semantic standardization using ontologies such as FHIR, SNOMED CT, LOINC, and RxNorm.

Establish auditability and reproducibility as non-negotiable standards

As documented across the reviewed literature, reproducibility challenges remain a major barrier to clinical AI deployment. All refinement rules, ontological mappings, threshold decisions,

and model parameters should be version-controlled. MADR-Net incorporates this requirement, ensuring that each stage of refinement is traceable and defensible, particularly in PHI-restricted environments.

Conclusion

This report demonstrates that the central bottleneck in healthcare AI lies not in model architecture but in the character and readiness of the underlying data. Whether examining CT denoising, orthopedic diagnostic systems, GAN-based synthesis, or multimodal EHR fusion, the literature consistently shows that AI performance degrades when built upon incomplete, noisy, inconsistent, or semantically misaligned data. MADR-Net was developed specifically to address this systemic gap. The framework consolidates five critical refinement stages such as noise correction, anomaly detection, data reconstruction, semantic harmonization, and multimodal fusion into a unified and adaptive pipeline. Unlike conventional preprocessing which treats refinement as a sequence of isolated cleaning steps MADR-Net embeds intelligence within each stage, allowing the system to adjust to variations in clinical context, input modality, and institutional documentation practices. By applying GAN-supported augmentation, transformer-driven enhancement, LSTM-based temporal analysis, and ontology-grounded standardization, the framework transforms raw clinical data into *clinically coherent and analysis-ready datasets*. This elevates both the statistical reliability and the clinical interpretability of downstream AI models.

The findings reinforce a broader conclusion:

Advances in healthcare AI will be meaningful and clinically safe only when data preparation is treated as a scientific and adaptive process rather than a preliminary technical formality. MADR-Net thus contributes not only a technical workflow but also a conceptual shift: the recognition that robust data refinement is the foundation for trustworthy, equitable, and scalable AI in healthcare.

Computational overhead and practical feasibility

Inference Performance: On commodity GPU infrastructure (e.g., NVIDIA T4/V100), the full pipeline processes approximately 1,000 records in the range of seconds to low minutes, depending on feature dimensionality.

Modular Execution: Each model component can be executed independently, allowing institutions to adopt only those stages aligned with their infrastructure and needs. Scalability: The framework supports distributed execution (e.g., Spark + Databricks), making it compatible with hospital data warehouses and cloud-based analytics platforms. These additions will clarify that MADR-Net is designed for scalable, retrospective data preparation rather than latency-sensitive clinical workflows.

Mitigating the Risk of generative hallucinations in clinical data

Several safeguards are employed to mitigate hallucination risks:

Constrained Generation: Generative components (VAE-GANs) are used strictly for imputing missing or corrupted values within existing clinical feature boundaries, not for synthesizing new diagnoses, encounters, or events.

Schema and Clinical Rule Validation: All generated outputs are validated against predefined clinical schemas, allowable value ranges, temporal constraints, and domain rules derived from EHR standards (e.g., ICD, LOINC, encounter sequencing). Confidence Thresholding: Low-confidence imputations are explicitly flagged and excluded from downstream clinical modeling unless corroborated by additional evidence. Human-in-the-Loop Design: MADR-Net is positioned as a preprocessing layer for analytics and research use, with clinician review recommended for any operational or decision-support deployment. These clarifications will be explicitly added to the manuscript to distinguish statistical reconstruction from clinical inference and to emphasize patient safety.

Limitations

While MADR-Net represents a significant advance toward intelligent data preparation, several limitations must be acknowledged to contextualize its practical deployment:

Limited real-world access to high-fidelity clinical data

Stringent privacy regulations restrict access to complete, labeled datasets. Much like the studies reviewed particularly those in CT imaging and orthopedics many experiments rely on de-identified or partially synthetic data, potentially reducing real-world generalizability.

High computational demands for multi-stage AI pipelines

The framework integrates resource-intensive modules such as transformers, GANs, LSTMs, and semantic alignment engines. Deployment in smaller hospitals or low-resource settings may require scaled outputs or model distillation strategies.

Dependence on variable institutional data quality

EHR documentation practices differ widely across healthcare systems. As also noted in several reviewed orthopedics and imaging studies, models trained under one environment may require further calibration for another.

Complexity and interpretability challenges

Although MADR-Net enhances data quality, the internal mechanics of generative and transformer models may be difficult for non-technical clinical staff to interpret. This challenge reflects broader interpretability concerns raised in orthopedic AI and multimodal fusion literature.

Evolving standards and ontologies

Semantic alignment relies on current clinical vocabularies. However, terminology in specialties such as orthopedics evolves quickly, potentially limiting the long-term sufficiency of existing ontologies.

Dynamic clinical contexts

Changing care pathways, new imaging technologies, updated diagnostic protocols, and emerging diseases require continuous refinement of the pipeline. As highlighted in recent medical imaging research, static models degrade rapidly under temporal shift.

Acknowledgements

I would like to express my deepest gratitude to Dr. Shankar Srinivasan, Chief Chair and Program Director at Rutgers University, whose unwavering guidance and encouragement have been the cornerstone of my academic journey. His wisdom, patience, and extensive experience in healthcare informatics have been instrumental in shaping both the direction and fruition of this research.

My heartfelt thanks also go to Dr. Dasantila Sherifi, Director of the Health Information Management Program, Assistant Professor, and my Project Professor at Rutgers University. Her exceptional mentorship, keen insights, and continuous guidance throughout every stage of this project have been invaluable. Her thoughtful feedback and attention to detail greatly strengthened the quality, structure, and depth of this research.

I am deeply grateful to all the professors in the Department of Health Informatics at Rutgers University for their continuous inspiration, encouragement, and faith in my abilities. Their collective wisdom has played an invaluable role in my academic and professional growth.

I would also like to extend my sincere appreciation to Dr. K. Paul Jayakar, PhD, from India, for his tremendous support in reviewing the code and providing valuable technical feedback throughout this project. His expertise and constructive guidance were crucial in refining the analytical components of this research.

My special thanks go to my Guru, Thulibaba, whose spiritual guidance, blessings, and inspiration have been a constant source of strength and clarity throughout this journey.

Finally, my heartfelt appreciation goes to my husband, Nathan Muthuswamy, whose endless patience, love, and encouragement have been my greatest support over these four years. His unwavering belief in me made every challenge achievable and every milestone meaningful.

References

1. Nathan M, Sherifi D, Muthusamy V. Diabetes patients' readmission prediction. *IJSAT–International Journal on Science and Technology.* 2025;16(4).
2. Nathan M, Srinivasan S. A fresh look: The role of a healthcare data fabric in AI-driven predictions. *IJSAT–International Journal on Science and Technology.* 2025;16(4).
3. Nathan M, Mital DP. Review of alternative medicine (AM) treatments for diabetes. *J Diabetes Metab Disord Control.* 2024;11(2):80–83.
4. Nathan L, Muthusamy V. Uses, benefits and future of artificial intelligence (AI) in orthopedics. *Indian Journal of Medical Sciences.* 2024;76(2):95–97.
5. Ortiz BL, Gupta V, Kumar R, et al. Data preprocessing techniques for AI and machine learning readiness: Scoping review of wearable sensor data in cancer care. *JMIR mHealth and uHealth*. 2024;12(1):e59587.
6. Shahidi S, Samadzai AW, Shahbazi H. Effective data pre-processing in data science: from method selection to domain-specific optimization. *Journal of Advanced Computer Knowledge and Algorithms.* 2025;2(4):84–90.
7. Mohammad-Rahimi H, Sohrabniya F, Ourang SA, et al. Artificial intelligence in endodontics: Data preparation, clinical applications, ethical considerations, limitations, and future directions. *Int Endod J*. 2024;57(11):1566–1595.
8. Shen W, Zhou W. A novel Internet of medical things framework for absorbing bioresorbable vascular scaffold towards healthcare monitoring based on improving YOLO paradigms. *Knowledge-Based Systems.* 2025;322:113696.
9. Ganapriya K, Poobalan A, Kalaivani K, et al. Performance improvement for reconfigurable processor system design in IoT health care monitoring applications. *Tehnički vjesnik.* 2024;31(1):222–227.
10. Ramasamy M, Acharjya K, Singh Y, et al. Artificial intelligence-based effective healthcare prediction system for diabetic patients. *Multidisciplinary Science Journal*. 2024;6(1):2024ss0303.
11. Thethi SK. Machine learning models for cost-effective healthcare delivery systems: A global perspective. *Digital Transformation in Healthcare*. 2024;5:199.
12. Anjum M, Min H, Ahmed Z. Trivial state fuzzy processing for error reduction in healthcare big data analysis towards precision diagnosis. *Bioengineering*. 2024;11(6):539.
13. Sharma C, Vaid A, Saini MK. Artificial intelligence-driven fraud detection in SAP for retail and healthcare. *International Journal of Science and Research (IJSR)*. 2024;13(11):312–315.
14. Kumari M, Gaikwad M, Chavan SA. A secure IoT-edge architecture with data-driven AI techniques for early detection of cyber threats in healthcare. *Discover Internet of Things*. 2025;5(1):54.
15. Gusain R, Shekhar S, Vidyarthi A, et al. Smart healthcare system using machine learning and IoT. In *Embedded Devices and Internet of Things: Technologies and Applications*. 2024;153–177.
16. Priya EM, Krishnan KS. LIFE-CARE: IoT–cloud–enabled smart heart disease prediction system for smart healthcare environment using deep learning. *International Journal of Distributed Sensor Networks.* 2025(1):6965319.
17. Lohani BP, Vishnoi A, Das L, et al. Development of a healthcare model using machine learning. In *Progressive Computational Intelligence, Information Technology and Networking*. CRC Press. 2025;899–904.
18. Master H, Annis J, Ching JH, et al. Capturing real-world habitual sleep patterns with a novel user-centric algorithm to preprocess Fitbit data in the All of Us Research Program: Retrospective observational longitudinal study. *J Med Internet Res*. 2025;27:e71718.
19. Sinha N, Kumar MG, Joshi AM, et al. DASMcC: Data augmented SMOTE multi-class classifier for prediction of cardiovascular diseases using time series features. *IEEE Access*. 2023;11:117643–117655.
20. Natarajan A, Shanthi N. Optimizing healthcare big data privacy with scalable subtree-based L-anonymization in cloud environments. *Wireless Networks*. 2025;31(3):2727–2742.
21. Essaid S, Andre J, Brooks IM, et al. MENDS-on-FHIR: Leveraging the OMOP common data model and FHIR standards for national chronic disease surveillance. *medRxiv*. 2023:2023.08.09.23293900.
22. Marfoglia A, Nardini F, Arcobelli VA, et al. Towards real-world clinical data standardization: A modular FHIR-driven transformation pipeline to enhance semantic interoperability in healthcare. *Comput Biol Med*. 2025;187:109745.
23. Williams E, Kienast M, Medawar E, et al. A standardized clinical data harmonization pipeline for scalable AI application deployment (FHIR-DHP): Validation and usability study. *JMIR Med Inform*. 2023;11:e43847.

24. Ahmadi N, Zoch M, Guengoeze O, et al. How to customize common data models for rare diseases: an OMOP-based implementation and lessons learned. *Orphanet J Rare Dis*. 2024;19(1):298.

25. Xiao G, Pfaff E, Prud'hommeaux E, et al. FHIR-Ontop-OMOP: Building clinical knowledge graphs in FHIR RDF with the OMOP common data model. *J Biomed Inform*. 2022;134:104201.

26. Bönisch C, Kesztyüs D, Kesztyüs T. Harvesting metadata in clinical care: A crosswalk between FHIR, OMOP, CDISC and openEHR metadata. *Sci Data*. 2022;9(1):659.

27. Maletzky A, Böck C, Tschoellitsch T, et al. Lifting hospital electronic health record data treasures: challenges and opportunities. *JMIR Med Inform.* 2022;10(10):e38557.

28. Sadr AV, Li J, Hwang W, et al. Flexible imputation toolkit for electronic health records. *Sci Rep*. 2025;15(1):17176.

29. Li H, Apathy NC, Holmgren AJ, et al. Imputation of missing aggregate EHR audit log data across individual and multiple organizations. *J Biomed Inform*. 2025;163:104805.

30. Zhang C, Chu X, Ma L, Zhu Y, et al. M3Care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM. 2022;2418–2428.

31. Deng O, Jin Q. Missing data imputation based on dynamically adaptable structural equation modeling with self-attention. *arXiv preprint arXiv:2308.12388.* 2023.

32. Liu Y, Qin S, Yepes AJ, et al. Integrated convolutional and recurrent neural networks for health risk prediction using patient journey data with many missing values. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* 2022;1658–1663.

33. Liao W, Zhu Y, Zhang Z, et al. Learnable prompt as pseudo-imputation: Rethinking the necessity of traditional EHR data imputation in downstream clinical prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2025;765–776.

34. Zhou Y, Shi J, Stein R, et al. Missing data matter: an empirical evaluation of the impacts of missing EHR data in comparative effectiveness research. *J Am Med Inform Assoc*. 2023;30(7):1246–1256.

35. Cesare N, Were LP. A multi-step approach to managing missing data in time and patient variant electronic health records. *BMC Research Notes,* 2022;15(1):64.

36. Ahmed W, Rasool A, Javed AR, et al. Security in next-generation mobile payment systems: A comprehensive survey. *IEEE Access.* 2021;9:3105450.

37. Centers for Disease Control and Prevention. *National diabetes statistics report: Prevalence of both diagnosed and undiagnosed diabetes*; 2021.

38. Centers for Medicare & Medicaid Services. *Hospital readmissions reduction program*. 2025.

39. Dhaliwal JS, Dang AK. Reducing hospital readmissions. In *StatPearls* [Internet]. *StatPearls Publishing*; 2025.

40. Emir B, Masters ET, Mardekian J, et al. Identification of a potential fibromyalgia diagnosis using random forest modeling applied to electronic medical records. *Journal of Pain Research*. 2015;8:277–288.

41. Hsieh CJ. High glucose variability increases 30-day readmission rates in patients with type 2 diabetes hospitalized in the department of surgery. *Scientific Reports*. 2019;9:14240.

42. Ivanov N, Yan Q. System-wide security for offline payment terminals. In *Proceedings of the 17th EAI International Conference on Security and Privacy in Communication Networks (SecureComm).* 2021;99–119.

43. Khan BUI, Shah A, Goh KW, et al. Decentralized payment framework for low-connectivity areas using Ethereum blockchains. *Engineering, Technology & Applied Science Research.* 2024;14(6):17798–17810.

44. Kukde RD, Chakraborty A, Shah J. A systematic review of recent studies on hospital readmissions of patients with diabetes. *Cureus.* 2024;16(8):e67513.

45. Lin Y, Gao Z, Wang Q, et al. BcMON: Blockchain middleware for offline networks. *arXiv preprint.* 2022;arXiv:2204.01964.

46. Li R, Wang Q, Zhang X, et al. An offline delegatable cryptocurrency system (DelegaCoin). *arXiv preprint.* 2021;arXiv:2103.12905.

47. Mainetti L, Aprile M, Mele E, et al. A sustainable approach to delivering programmable peer-to-peer offline payments. *Sensors. 2023;23*(3):1336.

48. Nuckols TK. County-level variation in readmission rates: Implications for the hospital readmission reduction program's potential to succeed. *Health Services Research.* 2015;50(1):12–19.

49. Ostling S, Wyckoff J, Ciarkowski, SL. The relationship between diabetes mellitus and 30-day readmission rates. *Clinical Diabetes and Endocrinology.* 2017;3(3):3.

50. Reddy AVSK, Banda G. ElasticPay: Instant peer-to-peer offline extended digital payment system. *Sensors.* 2024;24(24):8034.

51. Sravan SS, Mandal S, Alphonse PJA, et al. A partial offline payment system for connecting the unconnected using Internet of Things: A survey. *ACM Computing Surveys. 2024;57*(2):1–35.

52. UC Irvine Machine Learning Repository. *Diabetes 130-US hospitals for years 1999–2008;* 2025.

53. Yoon J, Kim Y. Offline payment of central bank digital currency based on a trusted platform module. *Journal of Cybersecurity and Privacy.* 2025;5(2):14.

54. Yi J, Kim J, Oh YK. Uncovering the quality factors driving the success of mobile payment apps. *Journal of Retailing and Consumer Services*. 2024;77:103641.

55. Alam MN, Kaur M, Kabir MS. Explainable AI in healthcare: Enhancing transparency and trust upon legal and ethical consideration. *International Research Journal of Engineering and Technology.* 2023;10:1–9.

56. AlZu'bi S, Elbes M, Mughaid A, et al. Diabetes monitoring system in smart health cities based on big data intelligence. *Future Internet*. 2023;15(2):85.

57. Bajwa J, Munir U, Nori A, et al. Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Health Journal.* 2021;8(2):e188–e194.

58. Benke K, Benke G. Artificial intelligence and big data in public health. *International Journal of Environmental Research and Public Health.* 2018;15(12):2796.

59. Boehm KM, Aherne EA, Ellenson L, et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nature Cancer.* 2022;3:723–733.

60. Budd J, Miller BS, Manning E, et al. Digital technologies in the public health response to COVID-19. *Nature Medicine.* 2020;26, 1183–1192.

61. Ferranti JM, Langman MK, Tanaka D, et al. Bridging the gap: Leveraging business intelligence tools in support of patient safety and financial effectiveness. *Journal of the American Medical Informatics Association. 2010;17*(2):136–143.

62. Global Market Insights. *Healthcare business intelligence market report.* Global Market Insights Inc; 2024.

63. Guo Y, Bian J, Modave F, et al. Assessing the effect of data integration on predictive ability of cancer survival models. *Health Informatics Journal.* 2019;26(1):8–20.

64. Isoviita VM, Salminen L, Azar J, et al. Open-source infrastructure for healthcare data integration and machine learning analyses. *JCO Clinical Cancer Informatics.* 2019;3:1–16.

65. Jensen LR. Using data integration to improve health and welfare insights. *International Journal of Environmental Research and Public Health.* 2022;19(3):836.

66. Kazemi-Arpanahi H, Shanbehzadeh M, Mirbagheri E, et al. Data integration in cardiac electrophysiology ablation toward achieving proper interoperability in health information systems. *Journal of Education and Health Promotion*. 2020;9:262.

67. Lee G, Kang B, Nho K, et al. MildInt: Deep learning-based multimodal longitudinal data integration framework. *Frontiers in Genetics.* 2019;10:617.

68. Li N, Zhu Q, Dang Y, et al. Development and implementation of a dynamically updated big data intelligence platform using electronic medical records for secondary hypertension. *Reviews in Cardiovascular Medicine.* 2024;25:104.

69. Lian W, Xue T, Lu Y. Research on hierarchical data fusion of intelligent medical monitoring. *IEEE Access*. 2020;8:38355–38367.

70. Lin L, Liang W, Li CF. Development and implementation of a dynamically updated big data intelligence platform from electronic health records for nasopharyngeal carcinoma research. *Br J Radiol*. 2019;92(1102):20190255.

71. Liu Y, Zhang L, Yang Y. A novel cloud-based framework for elderly healthcare services using digital twin. *IEEE Access*. 2019;7:49088–49101.

72. Liu Z, Wen H, Zhu Z. Diagnosis of significant liver fibrosis in patients with chronic hepatitis B using a deep learning-based data integration network. *Hepatology International*. 2022;16(3):526–536.

73. Mandl KD, Gottlieb D, Mandel JC. Integration of AI in healthcare requires an interoperable digital data ecosystem. *Nature Medicine*. 2024;30(3):1–4.

74. Manogaran G, Thota C, Lopez D. Big data security intelligence for healthcare industry 4.0. In *Cybersecurity for Industry.* 2017;4:103–126.

75. Martínez-García M, Hernández-Lemus E. Data integration challenges for machine learning in precision medicine. *Frontiers in Medicine*. 2022;8:784455.

76. Mirzaei A, Aslani P, Schneider CR. Healthcare data integration using machine learning: A case study evaluation with health information-seeking behavior databases. *Research in Social and Administrative Pharmacy*. 2022;18(12):4144–4149.

77. Norori N, Hu Q, Aellen FM. Addressing bias in big data and AI for healthcare: A call for open science. *Patterns*. 2021;2(1):100218.

78. Parciak M, Suhr M, Schmidt C. FAIRness through automation: Development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital. *BMC Med Inform Decis Mak*. 2023;23(1):94.

79. Perkins SW, Muste JC, Alam T. Improving clinical documentation with artificial intelligence: A systematic review. *Perspect Health Inf Manag.* 2024;21(2):1d.

80. Reda R, Piccinini F, Martinelli G. Heterogeneous self-tracked health and fitness data integration and sharing according to a linked open data approach. *Computing*. 2022;104(6):835–857.

81. Ross P, Spates K. Considering the safety and quality of artificial intelligence in health care. *Jt Comm J Qual Patient Saf.* 2020;46(10):596–603.

82. Shah V, Shukla S. Data distribution into distributed systems, integration, and advancing machine learning. *Revista Española de Documentación Científica*. 2017;11(1):1–17.

83. Syed K, Sleeman WC, Hagan M. Multi-view data integration methods for radiotherapy structure name standardization. *Cancers.* 2021;13(8):1796.

84. Tang H. Intelligent processing and classification of multisource health big data from the perspective of physical and medical integration. *Scientific Programming*. 2022;1–10.

85. Thapa C, Camtepe S. Precision health data: Requirements, challenges and existing techniques for data security and privacy. *Comput Biol Med*. 2021;129:104130.

86. Wegner P, Jose GM, Lage-Rupprecht V. Common data model for COVID-19 datasets. *Bioinformatics*. 2022;38(21):5466–5468.

87. World Health Organization. *Global action plan on the public health response to dementia 2017–2025.* World Health Organization. 2017.

88. Yang X, Chen A, PourNejatian N. A large language model for electronic health records. *NPJ Digital Medicine.* 2022;5(1):194.

89. Yaqoob I, Salah K, Jayaraman R. Blockchain for healthcare data management: Opportunities, challenges, and future recommendations. *Neural Computing and Applications*. 34(1):12275–12293.

90. Zhang Q, Lian B, Cao P. Multi-source medical data integration and mining for healthcare services. *IEEE Access.* 2020;8:165010–165017.

91. Zhao J, Feng Q, Wu P. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Sci Rep*. 2019;9(1):717.

92. Zoega H, Falster MO, Gillies MB. The Medicines Intelligence Data Platform: A population-based data resource from New South Wales, Australia. *Pharmacoepidemiol Drug Saf*. 2024;33(8):5887.

93. Chattopadhyay K, Wang H, Kaur J. Effectiveness and safety of Ayurvedic medicines in type 2 diabetes mellitus management: A systematic review and meta-analysis. *Front Pharmacol*. 2022;13:821810.

94. Gordon A, Buch Z, Baute V. Use of Ayurveda in the treatment of type 2 diabetes mellitus. *Glob Adv Health Med.* 2019;8:2164956119861094.

95. Hardy ML, Coulter I, Venuturupalli S. Ayurvedic interventions for diabetes mellitus: A systematic review. *AHRQ Evidence Report Summaries*. 2001(1):1–5.

96. Kumari S, Laxmikant SD, Sonika B. Efficacy of integrated Ayurveda treatment protocol in type 2 diabetes mellitus – A case report. *J Ayurveda Integr Med*. 2021;13(1):100512.

97. Thomas V. Ayurveda approach in the treatment of type 2 diabetes mellitus: A case report. *J Ayurveda Integr Med*. 2023;14(4):100744.

98. Guo Z, Gu Z, Zheng B, et al. Transformer for image harmonization and beyond. *IEEE Trans Pattern Anal Mach Intell*. 2021;45(11):12960–12977.

99. Hao X, Liu Y, Pei L, et al. Atmospheric temperature prediction based on a BiLSTM-attention model. *Symmetry*. 2021;14(11):2470.

100. Shao H, Jiang H, Zhao H, et al. A novel deep autoencoder feature learning method for rotating machinery fault diagnosis. *Mechanical Systems and Signal Processing*. 2017;95:187–204.

101. Yan K, Su J, Huang J, et al. Chiller fault diagnosis based on VAE-enabled generative adversarial networks. *IEEE Transactions on Automation Science and Engineering*. 2020;19(1):387–395.

102. Zhao Z, Zhang Z, Chen T, et al. Image augmentations for GAN training. *arXiv preprint arXiv:*2006.02595.