A Survey of Multimodal Temporal Data and Deep Learning in Healthcare

www.surveyx.cn

Abstract

In the rapidly evolving field of healthcare, the integration of multimodal temporal data with deep learning techniques is transforming clinical decision support systems (CDSS) and enhancing patient outcomes. This survey paper examines the intersection of these technologies, highlighting their roles in enriching data richness, predictive accuracy, and interpretability. Multimodal temporal data, encompassing diverse sources like clinical records and imaging, is critical for informed decision-making, especially in high-stakes environments such as COVID-19 management and chronic disease monitoring. Deep learning models, including Convolutional Neural Networks and Long Short-Term Memory networks, facilitate the analysis of complex datasets, offering improved diagnostic accuracy and personalized treatment strategies. The paper also explores the integration of reinforcement learning to optimize treatment strategies and the importance of uncertainty quantification in clinical decision-making. Despite these advancements, challenges in model interpretability and integration into clinical workflows persist. The survey underscores the need for robust interpretability frameworks and methodologies that align with clinical practices, ensuring transparency and trust in AI-driven healthcare solutions. Future directions include enhancing Neural Additive Models, expanding the applicability of Explainable AI, and integrating socio-structural understanding into AI development. By addressing these challenges, the field can advance toward more effective and reliable healthcare solutions, ultimately improving patient care and outcomes.

1 Introduction

1.1 Significance of Multimodal Temporal Data

Multimodal temporal data is essential for enriching clinical decision-making in healthcare by integrating visual, textual, and numerical data, leading to a more comprehensive understanding of patient health. For instance, in advanced heart failure management, such data is crucial for identifying patients who would benefit from specific treatments, thus optimizing therapeutic outcomes [1]. In high-stress scenarios, such as ventilator allocation during the COVID-19 pandemic, multimodal temporal data enhances decision-making by offering insights into patient needs and resource distribution [2]. The fusion of machine learning techniques with multimodal data has shown promise in diagnosing mental health conditions like depression, enriching diagnostic data and supporting effective decision-making processes [3].

The use of chest CT images within multimodal temporal data is particularly significant for diagnosing and managing COVID-19, enhancing data richness and informing clinical decisions [4]. The MAP model emphasizes the importance of user-centered design in AI systems, ensuring they meet the needs of medical professionals and improve clinical outcomes [5]. In predicting sepsis, the integration of electronic health records with time-series data showcases the critical role of multimodal temporal data in healthcare, enhancing prediction accuracy and facilitating timely interventions [6]. Similarly,

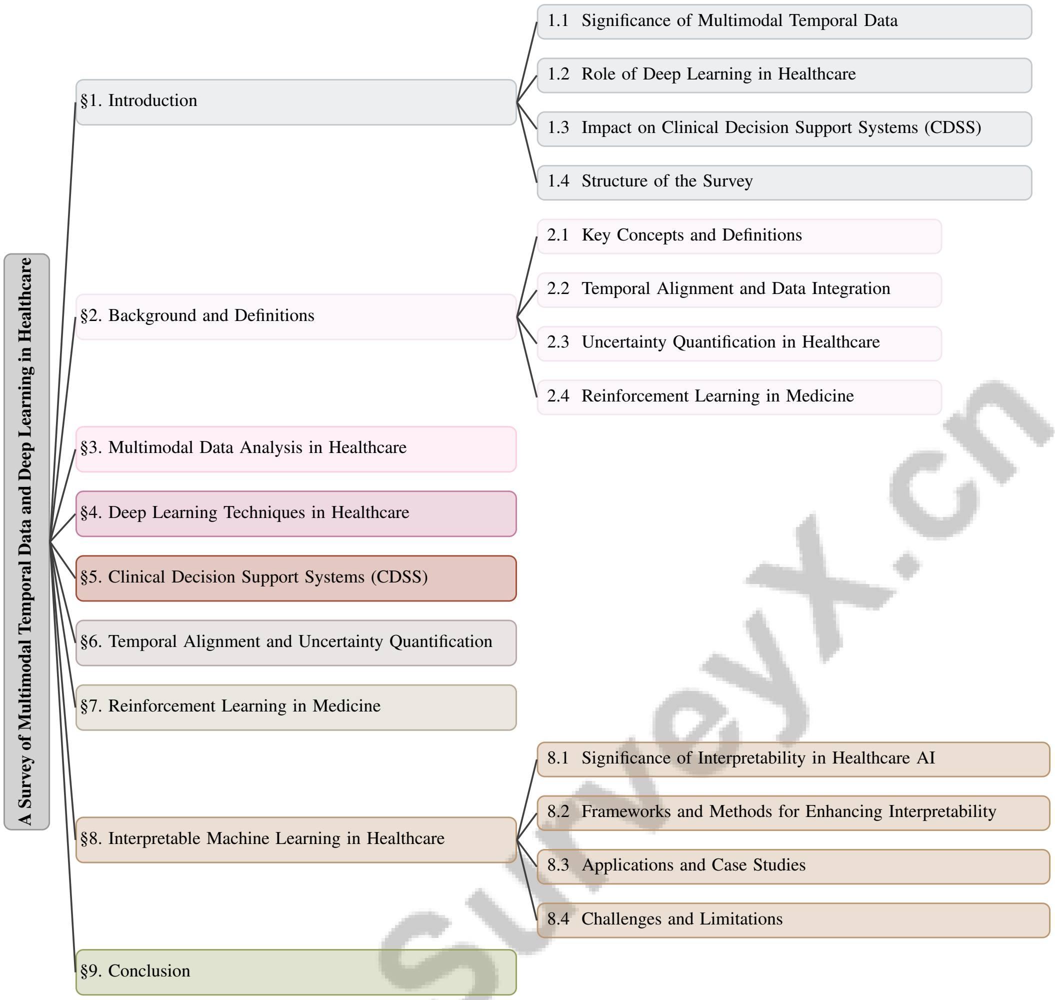

A Survey of Multimodal Temporal Data and Deep Learning in Healthcare
§1. Introduction
1.1 Significance of Multimodal Temporal Data
1.2 Role of Deep Learning in Healthcare
1.3 Impact on Clinical Decision Support Systems (CDSS)
1.4 Structure of the Survey
§2. Background and Definitions
2.1 Key Concepts and Definitions
2.2 Temporal Alignment and Data Integration
2.3 Uncertainty Quantification in Healthcare
2.4 Reinforcement Learning in Medicine
§3. Multimodal Data Analysis in Healthcare
§4. Deep Learning Techniques in Healthcare
§5. Clinical Decision Support Systems (CDSS)
§6. Temporal Alignment and Uncertainty Quantification
§7. Reinforcement Learning in Medicine
§8. Interpretable Machine Learning in Healthcare
8.1 Significance of Interpretability in Healthcare AI
8.2 Frameworks and Methods for Enhancing Interpretability
8.3 Applications and Case Studies
8.4 Challenges and Limitations
§9. Conclusion

Figure 1: chapter structure

for elderly ICU patients with multiple organ dysfunction syndrome (MODS), early detection through multimodal data is vital for accurate mortality prediction and improved clinical decision-making [7].

Moreover, the integration of diverse data sources, such as electronic health records and claims data, further illustrates how multimodal temporal data enhances decision-making in healthcare [8]. Addressing inconsistencies in labeling cough audio data through multimodal approaches also improves diagnostic accuracy in the context of COVID-19 [9]. Interpretable models utilizing multimodal temporal data are crucial in predicting severe maternal morbidity, enhancing the understanding of risk factors and supporting better clinical decision-making [10]. Collectively, these examples highlight the transformative potential of multimodal temporal data in improving healthcare outcomes through enhanced data richness and informed decision-making. Additionally, the increasing legalization of cannabis raises concerns about its potential negative health impacts, particularly among young adults, underscoring the need for comprehensive data to inform public health strategies [11].

1.2 Role of Deep Learning in Healthcare

Deep learning is a pivotal tool in healthcare, enabling the analysis of complex datasets and significantly enhancing predictive accuracy, thereby transforming clinical decision-making processes [12]. The MGP-AttTCN model exemplifies this by integrating multitask Gaussian Processes with attention mechanisms to improve early sepsis prediction while maintaining interpretability, showcasing deep learning's transformative potential in healthcare [6]. Additionally, interpretable machine learning models, such as those using XGBoost combined with SHAP, highlight deep learning's capability to enhance predictive accuracy through improved interpretability [7].

AI-generated, for reference only.

Incorporating deep learning into healthcare workflows is illustrated by the MAP model, which offers a structured approach to understanding Human-AI interaction, emphasizing AI's integration into healthcare processes [5]. This is further supported by two-tiered machine learning models leveraging extensive claims and eligibility data to enhance predictive accuracy [8]. The application of Large Language Models (LLMs) in Clinical Decision Support Systems (CDSS) significantly improves the identification of medication errors, enhancing patient safety and care quality [13].

Deep learning models have also been instrumental in forecasting COVID-19 infections using the Temporal Fusion Transformer model, which improves predictive accuracy by incorporating temporal data [14]. Furthermore, the integration of probabilistic scoring lists into deep learning models allows for uncertainty incorporation in decision-making processes, which is crucial in high-stakes environments [15]. Advanced AI techniques have been employed to analyze cannabis intoxication behaviors, providing personalized insights to assist clinicians in understanding patient-specific responses and tailoring interventions [11]. These advancements underscore deep learning's vital role in healthcare, improving predictive analytics and patient outcomes while facilitating the integration of diverse data sources into comprehensive clinical models.

1.3 Impact on Clinical Decision Support Systems (CDSS)

The integration of multimodal data and deep learning into Clinical Decision Support Systems (CDSS) marks a significant advancement in healthcare, enhancing decision support and patient outcomes through complex datasets and sophisticated algorithms. The MGP-AttTCN model aids clinicians by integrating multimodal data and deep learning, demonstrating the transformative potential of these technologies in CDSS [6]. Similarly, the UDC-Net method enhances decision support by automatically segmenting COVID-19 lesions from chest CT images, assisting in screening and treatment planning, thereby underscoring deep learning's practical benefits [4].

Models providing interpretable insights into mortality risk factors, particularly for elderly patients with MODS, further illustrate deep learning's role in enhancing decision support for clinicians [7]. These models improve predictive accuracy while offering transparent insights crucial for clinical decision-making. Additionally, proposed deep learning models for predicting post-COVID-19 fatigue aim to enhance diagnostic accuracy and patient outcomes through timely interventions, highlighting the relevance of predictive analytics in modern CDSS [16].

Despite these advancements, challenges remain, particularly regarding irrelevant alerts in CDSS that can lead to alert fatigue among healthcare providers. This issue is prevalent in rules-based systems, emphasizing the need for sophisticated, data-driven approaches to minimize unnecessary alerts and enhance clinical utility [13]. Furthermore, the implications of AI decision-making in high-stakes environments necessitate careful consideration of model accuracy and reliability [5].

Advancements in multimodal data integration and deep learning methodologies are transforming CDSS by enhancing accuracy, interpretability, and efficiency in clinical decision-making. These developments facilitate the synthesis of diverse data types—such as medical images and electronic health records—thereby improving predictive capabilities and enabling informed clinical judgments. Moreover, the integration of interactive frameworks and intuitive visualizations in systems like CarePre addresses interpretability challenges, ensuring effective tool utilization by healthcare professionals. Despite the promising potential of these innovations, ongoing challenges, including data biases and the need for robust validation, must be addressed for successful implementation in clinical settings [17, 18, 19, 12, 20]. By addressing existing challenges and leveraging advanced computational techniques, CDSS can significantly improve patient care and outcomes, marking a transformative shift in healthcare delivery.

1.4 Structure of the Survey

This survey is structured to explore the intersection of multimodal temporal data and deep learning within healthcare. It begins with an introduction that highlights the significance of these technologies in enhancing clinical decision-making and personalized medicine, emphasizing their roles in enriching data richness and predictive accuracy.

The second section provides foundational knowledge on key concepts such as multimodal temporal data, deep learning, and Clinical Decision Support Systems (CDSS), establishing a clear understanding of the terminologies and frameworks that underpin subsequent discussions.

The survey then investigates various challenges and methodologies involved in multimodal data analysis within healthcare, focusing on issues such as data biases, effective representation and fusion of diverse data types, and the importance of interpretability in machine learning models to enhance clinical decision support systems [21, 17, 12, 22]. This includes examining the integration and analysis of diverse data sources, emphasizing the need for effective data integration techniques to enhance predictive analytics. Case studies and applications illustrate successful implementations in real-world healthcare settings.

Deep learning techniques in healthcare are explored in the fourth section, focusing on medical image sequence analysis and predictive modeling, addressing advancements and challenges in applying deep learning to medical data, including the need for model interpretability and transparency.

The fifth section discusses the development and implementation of CDSS, emphasizing the integration of deep learning and multimodal data to improve decision-making and patient outcomes. It also addresses implementation challenges and proposes solutions, drawing on innovations such as the Clinical Evidence Engine, which retrieves relevant clinical trial reports and extracts key PICO elements [23].

Temporal alignment and uncertainty quantification are analyzed in the sixth section, highlighting their importance in managing time-varying data and enhancing clinical decision-making. Techniques for achieving temporal alignment and methods for quantifying uncertainty are discussed in detail.

The survey explores the application of reinforcement learning in medicine, focusing on its potential for optimizing treatment strategies and developing context-aware systems that enhance personalized medicine, examining frameworks that incorporate risk-awareness into decision-making processes.

Interpretable machine learning in healthcare is the focus of the eighth section, underscoring the importance of interpretability in healthcare AI. Various frameworks and methods designed to enhance model interpretability are explored, alongside real-world applications and case studies.

The survey concludes with a summary of key findings and a discussion of future directions and research opportunities in the field, reflecting on potential advancements in multimodal temporal data and deep learning, and highlighting areas for further exploration and innovation.

Throughout the survey, innovations such as the Modular Decision Network (MoDN), which enables effective learning from IIO datasets without data sharing, are referenced to illustrate cutting-edge approaches being developed in this field [24]. This structured approach ensures a thorough examination of the current landscape and future potential of these transformative technologies in healthcare.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Key Concepts and Definitions

Understanding key concepts in healthcare analytics, such as multimodal temporal data, deep learning, and Clinical Decision Support Systems (CDSS), is essential for advancing clinical decision-making. Multimodal temporal data involves integrating diverse data types, including clinical variables, imaging features, and time-series data, to enhance predictive capabilities. This integration tackles challenges like systematic missingness and heterogeneous measurements, which are crucial for effective analysis [9]. Complexities such as clock-drift and sensor data offsets require robust methodologies for temporal alignment [25].

Deep learning, a subset of AI, uses algorithms like CNNs and RNNs to analyze complex datasets, enabling medical condition classification and prediction from imaging data [12]. Interpretability in these models is vital for transparency and reliability, as seen in neonatal MRI analysis [26]. Developing interpretable models is crucial for mitigating ECG misdiagnosis risks [27], and the lack of transparency in AI models predicting cannabis intoxication behaviors limits clinical utility, emphasizing the need for algorithmic explainability [11].

AI-generated, for reference only.

CDSS leverage advanced methodologies to support data-driven decisions, improving patient outcomes. Usability and trust challenges among healthcare professionals must be addressed for effective AI adoption [5]. Imperfectly interoperable datasets due to missing features highlight the need for modular and flexible systems [24]. Interpretable risk prediction models for ordinal outcomes enhance CDSS utility by providing nuanced insights beyond binary outcomes [28].

Integrating interpretable AI into CDSS frameworks is crucial for addressing mental health disorders, establishing a foundational understanding of the challenges and opportunities in leveraging machine learning for diagnostics [3]. By overcoming these challenges and incorporating advanced computational techniques, CDSS can significantly enhance predictive accuracy and clinical utility across various healthcare settings.

2.2 Temporal Alignment and Data Integration

Temporal alignment is crucial in healthcare data analysis, enabling the integration of diverse data modalities into cohesive models that enhance predictive accuracy and decision-making. This process is vital for continuous physiological monitoring data, facilitating real-time analysis and interventions [29]. Effective temporal alignment synchronizes data from sources like electronic health records, wearable sensors, and imaging modalities, providing a comprehensive view of patient health [17].

Addressing the complexities of aligning non-parallel sequences, such as articulatory and acoustic data, is critical for accurate interpretation, as demonstrated in phonetic data integration [30]. In biomedical contexts, temporal alignment ensures consistency and reliability in data interpretation, particularly in telemonitoring and clinical trials [25].

Integrating historical data, such as COVID-19 infection rates with mobility data, illustrates the importance of temporal alignment for forecasting and decision-making. By aligning these datasets, researchers can develop models that accurately predict infection trends and inform public health strategies [31]. Integrating claims data over time enhances cost prediction accuracy, demonstrating its value in financial modeling and resource allocation [8].

In rural clinical settings, AI-based Clinical Decision Support Systems (AI-CDSS) require alignment with existing workflows. Temporal alignment is crucial for effectively incorporating AI-CDSS into unique clinical environments, addressing challenges like limited resources and varying practices [32]. By facilitating seamless multimodal data integration, temporal alignment enhances healthcare systems' effectiveness, supporting timely and informed clinical decision-making.

2.3 Uncertainty Quantification in Healthcare

Uncertainty quantification (UQ) is pivotal in healthcare analytics, enhancing the robustness and reliability of clinical decision-making by addressing inherent uncertainties in medical data and predictive models. In respiratory disorder classification, incorporating UQ into semi-supervised learning improves cough sound classification accuracy, supporting reliable diagnostics [9]. This underscores UQ's critical role in refining model outputs and ensuring consistency in automated assessments.

Challenges in uncertainty estimation are pronounced in small data regimes, where traditional models may yield overconfident predictions, misleading practitioners in high-stakes scenarios [33]. The degradation of uncertainty capabilities in state-of-the-art models with insufficient training data emphasizes the need for robust UQ strategies adaptable to varying data availability [33].

Incorporating UQ into healthcare models is essential for addressing ethical considerations like transparency and explainability, vital for building trust in AI-driven decision systems [11]. The integration of explainable AI (XAI) techniques with real-time sensor data enhances understanding and prediction of complex behaviors, such as cannabis intoxication, by providing interpretable insights for clinical interventions [11].

Developing interpretable machine learning methods that incorporate UQ is crucial for decision-making in safety-critical healthcare domains. These methods enable uncertainty integration into predictive models, enhancing interpretability and ensuring clinical decisions are informed by a comprehensive understanding of potential risks and outcomes [15]. This approach is indispensable

AI-generated, for reference only.

for capturing the nuances of ordinal data, avoiding pitfalls of binary classifications that overlook clinically relevant information.

Collectively, these advancements highlight UQ's importance in healthcare, enhancing predictive accuracy and clinical utility while ensuring informed and reliable decision-making. By addressing uncertainty estimation challenges and integrating advanced UQ techniques—such as Bayesian approximation and ensemble learning—into healthcare systems, practitioners can improve medical decision-making accuracy. This leads to better patient outcomes through enhanced classification and risk assessment, fostering resilient healthcare practices that adapt to real-world complexities. Utilizing robust evaluation metrics like the area under Confidence-Classification Characteristic curves (AUCCC) can further assess UQ methods' performance, promoting a more reliable healthcare environment [34, 35].

2.4 Reinforcement Learning in Medicine

Reinforcement learning (RL) is emerging as a transformative paradigm in medicine, providing innovative solutions for optimizing treatment strategies and enhancing patient care. By employing trial-and-error learning principles, RL enables healthcare systems to adaptively refine decision-making processes based on environmental feedback. This approach is beneficial in dynamic clinical environments, where patient care complexities often exceed traditional static models' capabilities, enabling real-time, interpretable insights that adapt to evolving patient data [21, 36].

A critical challenge in applying RL to medicine is ensuring robust uncertainty quantification, essential for reliable predictions and decisions across diverse patient populations. Developing benchmarks for uncertainty estimation in brain-computer interface (BCI) systems exemplifies efforts to enhance RL model trustworthiness in cross-subject scenarios [37]. This focus on uncertainty is crucial for RL applications in healthcare, given the significant implications of decisions on patient safety and outcomes.

Future RL research in healthcare should consider relaxing the independent and identically distributed (i.i.d.) assumptions prevalent in many current models. Improving the robustness of Probably Approximately Correct (PAC) guarantees in more complex and realistic scenarios is vital for advancing RL's applicability in clinical settings [38]. This involves developing sophisticated algorithms capable of managing inherent variability and uncertainty in medical data.

Integrating multi-level, multi-fidelity (MLMF) frameworks with RL can further enhance decision-making in healthcare by incorporating additional reduced-order models, extending applications to a broader array of cardiovascular scenarios [39]. Such advancements can lead to more accurate and efficient treatment strategies, ultimately improving patient outcomes.

The application of RL in medicine is bolstered by advancements in statistical methods for sequential decision-making. New safe, anytime-valid concentration bounds and frameworks for risk-aware contextual bandits represent significant progress in improving healthcare decision-making processes [40]. These innovations provide a foundation for RL algorithms to function effectively in uncertain and dynamic environments, ensuring adaptive and reliable patient care.

Furthermore, the AutoScore-Ordinal framework offers a systematic approach to generating interpretable and clinically useful scoring models for predicting ordinal outcomes, which can be integrated with RL to enhance interpretability and clinical utility [28]. By combining RL with interpretable machine learning techniques, healthcare providers can gain deeper insights into treatment strategies and patient responses, facilitating more personalized and effective interventions.

Recent advancements in deep learning (DL) and machine learning (ML) technologies underscore their transformative potential in healthcare, particularly in optimizing treatment strategies and enhancing patient care. These innovations enable the extraction of complex patterns from vast health data, facilitating personalized decision-making that improves service quality. Moreover, developing interpretable ML models allows healthcare professionals to better understand and explain predictions, supporting data-driven decisions that lead to more effective treatment outcomes and streamlined clinical processes [21, 41, 12]. By addressing challenges related to uncertainty, model robustness, and interpretability, RL can significantly enhance the efficacy and reliability of clinical decision-making processes, paving the way for more adaptive and patient-centered healthcare solutions.

3 Multimodal Data Analysis in Healthcare

The integration of multimodal data in healthcare is increasingly essential, addressing the complexities inherent in diverse data types such as clinical records and sensor outputs. This section explores the challenges and methodologies involved in multimodal data analysis, emphasizing their implications for clinical applications and predictive systems.

3.1 Challenges in Multimodal Data Analysis

The analysis of multimodal data in healthcare faces several challenges that impede the development of effective clinical decision-making systems. Traditional actuarial models often fail to predict costs accurately, particularly for smaller employer groups, due to the complexity of healthcare data [8, 12]. Discrepancies in expert interpretation of data types like cough audio samples further complicate decision-making processes [9]. Additionally, there is a tension between achieving high prediction accuracy and maintaining interpretability of individual feature contributions in non-stationary time series data, which is crucial for understanding health outcomes [14].

Benchmarks often lack contextually relevant alerts, leading to high rates of ignored safety alerts and ineffective medication error identification [13]. The high dimensionality and complexity of multimodal datasets can lead to cognitive overload for clinicians, reducing the effectiveness of clinical decision support systems. Addressing these issues requires advanced analytical frameworks capable of integrating and interpreting diverse data sources, ultimately enhancing predictive accuracy and informed clinical decision-making [21, 17, 24].

3.2 Methodologies for Effective Data Integration

Effective multimodal data integration is crucial for advancing predictive analytics in healthcare. Methods like the integration of structured epidemiological data with unstructured mobility data exemplify multimodal learning approaches that enhance predictive analytics [31]. Innovative methodologies such as the CCT-learner and CMC meta-learner generate predictive distributions of individual treatment effects through advanced statistical techniques, improving model accuracy and reliability [42].

The Deep Canonical Correlation Alignment (DCCA) method facilitates direct alignment of raw signals, simplifying the integration of diverse data sources [25]. For time series data, saliency map generation methodologies improve interpretability and robustness of predictive models [43]. Addressing inter-rater variability in medical image segmentation, modified loss functions enhance segmentation accuracy and uncertainty quantification [44].

Collectively, these methodologies underscore the importance of effective data integration in predictive analytics, leading to more accurate healthcare outcomes. By utilizing sophisticated computational techniques, these approaches foster the development of resilient predictive models that refine clinical decision-making processes and provide interpretable insights [21, 45, 17].

As illustrated in Figure 2, this figure illustrates methodologies for effective data integration in healthcare analytics, focusing on multimodal learning, predictive distributions, and data alignment and interpretation. Each category highlights innovative approaches and techniques that enhance predictive accuracy and model interpretability. Integrating multimodal data is vital for enhancing patient outcomes and advancing medical research. The methodologies highlighted provide a comprehensive overview of approaches to harmonize disparate data sources into a unified analytical framework. These examples underscore the importance of selecting appropriate models and understanding their implications, fostering transparency and comprehensibility in healthcare analytics [46, 47, 48].

3.3 Case Studies and Applications

The integration and analysis of multimodal data in healthcare have significantly advanced diagnostic accuracy and patient outcomes. For instance, uncertainty-based rejection techniques in melanoma detection have improved diagnostic accuracy and reduced misdiagnoses [49]. Similarly, integrating electronic health records with real-time sensor data has facilitated early sepsis detection, showcasing the potential of multimodal data for timely interventions and patient outcomes [21, 12].

AI-generated, for reference only.

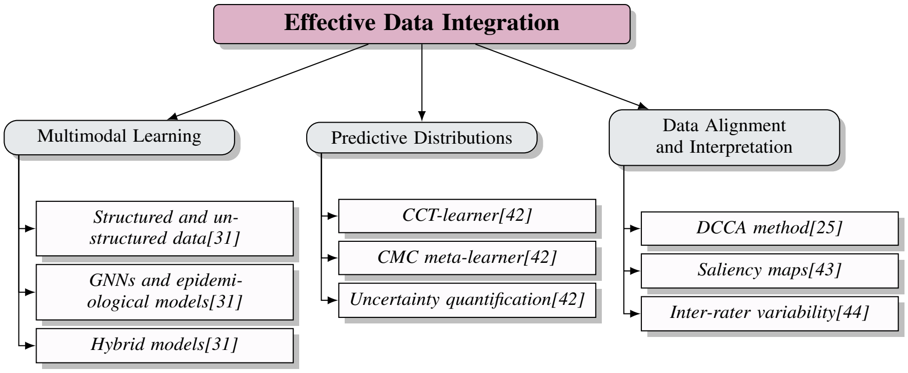
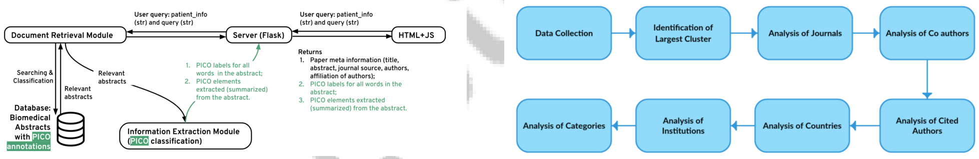

Effective Data Integration
Multimodal Learning
Predictive Distributions
Data Alignment and Interpretation
Structured and unstructured data[31]
GNNs and epidemiological models[31]
Hybrid models[31]
CCT-learner[42]
CMC meta-learner[42]
Uncertainty quantification[42]
DCCA method[25]
Saliency maps[43]
Inter-rater variability[44]

Figure 2: This figure illustrates methodologies for effective data integration in healthcare analytics, focusing on multimodal learning, predictive distributions, and data alignment and interpretation. Each category highlights innovative approaches and techniques that enhance predictive accuracy and model interpretability.

In chronic disease management, combining genomic data with lifestyle and environmental factors has led to personalized treatment strategies, improving precision in medical interventions [21, 12]. In mental health, integrating textual data from patient interviews with physiological measurements has enhanced depression diagnosis accuracy, demonstrating the effectiveness of multimodal approaches in clinical decision-making [21, 17, 19, 24].

These case studies illustrate the significant impact of multimodal data integration and analysis in healthcare, demonstrating how advanced computational techniques enhance diagnostic accuracy, facilitate personalized medicine, and improve overall patient care. By leveraging these approaches, healthcare providers can make more informed clinical predictions and decisions, ultimately advancing the quality of care delivered to patients [21, 17, 12].


User query: patient_info (str) and query (str)
Document Retrieval Module
Server (Flask)
User query: patient_info (str) and query (str)
HTML+JS
Returns
1. Paper meta information (title, abstract, journal source, authors, affiliation of authors);
2. PICO labels for all words in the abstract;
3. PICO elements extracted (summarized) from the abstract.
1. PICO labels for all words in the abstract;
2. PICO elements extracted (summarized) from the abstract.
Searching & Classification
Relevant abstracts
Relevant abstracts
Database: Biomedical Abstracts with PICO annotations
Information Extraction Module (PICO classification)
Data Collection
Identification of Largest Cluster
Analysis of Journals
Analysis of Co authors
Analysis of Cited Authors
Analysis of Countries
Analysis of Institutions
Analysis of Categories

(a) A diagram illustrating the process of information retrieval and extraction in a document-based system[23]

(b) Research Methodology Flowchart[19]

Figure 3: Examples of Case Studies and Applications

As depicted in Figure 3, the integration of multimodal data analysis is increasingly crucial for enhancing patient outcomes and streamlining clinical decision-making processes. These examples demonstrate the transformative potential of multimodal data analysis in healthcare, highlighting its ability to harness diverse data sources for improved clinical decision support systems [23, 19].

4 Deep Learning Techniques in Healthcare

The incorporation of deep learning techniques into healthcare has significantly revolutionized medical data analysis and interpretation. Table 1 provides a detailed summary of the deep learning methods employed in healthcare, emphasizing their application in medical image sequence analysis, predictive modeling, and addressing challenges in model interpretability. This section delves into applications in medical image sequence analysis, predictive modeling, and the challenges of model interpretability.

4.1 Medical Image Sequence Analysis

Deep learning has enhanced medical image sequence analysis by improving the interpretation of complex imaging data. The Monte Carlo Dropout U-Net exemplifies this by capturing segmentation

AI-generated, for reference only.

Category	Feature	Method
Predictive Modeling and Personalized Medicine	Reliability and Robustness Temporal and Data Integration	SSL-CAC[9] TFT[14]
Challenges and Solutions in Model Interpretability	Interpretability Enhancement	XAI-CIP[11], EBM[10], PSL[15]

Table 1: This table provides a comprehensive overview of various deep learning methods and their corresponding features applied in the domains of predictive modeling, personalized medicine, and model interpretability in healthcare. It highlights specific techniques such as SSL-CAC and TFT for improving reliability, robustness, and temporal data integration, as well as interpretability enhancement methods like XAI-CIP, EBM, and PSL. These methods illustrate the advancements in addressing challenges associated with model interpretability and predictive accuracy in clinical applications.

uncertainty through multiple outputs, crucial for clinical diagnostics and treatment planning [50]. Deep learning models, such as CNNs and LSTMs, are effective across various medical data types, including ECG data, showcasing their adaptability in sequential imaging tasks [12]. Frameworks like MONAI within PyTorch address challenges like abnormality detection, where deep learning has significantly improved performance. Bayesian neural networks and ensemble methods emphasize uncertainty quantification, enhancing prediction reliability in critical medical applications [51, 34, 33, 52, 53]. Explainable AI techniques, including LIME, SHAP, and CIU, improve model interpretability, aiding clinical decision-making and advancing diagnostic tools and personalized patient care [22, 12, 54].

4.2 Predictive Modeling and Personalized Medicine

Predictive modeling is central to advancing personalized medicine by enabling tailored interventions based on individual profiles. The Temporal Fusion Transformer model enhances predictive accuracy through temporal data integration for forecasting COVID-19 infections, highlighting feature sensitivity's role in predictive analytics [14]. This approach optimizes patient outcomes by refining treatment plans according to individual variability. Multitask learning models with attention mechanisms provide robust predictions while maintaining interpretability crucial for clinical applications. Integrating uncertainty quantification into predictive models enhances reliability; a semi-supervised learning approach for COVID-19 detection refines cough audio labeling consistency [9]. RAG-LLM based Clinical Decision Support Systems (CDSS) illustrate predictive modeling's potential in personalized medicine, improving medication error identification accuracy [13]. Recent advancements in interpretable machine learning emphasize transparency, enabling informed decisions that enhance patient care [21, 55, 56, 41, 57]. By leveraging deep learning and uncertainty quantification, predictive models facilitate accurate, individualized interventions, advancing personalized healthcare.

4.3 Challenges and Solutions in Model Interpretability

Method Name	**Interpretability Challenges**	**Innovative Solutions**	**Application Context**
EBM[10]	Black-box Models	Explainable Boosting Machines	Clinical Decision-making
XAI-CIP[11]	Black Boxes	Shap, Skoperules	Clinical Interventions
PSL[15]	Complexity And Opacity	Probabilistic Scoring Lists	Healthcare And Justice

Table 2: This table presents an overview of various interpretability methods used in deep learning models, highlighting the challenges they address, the innovative solutions they offer, and their specific application contexts. The methods include Explainable Boosting Machines, SHAP and Skoperules, and Probabilistic Scoring Lists, each contributing to enhanced transparency and understanding in clinical decision-making and other healthcare applications.

The interpretability of deep learning models in healthcare is challenged by their complexity and opacity, which can obscure decision-making and hinder clinical adoption. Ambiguities in target explanations can lead to misinterpretations, complicating trust in healthcare applications [10]. Existing algorithms often struggle with missing data, highlighting the need for improved interpretability methods. Innovative methodologies, such as Explainable Boosting Machines (EBMs), enhance transparency by clarifying feature contributions and decisions [10]. XAI methods like SHAP and counterfactual explanations further improve AI prediction interpretability, providing actionable insights for clinicians [11]. Probabilistic Scoring Lists (PSLs) offer nuanced understanding by producing probability distributions for decisions rather than deterministic outputs [15]. Monte Carlo sampling with dropout and variational inference estimate prediction uncertainty in deep learning models for medical imaging, guiding cautious decision-making [52]. These solutions reflect ongoing efforts to enhance model interpretability in healthcare, addressing the critical need for transparency in clinical decision-making. As deep learning continues to revolutionize healthcare through analyzing vast electronic health record data, developing interpretable models is essential for enabling professionals to trust and understand predictions generated by complex algorithms. This research examines interpretability methods and emphasizes their adaptation for healthcare, facilitating better AI integration in clinical practice [22, 12]. By addressing model complexity challenges and integrating interpretability techniques, these approaches support deep learning model adoption in clinical practice, enhancing patient care and outcomes. Table 2 provides a comprehensive summary of interpretability methods in deep learning, detailing the challenges they tackle, the innovative solutions they propose, and their application in healthcare and justice contexts.

Feature	Medical Image Sequence Analysis	Predictive Modeling and Personalized Medicine	Challenges and Solutions in Model Interpretability
Application Focus	Image Interpretation	Personalized Interventions	Model Transparency
Key Technique	Monte Carlo Dropout	Temporal Fusion Transformer	Explainable Boosting Machines
Interpretability Approach	Explainable AI	Multitask Learning	Shap, Counterfactuals

Table 3: This table provides a comparative analysis of deep learning methodologies applied in healthcare, focusing on medical image sequence analysis, predictive modeling and personalized medicine, and the challenges and solutions in model interpretability. It highlights the application focus, key techniques, and interpretability approaches for each domain, offering insights into the current landscape of deep learning applications in medical contexts.

5 Clinical Decision Support Systems (CDSS)

5.1 Integration of Deep Learning in CDSS

The integration of deep learning into Clinical Decision Support Systems (CDSS) represents a major advancement in healthcare, enhancing decision-making and patient outcomes through advanced computational techniques. For instance, the hybrid active learning model (HALM-DR) employs Bayesian inference and active learning to improve diabetic retinopathy classification, thus refining diagnostic precision [58]. Similarly, the ML-HIP model underscores machine learning's role in healthcare financial modeling by providing accurate cost predictions [8].

Deep learning applications in CDSS also include the Temporal Fusion Transformer model for COVID-19 forecasting, utilizing feature importance to derive actionable insights [14]. Semi-supervised learning (SSL) techniques that combine expert and user labels have enhanced cough audio classification, showcasing deep learning's potential to improve diagnostic accuracy [9].

Incorporating explainable boosting machines (EBMs) into CDSS offers interpretable predictions that improve healthcare outcomes, particularly in obstetrics, by providing transparent insights into model decisions [10]. This aligns with the broader trend towards interpretable AI in CDSS to foster transparency and trust among healthcare providers [11]. Additionally, probabilistic scoring lists (PSLs) can enhance CDSS by offering a flexible framework that accommodates uncertainty, thereby improving the reliability of clinical predictions [15].

The development of large language model (LLM)-based CDSS frameworks aims to enhance medication safety and reduce alert fatigue, ultimately improving patient care [13]. These advancements illustrate deep learning's potential to optimize clinical workflows and enhance patient outcomes through more efficient and accurate decision-making processes.

These examples highlight the profound impact of deep learning in CDSS, facilitating precise, interpretable, and effective clinical decision-making. By leveraging advanced algorithms and enhancing model transparency, CDSS can significantly improve healthcare delivery and patient outcomes, marking a pivotal advancement in medical informatics [12].

5.2 Implementation Challenges and Solutions

The implementation of Clinical Decision Support Systems (CDSS) in healthcare faces challenges that can hinder their adoption and effectiveness. A significant barrier is the variability in physicians' evaluations of AI-generated explanations, influenced by personal experiences and cultural contexts, complicating the assessment of explanation quality [36]. This variability necessitates standardized evaluation frameworks that accommodate diverse clinical perspectives, enhancing the reliability of AI-generated insights.

Trust issues and lack of transparency in AI decision-making processes contribute to low adoption rates of AI-CDSS, disrupting clinical workflows and deterring healthcare professionals from fully embracing these technologies [20]. To address these concerns, integrating explainable AI methods that provide clear, interpretable insights into model decisions is vital for building trust and facilitating smoother integration into clinical practices.

The integration of the Unified Theory of Acceptance and Use of Technology (UTAUT) with Task-Technology Fit (TTF) offers a robust framework for understanding user acceptance, guiding developers in creating effective CDSS that align with user needs and workflows [59]. By enhancing perceived usefulness and ease of use, developers can boost user acceptance and increase adoption rates among healthcare professionals.

Addressing disruptions in clinical workflows caused by AI-CDSS requires careful consideration of system design and implementation strategies. Engaging healthcare providers during development and testing phases enhances system integration into existing workflows. Research indicates that incorporating feedback from both physicians and patients, alongside validating tools across diverse clinical settings, leads to more effective and trusted AI-enhanced CDSS solutions, fostering wider adoption and improved patient outcomes [20, 60, 19]. Additionally, ongoing training and support for healthcare professionals can facilitate the effective use of CDSS, ensuring these systems complement rather than complicate clinical practice.

The solutions presented in the referenced works underscore the critical need to address implementation challenges by integrating technological innovation, user-centered design, and collaborative development approaches. This multifaceted strategy is essential for ensuring effective interaction with complex machine learning systems while addressing privacy concerns and enhancing trust in AI applications. By tailoring interpretability frameworks to the specific needs of various stakeholders, these solutions aim to bridge gaps in current practices and promote responsible deployment of AI technologies across diverse contexts, including healthcare and pervasive systems [55, 61, 62, 32, 63]. Overcoming these barriers allows CDSS to realize their full potential in enhancing clinical decision-making and improving patient outcomes.

5.3 Knowledge Integration and Predictive Modeling

Integrating knowledge and predictive modeling within Clinical Decision Support Systems (CDSS) is crucial for enhancing healthcare providers' decision-making capabilities. This integration synthesizes complex clinical data into actionable insights, improving patient outcomes and streamlining workflows. The Clinical Evidence Engine exemplifies advanced computational tools' potential to process complex queries and rapidly extract relevant clinical information, enabling clinicians to make informed decisions more efficiently [23].

Predictive modeling significantly enhances CDSS capabilities to forecast patient outcomes and facilitate proactive healthcare interventions. Utilizing advanced deep learning techniques and electronic health records (EHR), modern CDSS like CarePre provide timely and precise predictions of medical events, offering interpretable insights through interactive frameworks and visualizations. This approach supports clinicians in diagnosing and analyzing treatment outcomes while addressing interpretability challenges, promoting informed decision-making [21, 18, 24, 19, 20]. By leveraging machine learning algorithms, CDSS can analyze vast datasets to identify patterns and predict future health events, enabling personalized treatment plans and timely interventions, particularly in chronic disease management.

The use of privacy-preserving synthetic data, as demonstrated by SyntHIR, further emphasizes the importance of integrating predictive modeling with knowledge management in CDSS. SyntHIR generates synthetic health data that closely resembles real patient data, allowing for the testing and validation of CDSS tools without compromising patient privacy. This capability is essential for developing and deploying robust CDSS, enabling safe and ethical evaluation of predictive models across diverse healthcare settings [60].

Advancements in integrating knowledge and predictive modeling within CDSS highlight the necessity of tailoring interpretability to the diverse needs of stakeholders. These systems must deliver accurate predictions while providing insights that are comprehensible and relevant to various users. This approach underscores the importance of understanding stakeholder roles and their specific interpretability requirements, which is vital for fostering accountability and enhancing the usability of machine learning models in clinical contexts [55, 64, 65, 56]. By harnessing sophisticated algorithms and privacy-preserving technologies, CDSS can significantly enhance clinical decision-making, improve patient outcomes, and support personalized healthcare delivery.

6 Temporal Alignment and Uncertainty Quantification

6.1 Importance of Temporal Alignment in Healthcare

Temporal alignment is vital in healthcare data analysis, synchronizing diverse data modalities to ensure accurate clinical decision-making. This is particularly crucial for continuous physiological data, enabling real-time analysis and interventions [29]. Capturing temporal dependencies in time series data maintains the consistency and robustness of saliency explanations, essential for interpreting dynamic interactions [43]. The Temporal Fusion Transformer model exemplifies this by enhancing predictive accuracy in COVID-19 data analysis through capturing complex interactions over time, providing actionable public health insights [14]. In diagnostic imaging, temporal alignment is critical for processing whole-slide images in colorectal cancer diagnosis, ensuring reliable analysis of time-varying data [66]. Techniques like Deep Canonical Correlation Alignment (DCCA) manage long and noisy signals, preserving data integrity through automatic filtering and transformation [25].

Integrating uncertainty quantification methodologies further enhances temporal alignment's clinical relevance by presenting uncertainty meaningfully, improving segmentation accuracy in 3D contexts and reducing misinterpretation risks [67, 68]. Methods emulating emergency department team collaboration highlight temporal alignment's role in facilitating accurate decision-making [69]. Recent advancements underscore temporal alignment's importance in integrating diverse data sources, significantly enhancing the accuracy and reliability of clinical analyses. Techniques like subsequence alignment improve patient similarity assessments, enabling precise risk stratification for chronic diseases with heterogeneous progression. The application of deep learning techniques in health informatics leverages extensive electronic health record data to extract complex patterns, necessitating a focus on interpretability to ensure healthcare professionals trust these systems' predictions. Effective temporal alignment is thus vital for patient comparisons and harnessing advanced analytical tools to improve clinical decision-making [21, 22, 12, 70]. Addressing challenges related to temporal variability and misalignment leads to more accurate and informed decision-making, enhancing patient outcomes.

6.2 Techniques for Temporal Alignment

Temporal alignment is crucial in healthcare data analysis, integrating diverse data modalities to enhance predictive accuracy and clinical decision-making. Probabilistic methods like Probabilistic Neighbourhood Component Analysis (PNCA) utilize a probabilistic k-Nearest Neighbors (kNN) classifier to quantify uncertainties in predictions, mapping inputs into a latent space as distributions for accurate temporal alignment [33]. Dynamic time warping (DTW) aligns time-series data from different modalities, projecting non-parallel sequences into a shared latent space. This enhances temporal alignment, crucial for interpreting biomedical data, particularly in articulatory-to-acoustic speech synthesis and sensor signal alignment. DTW effectively aligns signals from various sources, even amid noise or desynchronized data, improving output quality and facilitating robust longitudinal patient data comparisons [30, 56, 70, 71, 25].

Bayesian neural networks combined with Transductive Dropout enhance uncertainty calibration by leveraging unlabelled target data, incorporating a regularization term to quantify model confidence. This approach improves the consistency and accuracy of temporal alignment in predictive models, leading to more reliable interpretations of underlying data patterns [72, 56, 57, 73, 65]. Monte Carlo sampling with dropout and variational inference supports uncertainty modeling in predictions, maintaining alignment accuracy and reliability in clinical settings.

In body composition analysis, training ResNet50 neural networks to predict both mean and variance of measurements aids in uncertainty quantification, promoting precise temporal alignment in healthcare applications. Advanced modeling techniques, such as a 3D U-Net enhanced by a 3D conditional Variational Autoencoder (VAE) with Normalizing Flows, generate diverse and anatomically plausible segmentations, addressing challenges like inter-rater variability and class imbalance in medical imaging. This approach enhances segmentation precision, crucial for clinical analysis, and emphasizes effective temporal alignment in applications like articulatory-to-acoustic speech synthesis and cardiac structure assessment, where accurate segmentation directly impacts diagnostics and treatment outcomes [44, 74, 30].

These methodologies underscore the critical role of temporal alignment in healthcare, facilitating multimodal data synchronization, thereby improving clinical analysis accuracy and reliability. Employing techniques like subsequence alignment and multi-view learning addresses challenges posed by heterogeneous disease progression and non-parallel data collection. This synchronization enhances patient similarity assessments and optimizes clinical decision-making processes, leading to more precise risk stratification and improved patient outcomes in complex medical scenarios [21, 30, 17, 70, 12]. By leveraging advanced computational techniques and tackling temporal variability challenges, healthcare systems can achieve more accurate and informed decision-making, ultimately enhancing patient outcomes.

6.3 Uncertainty Quantification in Clinical Decision-Making

Uncertainty quantification (UQ) is pivotal in enhancing the reliability of clinical decision-making by addressing inherent uncertainties in medical data. Integrating UQ techniques into predictive models allows for more informed decisions, especially in high-stakes clinical contexts where AI-generated recommendations' reliability is critical. Probabilistic Scoring Lists (PSLs) exemplify methods for quantifying uncertainty, facilitating dynamic decision-making based on evidence accumulation, essential for maintaining clinical assessment integrity [15].

In medical imaging, UQ is crucial for identifying low-quality segmentations, ensuring clinical decisions rely on reliable data and enhancing patient safety. Interpretable machine learning models elucidate risk factors and quantify uncertainty in clinical predictions, adhering to the Predictive, Descriptive, and Relevant (PDR) framework. This interpretability fosters data-driven, personalized decision-making, improving patient care quality. Furthermore, categorizing interpretability techniques into model-specific and model-agnostic approaches enables tailored explanations understood at individual and population levels, supporting effective treatment optimization in high-stakes healthcare environments [21, 41, 56]. This capability is particularly relevant in conditions like diabetic retinopathy, where understanding variable influences over time is critical for refining decision-making.

Investigating UQ techniques in deep learning for computer-aided diagnosis highlights their role in improving diagnostic outcomes and enhancing patient safety. By employing Bayesian approximation and Monte Carlo sampling, these models assess prediction uncertainties, indicating when they are uncertain about classifications. This capability is vital in medical applications, enabling healthcare professionals to identify ambiguous cases and make informed decisions, ultimately enhancing patient care [34, 52, 50, 35, 53]. Clear insights into model predictions and their uncertainties enhance AI-driven clinical decisions' transparency and trustworthiness.

The integration of UQ into predictive models is increasingly recognized for enhancing healthcare decision-making, particularly in identifying concession opportunities and refining pricing strategies. For example, machine learning models significantly outperformed traditional actuarial methods, achieving a 20

Recent advancements in artificial intelligence and deep learning have transformed clinical decision-making processes by enhancing Clinical Decision Support Systems (CDSS). These innovations enable complex pattern extraction from vast datasets, facilitating informed treatment choices. For instance, systems like Aifred Health assist in managing major depressive disorder, while the Clinical Evidence Engine provides clinicians with relevant scientific evidence from biomedical literature, bridging the gap between clinical questions and trial data. Collectively, these developments highlight UQ's critical role in improving clinical decisions' reliability and effectiveness, ultimately leading to better patient outcomes [23, 12, 20, 19]. By integrating advanced UQ techniques and addressing uncertainty estimation challenges, healthcare systems can significantly enhance predictive accuracy

AI-generated, for reference only.

and clinical utility, ensuring decisions are informed by a comprehensive understanding of potential risks and outcomes.

7 Reinforcement Learning in Medicine

7.1 Knowledge Transfer and Treatment Optimization

Reinforcement learning (RL) is revolutionizing healthcare by optimizing treatment strategies and facilitating knowledge transfer across various medical domains. By leveraging extensive datasets from electronic health records and deep learning techniques, RL enhances clinical decision-making, addressing challenges in data interpretability and collaboration among healthcare providers [21, 22, 41, 12, 75]. Its trial-and-error learning approach is particularly effective in dynamic clinical environments where static models may be inadequate.

A significant benefit of RL is its ability to transfer knowledge, adapting learned policies to new, related tasks. This capability enhances treatment efficiency by applying insights from one context to improve predictions and decisions in another. Advances in transfer learning, such as the RECaST framework, highlight the importance of uncertainty quantification, ensuring reliable predictions during policy adaptation across diverse populations and data models [55, 76]. This is particularly advantageous in scenarios with limited data, enabling the reuse of prior knowledge and reducing retraining needs. In chronic disease management, RL can personalize treatment plans by adapting to patients' evolving health statuses through real-time feedback.

RL's potential in personalized medicine is evident in its ability to tailor interventions to individual needs by integrating factors like genetic profiles, lifestyle, and environment. This adaptability fosters targeted care and improves interpretability in clinical decision-making, empowering healthcare professionals to make informed, data-driven choices that enhance patient outcomes [11, 41, 77]. By modeling the sequential decision-making process in medical treatments, RL identifies effective treatment pathways, improving outcomes and reducing costs.

The integration of RL with deep learning (DL) expands its applicability in medicine, enabling high-dimensional data analysis and complex decision-making. This synergy allows for extracting intricate patterns from large datasets, such as electronic health records and medical images, enhancing clinical decision support systems for diagnosis, prognosis, and treatment. DL's quantitative assessment capabilities in medical imaging drive advancements in fields like radiology and oncology, while research into interpretable DL models addresses transparency and trust in AI-driven healthcare [78, 12, 22]. This collaboration develops sophisticated models capturing complex patterns in patient data, informing precise and effective treatment strategies.

RL's application in healthcare promises significant advancements in knowledge transfer and treatment optimization. By harnessing its unique capabilities, RL supports the development of adaptive and personalized healthcare solutions, crucial for analyzing vast datasets and uncovering complex patterns. This approach aids healthcare professionals in making informed decisions through interpretable algorithms, significantly improving patient care and outcomes by tailoring interventions to individual needs [21, 12].

7.2 Context-Aware Systems and Personalized Medicine

The advancement of context-aware systems in healthcare, particularly through RL, marks a significant step forward in personalized medicine. These systems dynamically adapt to individual patient contexts, optimizing treatment strategies and improving outcomes. RL's advantages for clinical decision support systems (CDSS) lie in its ability to learn adaptively from real-time interactions with patients and their environments. This continuous refinement of treatment strategies based on immediate patient data and feedback is crucial for addressing complexities in mental health care, exemplified by AI models like Aifred Health, which enhance personalized care by integrating diverse data insights while overcoming interoperability and clinical workflow challenges [12, 24, 20].

Context-aware systems use RL to customize medical interventions by analyzing a wide range of patient-specific factors, including genetic profiles, lifestyle choices, and environmental conditions. This approach enhances treatment personalization and incorporates interpretable artificial intelligence (XAI) techniques, enabling healthcare professionals to understand the influence of specific factors—such as environmental noise and physiological responses—on patient outcomes. By leveraging these insights, clinicians can develop targeted intervention strategies catering to individual patients' unique needs and behaviors, ultimately improving the effectiveness of CDSS in addressing substance misuse and other health-related issues [55, 21, 11]. For instance, in chronic disease management, RL can dynamically adjust treatment regimens based on changes in patients' health statuses and responses to previous treatments.

Integrating RL with context-aware systems enhances the creation of adaptive models capable of predicting patient needs and optimizing resource allocation in real-time, thereby improving clinical decision-making processes. These advanced models leverage historical medical data and real-time inputs to generate actionable insights, ensuring prompt and effective responses to individual patient requirements. This not only streamlines resource management but also contributes to the overall efficacy of clinical interventions, as demonstrated by systems like CarePre, which employs deep learning for precise medical event predictions and supports interpretability through intuitive visualizations [55, 17, 18, 79]. Such capabilities are invaluable in complex clinical environments where timely and accurate decision-making is critical. By utilizing RL, these systems learn optimal policies that maximize patient outcomes while minimizing unnecessary interventions and costs.

The application of RL in context-aware systems significantly enhances healthcare delivery processes by enabling adaptive decision-making that accounts for the dynamic nature of patient needs and environmental conditions, ultimately leading to improved patient outcomes and more efficient resource utilization [21, 55, 22, 41, 12]. By analyzing patient data patterns, RL identifies inefficiencies in treatment pathways and suggests improvements, thereby enhancing overall care quality. Additionally, employing RL in developing context-aware systems supports preventive measures by identifying potential health risks and recommending proactive interventions.

The development of context-aware systems using RL significantly enhances personalized medicine by providing tailored, efficient, and adaptive healthcare solutions. By integrating RL with advanced Clinical Decision Support Systems (CDSS), such as Retrieval Augmented Generation (RAG)-Large Language Models (LLMs), these systems can substantially improve patient outcomes by accurately identifying medication errors, optimizing resource utilization through predictive analytics, and facilitating high-quality, individualized care tailored to each patient's medical history and needs [13, 36, 18].

7.3 Risk-Aware Decision-Making Frameworks

Risk-aware decision-making frameworks in healthcare leverage RL to enhance the reliability and safety of clinical decisions by systematically incorporating risk assessments into the decision-making process. These frameworks are crucial in high-stakes medical environments, where decision consequences can significantly impact patient outcomes. By integrating risk-awareness into RL models, these systems effectively address uncertainties present in medical data, facilitating well-informed and strategically cautious decision-making. This approach improves models' capabilities to evaluate potential outcomes and prioritize patient safety, particularly in complex healthcare environments characterized by high stakes and ambiguous data [13, 15, 40].

A key aspect of risk-aware frameworks is their ability to quantify and manage uncertainty, essential for making robust decisions in dynamic clinical settings. The development of safe, anytime-valid concentration bounds and frameworks for risk-aware contextual bandits exemplifies advancements in statistical methods that support sequential decision-making in healthcare [40]. These methods provide a foundation for RL algorithms to operate effectively in uncertain and dynamic environments, ensuring that patient care remains adaptive and reliable.

Integrating multi-level, multi-fidelity (MLMF) frameworks with RL further enhances decision-making by incorporating additional reduced-order models and extending applications to a broader range of medical scenarios [39]. Such enhancements enable more accurate and efficient treatment strategies, ultimately improving patient outcomes. By leveraging advanced computational techniques, risk-aware frameworks can identify optimal treatment pathways while accounting for potential risks and uncertainties.

Moreover, incorporating uncertainty quantification into RL models is vital for ensuring the reliability of predictions and decisions. Techniques such as Bayesian inference and probabilistic modeling enhance the interpretability of machine learning outputs in healthcare by quantifying confidence

AI-generated, for reference only.

levels associated with model predictions. This allows healthcare providers to better assess risks linked to treatment options, elucidating relationships between input features and outcomes, identifying significant influencing factors, and clarifying uncertainty surrounding diagnostic decisions. By employing these approaches, clinicians can make more informed choices, ultimately improving patient care and treatment efficacy [53, 21, 80]. This is particularly beneficial in personalized medicine, where treatment plans must be tailored to individual patient profiles while considering potential risks.

Recent advancements in interpretable machine learning, deep learning applications, and stakeholder engagement frameworks highlight the significant impact of risk-aware decision-making frameworks in healthcare, emphasizing their ability to enhance transparency, improve clinical outcomes, and facilitate informed decision-making among healthcare professionals [55, 21, 22, 41, 12]. By integrating RL with sophisticated risk assessment methodologies, these frameworks enhance the safety and efficacy of clinical decisions, paving the way for more adaptive and patient-centered healthcare solutions.

8 Interpretable Machine Learning in Healthcare

The integration of interpretable machine learning (IML) in healthcare transcends mere technical advancement, addressing the vital intersection between technology and clinical practice. The importance of interpretability in healthcare AI lies in clarifying the rationale behind AI-driven decisions, fostering trust among healthcare professionals, and ensuring alignment with established medical reasoning and ethical standards. This section explores the implications of interpretability for clinical adoption and decision-making processes.

To further illustrate these concepts, Figure 4 presents a hierarchical structure of key ideas in interpretable machine learning within healthcare. This figure highlights the significance, frameworks, applications, and challenges associated with enhancing interpretability in AI models for clinical settings, thereby providing a visual representation that complements the discussion of interpretability's critical role in fostering effective healthcare solutions.

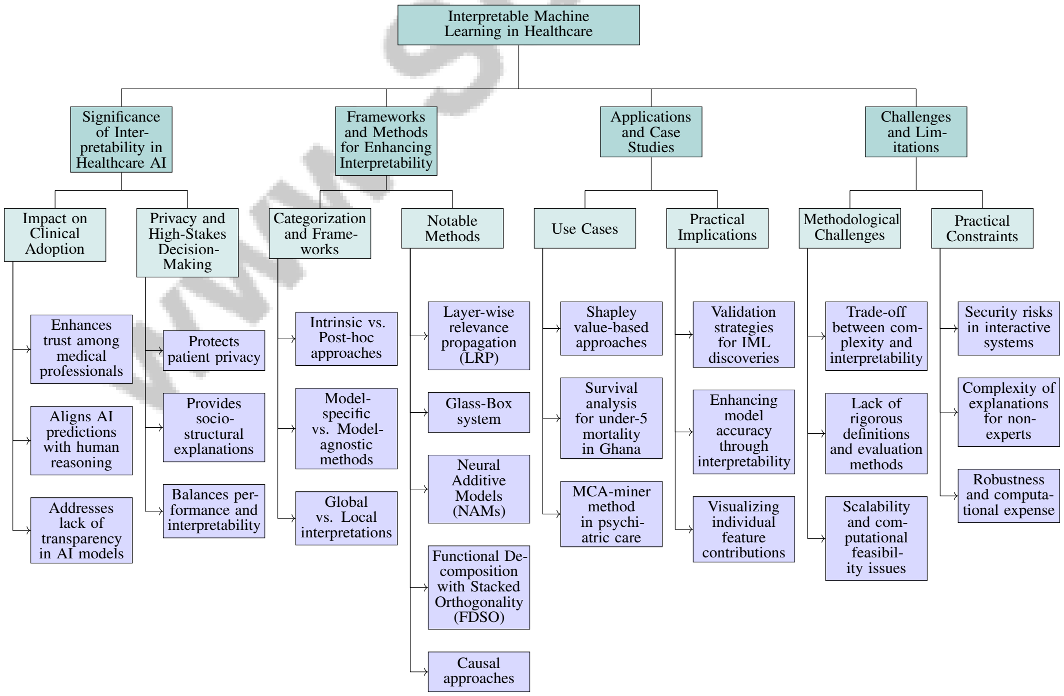

Interpretable Machine Learning in Healthcare
Significance of Interpretability in Healthcare AI
Frameworks and Methods for Enhancing Interpretability
Applications and Case Studies
Challenges and Limitations
Impact on Clinical Adoption
Privacy and High-Stakes Decision-Making
Categorization and Frameworks
Notable Methods
Use Cases
Practical Implications
Methodological Challenges
Practical Constraints
Enhances trust among medical professionals
Aligns AI predictions with human reasoning
Addresses lack of transparency in AI models
Protects patient privacy
Provides socio-structural explanations
Balances performance and interpretability
Intrinsic vs. Post-hoc approaches
Model-specific vs. Model-agnostic methods
Global vs. Local interpretations
Layer-wise relevance propagation (LRP)
Glass-Box system
Neural Additive Models (NAMs)
Functional Decomposition with Stacked Orthogonality (FDSO)
Causal approaches
Shapley value-based approaches
Survival analysis for under-5 mortality in Ghana
MCA-miner method in psychiatric care
Validation strategies for IML discoveries
Enhancing model accuracy through interpretability
Visualizing individual feature contributions
Trade-off between complexity and interpretability
Lack of rigorous definitions and evaluation methods
Scalability and computational feasibility issues
Security risks in interactive systems
Complexity of explanations for non-experts
Robustness and computational expense

Figure 4: This figure illustrates the hierarchical structure of key concepts in interpretable machine learning within healthcare, highlighting the significance, frameworks, applications, and challenges associated with enhancing interpretability in AI models for clinical settings.

8.1 Significance of Interpretability in Healthcare AI

Interpretability is crucial in healthcare AI due to its significant impact on clinical adoption and decision-making. Understanding AI model decision processes enhances trust among medical professionals and aligns AI-generated predictions with human medical reasoning [81, 82]. The opaque nature of many AI models, especially deep learning models, impedes their acceptance in healthcare due to a lack of transparency [22]. Interpretable machine learning offers a solution by emphasizing transparency and understanding in model predictions [57]. This transparency not only builds trust but also influences the adoption of AI technologies in clinical settings by providing clear insights into model reasoning [82].

Moreover, interpretability protects patient privacy by informing users of potential privacy risks without revealing sensitive information [62]. Socio-structural explanations from interpretable models are essential in high-stakes situations, offering insights necessary for informed decision-making [83]. Balancing performance and interpretability is critical, as interpretable models must maintain competitive accuracy while providing transparency, particularly in high-stakes environments where understanding model predictions can significantly influence clinical decisions [81].

8.2 Frameworks and Methods for Enhancing Interpretability

Enhancing the interpretability of machine learning models is essential in healthcare, where understanding model predictions is critical for clinical adoption. Interpretability methods are categorized into intrinsic and post-hoc approaches, further divided into model-specific and model-agnostic, as well as global versus local interpretations [84]. This classification aids in selecting suitable interpretability techniques for specific healthcare applications.

A notable framework is the taxonomy by Doshi-Velez and Kim, categorizing interpretability into application-grounded, human-grounded, and functionally-grounded approaches, emphasizing the need for rigorous evaluation [85]. Layer-wise relevance propagation (LRP) illustrates network decisions, enhancing trust among medical experts by highlighting the contribution of each input feature to the final prediction [26].

The Glass-Box system proposed by Sokol et al. allows users to interactively modify counterfactual explanations and pose follow-up questions, enhancing interpretability [63]. Neural Additive Models (NAMs) offer a flexible and interpretable approach by decomposing predictions into additive components corresponding to individual features [82]. The Functional Decomposition with Stacked Orthogonality (FDSO) method simplifies black-box prediction functions into clearer subfunctions, enhancing interpretability by elucidating feature influences [81].

Causal approaches in IML, emphasized by Xu et al., focus on causality rather than correlation, providing more meaningful insights into model predictions [73]. The Hybrid Predictive Model (HPM) integrates interpretable models with black-box models, balancing interpretability and predictive performance [57]. Collectively, these frameworks and methods highlight the importance of enhancing interpretability in machine learning models, particularly in healthcare, where transparency and trust are paramount.

8.3 Applications and Case Studies

Interpretable machine learning (IML) has been effectively applied across various healthcare domains, significantly enhancing decision-making and patient outcomes. Shapley value-based approaches improve the interpretability of concept-based machine learning models, enabling better understanding and validation of classification outcomes, crucial in clinical settings [86].

Recent studies underscore the necessity of validation for IML discoveries, emphasizing practical validation strategies and the importance of statistical theory in supporting reliable findings [65]. In high-stakes scenarios, interpretable models ensure transparency and can enhance model accuracy through better troubleshooting [84].

Causal methods in IML offer deeper insights into model behavior by focusing on causality, providing valuable explanations in healthcare [73]. Visualizing individual feature contributions further enhances trust and understanding in model predictions, supporting high-stakes decision-making processes [82].

Practical applications of IML, such as survival analysis for under-5 mortality in Ghana, demonstrate how tailored techniques can yield actionable insights into mortality risk factors [87]. The MCA-miner method in psychiatric care matches existing prediction accuracy while reducing computation time, showcasing its practical application potential [45].

The significance of balancing model performance and interpretability is emphasized in healthcare applications, where selecting appropriate methods is crucial for achieving effective outcomes [22]. Providing clear visualizations of feature effects enhances the interpretability of machine learning models, facilitating better clinical decision-making [81].

8.4 Challenges and Limitations

Implementing interpretable machine learning (IML) in healthcare faces several challenges, primarily due to methodological limitations and practical constraints. A significant challenge is the trade-off between model complexity and interpretability, which can lead to misinterpretations of model outputs. In healthcare, where complex models capture intricate patterns in biomedical data, simpler, more interpretable models are often preferred [22]. This complexity-interpretability trade-off is further complicated by the limited applicability of certain methods to low-dimensional feature sets, as high-dimensional cases introduce computational challenges [81].

The lack of rigorous definitions and evaluation methods for interpretability remains a significant hurdle. Many current approaches are informal, complicating meaningful comparisons between methodologies [85]. This absence of standard definitions hinders the development of IML systems and makes evaluating their effectiveness in clinical settings challenging [65].

Practical constraints also challenge the scalability of IML applications. For instance, the computational feasibility of calculating Shapley values can be limited in contexts with numerous attributes [86]. Additionally, the potential for malicious users to exploit interactive systems like the Glass-Box poses security risks [63]. The complexity of explanations generated by some methods may overwhelm non-expert users, complicating understanding of privacy implications [62].

Moreover, current IML methods may lack robustness and can be computationally expensive, limiting their practical application in clinical settings. They may not always align with clinical needs or provide consistent explanations, complicating their integration into healthcare workflows [22]. The hybrid predictive model, while balancing interpretability and performance, may not perform optimally in all data regions, particularly if the interpretable model fails to capture complex relationships adequately [57].

Producing rigorous socio-structural explanations requires interdisciplinary expertise, which can be challenging to achieve [83]. Additionally, limited error rate control and inflexible feature resolution of current IML methods can hinder effective implementation, highlighting the need for ongoing research and collaboration between researchers and practitioners [88].

These challenges emphasize the need for continued development and refinement of IML techniques to enhance their accuracy and understandability, thereby improving their utility in clinical decision-making. Addressing interpretability challenges in healthcare machine learning will facilitate personalized, data-driven decision-making, ultimately enhancing care quality by enabling clearer explanations of model predictions. Exploring both model-specific and model-agnostic interpretability techniques will be crucial for ensuring that machine learning algorithms are understandable and actionable in high-stakes healthcare environments [21, 41].

9 Conclusion

9.1 Future Directions and Research Opportunities

The advancement of multimodal temporal data and deep learning in healthcare offers substantial potential for innovation, promising to significantly improve clinical outcomes and the field of medical informatics. A key area for future exploration is the enhancement of Neural Additive Models (NAMs) by incorporating higher-order feature interactions and optimizing activation functions, which would extend their utility in analyzing complex healthcare datasets. This could provide deeper insights into patient-specific variables.

Strengthening the robustness of interpretability methods is crucial for their alignment with clinical practices, thereby improving user interaction and facilitating the integration of AI tools into healthcare environments. Such improvements are vital for increasing the practical utility and acceptance of AI systems among healthcare professionals.

Additionally, increasing sample sizes and diversifying population groups are imperative for refining Explainable AI (XAI) techniques in clinical settings. This ensures that AI models are representative and inclusive, enhancing their applicability across diverse demographic groups.

Further research should focus on refining sensitivity analysis methods and applying them to various public health challenges. This could lead to more accurate modeling of public health data, supporting better-informed policy decisions and interventions.

The integration of Large Language Models (LLMs) into clinical settings and their potential for real-time application present exciting research opportunities. This could transform clinical decision support systems by providing timely, contextually relevant insights that enhance patient care.

Improving learning algorithms for Probabilistic Scoring Lists (PSLs) and exploring the integration of feature costs remain promising areas for innovation within decision-making frameworks. These advancements are expected to refine decision-making tools, ensuring their precision and efficiency.

Finally, incorporating socio-structural understanding throughout the machine learning lifecycle is essential for developing AI systems that are both technically robust and socially responsible. This comprehensive approach is critical for the ethical and equitable development and deployment of AI technologies, fostering trust and acceptance among users.

These future directions underscore the importance of continuous innovation and exploration in multimodal temporal data and deep learning within healthcare. By addressing current challenges and leveraging advanced computational techniques, significant enhancements in patient care and medical informatics can be achieved.

References

[1] Heming Yao, Harm Derksen, Jessica R. Golbus, Justin Zhang, Keith D. Aaronson, Jonathan Gryak, and Kayvan Najarian. A novel tropical geometry-based interpretable machine learning method: Application in prognosis of advanced heart failure, 2021.

[2] Julien Grand-Clément, You Hui Goh, Carri Chan, Vineet Goyal, and Elizabeth Chuang. Interpretable machine learning for resource allocation with application to ventilator triage, 2024.

[3] Hossein Simchi and Samira Tajik. The pros and cons of using machine learning and interpretable machine learning methods in psychiatry detection applications, specifically depression disorder: A brief review, 2023.

[4] Yanwen Li, Luyang Luo, Huangjing Lin, Hao Chen, and Pheng-Ann Heng. Dual-consistency semi-supervised learning with uncertainty quantification for covid-19 lesion segmentation from ct images, 2021.

[5] Niels Van Berkel, Maura Bellio, Mikael B Skov, and Ann Blandford. Measurements, algorithms, and presentations of reality: Framing interactions with ai-enabled decision support. *ACM Transactions on Computer-Human Interaction*, 30(2):1–33, 2023.

[6] Margherita Rosnati and Vincent Fortuin. Mgp-atttcn: An interpretable machine learning model for the prediction of sepsis, 2021.

[7] Xiaoli Liu, Pan Hu, Zhi Mao, Po-Chih Kuo, Peiyao Li, Chao Liu, Jie Hu, Deyu Li, Desen Cao, Roger G. Mark, Leo Anthony Celi, Zhengbo Zhang, and Feihu Zhou. Interpretable machine learning model for early prediction of mortality in elderly patients with multiple organ dysfunction syndrome (mods): a multicenter retrospective study and cross validation, 2020.

[8] Rohun Kshirsagar, Li-Yen Hsu, Vatshank Chaturvedi, Charles H. Greenberg, Matthew McClelland, Anushadevi Mohan, Wideet Shende, Nicolas P. Tilmans, Renzo Frigato, Min Guo, Ankit Chheda, Meredith Trotter, Shonket Ray, Arnold Lee, and Miguel Alvarado. Accurate and interpretable machine learning for transparent pricing of health insurance plans, 2021.

[9] Lara Orlandic, Tomas Teijeiro, and David Atienza. A semi-supervised algorithm for improving the consistency of crowdsourced datasets: The covid-19 case study on respiratory disorder classification, 2022.

[10] Tomas M. Bosschieter, Zifei Xu, Hui Lan, Benjamin J. Lengerich, Harsha Nori, Kristin Sitcov, Vivienne Souter, and Rich Caruana. Using interpretable machine learning to predict maternal and fetal outcomes, 2022.

[11] Tongze Zhang, Tammy Chung, Anind Dey, and Sang Won Bae. Exploring algorithmic explainability: Generating explainable ai insights for personalized clinical decision support focused on cannabis intoxication in young adults, 2024.

[12] Farzan Shenavarmasouleh, Farid Ghareh Mohammadi, Khaled M. Rasheed, and Hamid R. Arabnia. Deep learning in healthcare: An in-depth analysis, 2023.

[13] Jasmine Chiat Ling Ong, Liyuan Jin, Kabilan Elangovan, Gilbert Yong San Lim, Daniel Yan Zheng Lim, Gerald Gui Ren Sng, Yuhe Ke, Joshua Yi Min Tung, Ryan Jian Zhong, Christopher Ming Yao Koh, Keane Zhi Hao Lee, Xiang Chen, Jack Kian Chng, Aung Than, Ken Junyang Goh, and Daniel Shu Wei Ting. Development and testing of a novel large language model-based clinical decision support systems for medication safety in 12 clinical specialties, 2024.

[14] Md Khairul Islam, Di Zhu, Yingzheng Liu, Andrej Erkelens, Nick Daniello, and Judy Fox. Interpreting county level covid-19 infection and feature sensitivity using deep learning time series models, 2022.

[15] Jonas Hanselle, Stefan Heid, Johannes Fürnkranz, and Eyke Hüllermeier. Probabilistic scoring lists for interpretable machine learning, 2024.

[16] Fadhil G. Al-Amran, Salman Rawaf, and Maitham G. Yousif. Early detection of post-covid-19 fatigue syndrome using deep learning models, 2023.

[17] Elisa Warner, Joonsang Lee, William Hsu, Tanveer Syeda-Mahmood, Charles E Kahn Jr, Olivier Gevaert, and Arvind Rao. Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects. *International Journal of Computer Vision*, 132(9):3753–3769, 2024.

[18] Zhuochen Jin, Jingshun Yang, Shuyuan Cui, David Gotz, Jimeng Sun, and Nan Cao. Carepre: An intelligent clinical decision assistance system, 2018.

[19] Kamran Farooq, Bisma S Khan, Muaz A Niazi, Stephen J Leslie, and Amir Hussain. Clinical decision support systems: A visual survey, 2017.

[20] Grace Golden, Christina Popescu, Sonia Israel, Kelly Perlman, Caitrin Armstrong, Robert Fratila, Myriam Tanguay-Sela, and David Benrimoh. Applying artificial intelligence to clinical decision support in mental health: What have we learned?, 2023.

[21] Daniel Sierra-Botero, Ana Molina-Taborda, Mario S. Valdés-Tresanco, Alejandro Hernández-Arango, Leonardo Espinosa-Leal, Alexander Karpenko, and Olga Lopez-Acevedo. Selecting interpretability techniques for healthcare machine learning models, 2024.

[22] Di Jin, Elena Sergeeva, Wei-Hung Weng, Geeticka Chauhan, and Peter Szolovits. Explainable deep learning in healthcare: A methodological survey from an attribution view, 2021.

[23] Bojian Hou, Hao Zhang, Gur Ladizhinsky, Gur Ladizhinsky, Stephen Yang, Volodymyr Kuleshov, Fei Wang, and Qian Yang. Clinical evidence engine: Proof-of-concept for a clinical-domain-agnostic decision support infrastructure, 2021.

[24] Cécile Trottet, Thijs Vogels, Martin Jaggi, and Mary-Anne Hartley. Modular clinical decision support networks (modn) – updatable, interpretable, and portable predictions for evolving clinical environments, 2022.

[25] Narayan Schütz, Angela Botros, Michael Single, Aileen C. Naef, Philipp Buluschek, and Tobias Nef. Deep canonical correlation alignment for sensor signals, 2021.

[26] Irina Grigorescu, Lucilio Cordero-Grande, A David Edwards, Jo Hajnal, Marc Modat, and Maria Deprez. Interpretable convolutional neural networks for preterm birth classification, 2019.

[27] Shourya Verma. Development of interpretable machine learning models to detect arrhythmia based on ecg data, 2022.

[28] Seyed Ehsan Saffari, Yilin Ning, Xie Feng, Bibhas Chakraborty, Victor Volovici, Roger Vaughan, Marcus Eng Hock Ong, and Nan Liu. Autoscore-ordinal: An interpretable machine learning framework for generating scoring models for ordinal outcomes, 2022.

[29] Olivia Pifer Alge. *Dynamic Machine Learning using Signal Processing and Tensor-Based Methods to Predict Clinical Outcomes*. PhD thesis, 2024.

[30] Jose A. Gonzalez-Lopez, Miriam Gonzalez-Atienza, Alejandro Gomez-Alanis, Jose L. Perez-Cordoba, and Phil D. Green. Multi-view temporal alignment for non-parallel articulatory-to-acoustic speech synthesis, 2020.

[31] Cornelius Fritz, Emilio Dorigatti, and David Rügamer. Combining graph neural networks and spatio-temporal disease models to predict covid-19 cases in germany, 2021.

[32] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. "brilliant ai doctor" in rural china: Tensions and challenges in ai-powered cdss deployment, 2021.

[33] Ankur Mallick, Chaitanya Dwivedi, Bhavya Kailkhura, Gauri Joshi, and T. Yong-Jin Han. Probabilistic neighbourhood component analysis: Sample efficient uncertainty estimation in deep learning, 2020.

[34] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges, 2021.

[35] Xiaoyang Huang, Jiancheng Yang, Linguo Li, Haoran Deng, Bingbing Ni, and Yi Xu. Evaluating and boosting uncertainty quantification in classification, 2019.

[36] D. Umerenkov, G. Zubkova, and A. Nesterov. Deciphering diagnoses: How large language models explanations influence clinical decision making, 2023.

[37] Prithviraj Manivannan, Ivo Pascal de Jong, Matias Valdenegro-Toro, and Andreea Ioana Sburlea. Uncertainty quantification for cross-subject motor imagery classification, 2024.

[38] Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. Pac prediction sets for meta-learning, 2022.

[39] Casey M. Fleeter, Gianluca Geraci, Daniele E. Schiavazzi, Andrew M. Kahn, and Alison L. Marsden. Multilevel and multifidelity uncertainty quantification for cardiovascular hemodynamics, 2020.

[40] Patrick Saux. Mathematics of statistical sequential decision-making: concentration, risk-awareness and modelling in stochastic bandits, with applications to bariatric surgery, 2024.

[41] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning based prediction models in healthcare, 2020.

[42] Jef Jonkers, Jarne Verhaeghe, Glenn Van Wallendael, Luc Duchateau, and Sofie Van Hoecke. Conformal convolution and monte carlo meta-learners for predictive inference of individual treatment effects, 2024.

[43] Chiara Balestra, Bin Li, and Emmanuel Müller. On the consistency and robustness of saliency explanations for time series classification, 2023.

[44] Soumick Chatterjee, Franziska Gaidzik, Alessandro Sciarra, Hendrik Mattern, Gábor Janiga, Oliver Speck, Andreas Nürnberger, and Sahani Pathiraja. Pulaski: Learning inter-rater variability using statistical distances to improve probabilistic segmentation, 2023.

[45] Qingzhu Gao, Humberto Gonzalez, and Parvez Ahammad. Mca-based rule mining enables interpretable inference in clinical psychiatry, 2018.

[46] Sven Kruschel, Nico Hambauer, Sven Weinzierl, Sandra Zilker, Mathias Kraus, and Patrick Zschech. Challenging the performance-interpretability trade-off: An evaluation of interpretable machine learning models, 2024.

[47] Anna Karanika, Panagiotis Oikonomou, Kostas Kolomvatsos, and Christos Anagnostopoulos. On the use of interpretable machine learning for the management of data quality, 2020.

[48] Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Interpretable machine learning: Moving from mythos to diagnostics, 2021.

[49] SangHyuk Kim, Edward Gaibor, Brian Matejek, and Daniel Haehn. Melanoma detection with uncertainty quantification, 2024.

[50] Katharina Hoebel, Ken Chang, Jay Patel, Praveer Singh, and Jayashree Kalpathy-Cramer. Give me (un)certainty – an exploration of parameters that affect segmentation uncertainty, 2019.

[51] Alex J. Chan, Ahmed M. Alaa, Zhaozhi Qian, and Mihaela van der Schaar. Unlabelled data improves bayesian uncertainty calibration under covariate shift, 2020.

[52] Max-Heinrich Laves, Sontje Ihler, and Tobias Ortmaier. Uncertainty quantification in computer-aided diagnosis: Make your model say "i don't know" for ambiguous cases, 2019.

[53] Masoumeh Javanbakhat, Md Tasnimul Hasan, and Cristoph Lippert. Assessing uncertainty estimation methods for 3d image segmentation under distribution shifts, 2024.

[54] S. Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises, 2021.

[55] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs, 2021.

[56] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications, 2019.

[57] Tong Wang and Qihang Lin. Hybrid predictive model: When an interpretable model collaborates with a black-box model, 2019.

[58] Muhammad Ahtazaz Ahsan, Adnan Qayyum, Junaid Qadir, and Adeel Razi. An active learning method for diabetic retinopathy classification with uncertainty quantification, 2020.

[59] Soliman Aljarboa and Shah J. Miah. An integration of utaut and task-technology fit frameworks for assessing the acceptance of clinical decision support systems in the context of a developing country, 2020.

[60] Pavitra Chauhan, Mohsen Gamal Saad Askar, Bjørn Fjukstad, Lars Ailo Bongo, and Edvard Pedersen. Interoperable synthetic health data with synthir to enable the development of cdss tools, 2023.

[61] Ghita Ghislat, Saiveth Hernandez-Hernandez, Chayanit Piyawajanusorn, and Pedro J. Ballester. Data-centric challenges with the application and adoption of artificial intelligence for drug discovery, 2024.

[62] Benjamin Baron and Mirco Musolesi. Interpretable machine learning for privacy-preserving pervasive systems, 2019.

[63] Kacper Sokol and Peter Flach. One explanation does not fit all: The promise of interactive explanations for machine learning transparency, 2020.

[64] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to whom? a role-based model for analyzing interpretable machine learning systems, 2018.

[65] Genevera I. Allen, Luqin Gan, and Lili Zheng. Interpretable machine learning for discovery: Statistical challenges & opportunities, 2023.

[66] Pedro C. Neto, Diana Montezuma, Sara P. Oliveira, Domingos Oliveira, João Fraga, Ana Monteiro, João Monteiro, Liliana Ribeiro, Sofia Gonçalves, Stefan Reinhard, Inti Zlobec, Isabel M. Pinto, and Jaime S. Cardoso. An interpretable machine learning system for colorectal cancer diagnosis from pathology slides, 2024.

[67] Christiaan G. A. Viviers, Amaan M. M. Valiuddin, Peter H. N. de With, and Fons van der Sommen. Probabilistic 3d segmentation for aleatoric uncertainty quantification in full 3d medical data, 2023.

[68] Kumud Lakara and Matias Valdenegro-Toro. Disentangled uncertainty and out of distribution detection in medical generative models, 2022.

[69] Seungjun Han and Wongyung Choi. Development of a large language model-based multi-agent clinical decision support system for korean triage and acuity scale (ktas)-based triage and treatment planning in emergency departments, 2024.

[70] Dev Goyal, Zeeshan Syed, and Jenna Wiens. Clinically meaningful comparisons over time: An approach to measuring patient similarity based on subsequence alignment, 2018.

[71] Huy Phan, Kaare Mikkelsen, Oliver Y. Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification, 2022.

[72] Chao Min, Guoyong Liao, Guoquan Wen, Yingjun Li, and Xing Guo. Ensemble interpretation: A unified method for interpretable machine learning, 2023.

[73] Guandong Xu, Tri Dung Duong, Qian Li, Shaowu Liu, and Xianzhi Wang. Causality learning: A new perspective for interpretable machine learning, 2021.

[74] Xiaofeng Liu, Fangxu Xing, Hanna K. Gaggin, Weichung Wang, C. C. Jay Kuo, Georges El Fakhri, and Jonghye Woo. Segmentation of cardiac structures via successive subspace learning with saab transform from cine mri, 2021.

[75] Maarten G. Poirot, Praneeth Vepakomma, Ken Chang, Jayashree Kalpathy-Cramer, Rajiv Gupta, and Ramesh Raskar. Split learning for collaborative deep learning in healthcare, 2019.

[76] Jimmy Hickey, Jonathan P. Williams, and Emily C. Hector. Transfer learning with uncertainty quantification: Random effect calibration of source to target (recast), 2022.

[77] Alek Fröhlich, Thiago Ramos, Gustavo Cabello, Isabela Buzatto, Rafael Izbicki, and Daniel Tiezzi. Personalizedus: Interpretable breast cancer risk assessment with local coverage uncertainty quantification, 2024.

[78] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510, 2018.

[79] Francisco de Arriba-Pérez and Silvia García-Méndez. Leveraging large language models through natural language processing to provide interpretable machine learning predictions of mental deterioration in real time, 2024.

[80] Catarina Moreira, Yu-Liang Chou, Mythreyi Velmurugan, Chun Ouyang, Renuka Sindhgatta, and Peter Bruza. An interpretable probabilistic approach for demystifying black-box predictive models, 2020.

[81] David Köhler, David Rügamer, and Matthias Schmid. Achieving interpretable machine learning by functional decomposition of black-box models into explainable predictor effects, 2024.

[82] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey Hinton. Neural additive models: Interpretable machine learning with neural nets, 2021.

[83] Andrew Smart and Atoosa Kasirzadeh. Beyond model interpretability: Socio-structural explanations in machine learning, 2024.

[84] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges, 2021.

[85] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[86] Léonard Kwuida and Dmitry I. Ignatov. On interpretability and similarity in concept-based machine learning, 2021.

[87] Sophie Hanna Langbein, Mateusz Krzyziński, Mikołaj Spytek, Hubert Baniecki, Przemysław Biecek, and Marvin N. Wright. Interpretable machine learning for survival analysis, 2024.

[88] David S. Watson. Interpretable machine learning for genomics, 2021.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.