

Information Theory for Medical Data Fusion

Magdalena Punceva¹, Ninoslav Marina²

MIT Univerzitet SkopjeSkopje, Macedonia¹, University of Information Science and Technology Skopje, Macedonia²

Abstract: Data fusion is the process of integrating multiple heterogeneous data sources to produce more accurate, comprehensive, and useful information than any single source alone. For the medical data fusion, this information may come from imaging, genomic data, clinical records, and physiological signals. It has emerged as a cornerstone of modern precision medicine. Information theory provides a rigorous mathematical framework for quantifying uncertainty, measuring information gain, and can thus be used to optimize fusion strategies across diverse clinical contexts. This paper presents a review of information-theoretic approaches suitable for medical data fusion, covering: fundamental concepts (entropy, mutual information, Kullback-Leibler divergence, rate-distortion), theoretical frameworks (information bottleneck, transfer entropy, partial information decomposition), and their application across major clinical domains. We synthesize recent advances in fusion-based architectures, discuss the critical challenges of uncertainty quantification, and provide practical guidelines for implementing information-theoretic fusion in clinical settings. Through a systematic analysis, we identify key challenges, including parameter sensitivity, missing modalities, and clinical interpretability, to outline promising directions for future research. This review aims to provide clinicians, researchers, and developers with a comprehensive understanding of how information theory can transform medical data for improved diagnostic accuracy, prognostic precision, and personalized patient care.

KEYWORDS—INFORMATION THEORY, DATA FUSION, ENTROPY, MUTUAL INFORMATION, RATE-DISTORTION THEORY, INFORMATION BOTTLENECK, TRANSFER ENTROPY, PARTIAL INFORMATION DECOMPOSITION, DEEP LEARNING.

I. INTRODUCTION

Modern healthcare generates unprecedented volumes of heterogeneous data: high-resolution medical images (CT, MRI, PET), molecular profiles (genomics, proteomics, metabolomics), continuous physiological signals (ECG, EEG, EMG), and structured clinical data from electronic health records. Each data type provides a partial view of patient health, and thus their integration by medical data fusion holds the promise of a holistic understanding that can transform diagnosis, prognosis, and treatment selection [Duan et al., 2024, Salvi et al., 2024]. However, the fusion of medical data presents some fundamental challenges: heterogeneity (data types span continuous signals, categorical variables, images, and unstructured text), high dimensionality (genomic and imaging data often contain plenty of features), missing modalities (real-world clinical data frequently have incomplete measurements), temporal dynamics (patient status evolves over time, requiring longitudinal integration, interpretability (clinicians require transparent, explainable fusion results).

Information theory is the mathematical study of the quantification, storage, and communication of a particular type of mathematically defined information. The field was established and formalized by Shannon in his seminal work [Shannon, 1948]. It proposed a mathematical framework for the theory of communication; however, it soon found a wide range of other applications. It is now at the intersection of mathematics, statistics and computer science, and has applications in diverse fields ranging from electrical engineering to neurobiology. Shannon's paper is quite influential as it defines the fundamentals of digital communications and information theory. At its core, information theory quantifies uncertainty, measures the quantity of information, and characterizes the information content between two signals. In addition, information theory offers a fundamental approach to cryptography and cybersecurity. Such quantities are essential to understand how different medical data sources can be combined to inform clinical decisions.

Key information-theoretic concepts for medical data fusion, among others, include:

- **Entropy** $H(X)$ - Quantifies uncertainty in a random variable X , thus in the context of medical data it can be used, for example, to assess diagnostic uncertainty before and after a test.
- **Mutual Information** $I(X; Y)$ - Measures shared information between variables, the two random variables X and Y , and is ideal for identifying redundant tests or discovering novel biomarker relationships.
- **Kullback-Leibler (KL) divergence** $D_{KL}(P \parallel Q)$ - Quantifies the difference between probability distributions, more precisely, it measures how much a probability distribution Q is different from a probability distribution P . It is essential for validating clinical models.

- **Rate-distortion function** $R(D)$ is a central concept in rate-distortion theory, which deals with lossy compression. It quantifies the trade-off between the bit rate R used to represent a source and the distortion D incurred when reconstructing the source from the compressed data. It is the minimum rate R at which a source can be represented while ensuring the average distortion does not exceed D . This is important for medical data since we can define the acceptable level of quality of the medical measurements.
- **Directed information** $I(\mathbf{X}^n \rightarrow \mathbf{Y}^n)$ is an information-theoretic measure that quantifies the information flow from the random sequence \mathbf{X}^n to another random sequence \mathbf{Y}^n . Directed information has applications to problems where causality plays an important role, such as in situations where feedback from \mathbf{Y}^n to \mathbf{X}^n is important. It is arguably important for medical applications, as the feedback for some type of patient parameters might improve the accuracy of the measurement of that specific parameter.

In order to use some of these concepts that are important in information theory, many scientists introduced and used several concepts that help in various aspects of data fusion, feature selection and machine learning.

The most sound and interesting concepts are the following:

- **Information Bottleneck** is a technique designed for finding the best tradeoff between accuracy and complexity when summarizing a random variable X with an observed relevant variable Y , given their joint probability distribution $P(X, Y)$. Applications include distributional clustering and dimension reduction, as well as deep learning. The information bottleneck can also be viewed as a rate distortion problem, with a distortion function that measures how well Y is predicted from a compressed representation T compared to its direct prediction from X .
- **Transfer entropy** is a non-parametric statistic measurement of the amount of directed transfer of information between two random processes. Transfer entropy from a process X to a process Y is the amount of uncertainty reduced in future values of Y by knowing the past values of X given past values of Y . Transfer entropy has been used for estimation of functional connectivity of neurons social influence in social networks and statistical causality between two processes. Hence, it has a great potential in evaluation of the causes of various diseases while processing medical data. Transfer entropy is a finite version of the directed information.
- **Partial Information Decomposition (PID)** decomposes multivariate information into redundancy, uniqueness, and synergy, thus providing theoretical guidance on fusion strategies. This measure has been introduced to generalize the pairwise relations between two variables described by, for

example, mutual information, to interactions between multiple variables.

This paper summarizes the existing approaches on information-theoretic concepts in the field of medical data fusion. Our contributions are focused on the following methods: (1) a systematic presentation of fundamental information theoretic concepts with clinical interpretations; (2) detailed analysis of theoretical frameworks, including partial information decomposition for medical fusion (3) comprehensive coverage of clinical applications across several medical domains; (4) a framework for evaluation of current challenges and future research directions; (5) practical guidelines for implementing information theoretic fusion in clinical settings.

Data fusion combines data from multiple and usually heterogeneous sources in order to obtain improved data quality and reduce data uncertainty. It is also essential for feature extraction and predictive analytics, and leads to reduced cost for the corresponding data processing. Data fusion has been used in various domains including sensor data fusion, geographical information systems, bioinformatics, and business intelligence. In this work, our focus is on data fusion techniques that can primarily increase data quality and help in data processing, feature selection and decision making in medical applications.

The remainder of this paper is organized such that Section II presents the mathematical foundations of information theory and their clinical interpretations. Section III introduces theoretical frameworks for multimodal medical data fusion. Section IV provides a comprehensive review of clinical applications. Section V addresses challenges and future directions, while Section VI concludes the review with a summary of the key contributions.

II. MATHEMATICAL FOUNDATIONS

In the current section, we elaborate on several theoretical concepts that show potential to be applied in the context of medical data fusion.

A. Entropy

In information theory, *entropy* is a measure of the uncertainty of a random variable. Thus, in medical data fusion, it can be used to quantify clinical uncertainty.

Definition II.1 (Shannon Entropy).

For a discrete random variable X with probability distribution $p(x)$, the Shannon entropy is

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x), \quad (1)$$

and is measured in bits, when the logarithm is base 2.

Clinical interpretation: Entropy may quantify diagnostic uncertainty. A patient presenting with chest pain might have $H(\text{Diagnosis}) = 2.1$ bits when considering eight possible diagnoses ($2^3 = 8$); after an ECG, entropy might drop to $H(\text{Diagnosis} | \text{ECG}) = 1.2$ bits, reflecting reduced uncertainty, or in other words, more certainty, hence, a more precise diagnosis. In this particular case, with eight possible diagnoses, the situation with the highest uncertainty, or highest entropy, is when all eight diagnoses are equiprobable. In that case, the probability of each case is $1/8$, and if we replace it in (1), we get an entropy of $\log_2(8) = 3$ bits. More generally, the maximal entropy in a case with N outcomes is $\log_2(N)$.

B. Mutual Information

Mutual information of two random variables is a measure of the mutual dependence between the two variables. Thus, we could use mutual information to measure shared information between any two random variables.

Definition II.2 (Mutual Information).

The mutual information between variables X and Y is

$$I(X; Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \quad (2)$$

These are the properties relevant to clinical applications:

- $I(X; Y) \geq 0$ with equality iff X, Y are independent,
- $I(X; Y) = I(Y; X)$ (symmetry),
- Non-linear relations missed by correlation,
- Invariant to monotonic transformations of data.

Clinical interpretation: Mutual information measures how much a diagnostic test reduces uncertainty about a disease. For example, $I(\text{Troponin}; \text{ACS}) = 0.82$ bits indicates that the diagnostic test "Troponin measurement" provides substantial information about the disease acute coronary syndrome (ACS). The Troponin test is a crucial tool in the diagnosis of Acute Coronary Syndrome (ACS). It is a specific marker for myocardial necrosis and is preferred for diagnosing myocardial infarction due to its high cardiac specificity. If the mutual information between another test and ACS is lower, then we can conclude that the Troponin test is preferred.

C. Kullback-Leibler Divergence

Kullback-Leibler (KL) Divergence is a type of statistical distance, or in other words, a measure of how much an approximating probability distribution Q is different from a true probability distribution P . It was introduced by Kullback and Leibler [1951]. Sometimes it is called relative entropy or I -divergence, like in Csiszar [1975]. Consider two probability distributions, a true P and an approximating Q . Often, P represents the data, the observations, or a measured probability distribution, and distribution Q represents a theory, a model, a description, or another approximation of P . However, sometimes the true distribution P represents a model and the approximating distribution Q represents (simulated) data that are intended to match the true distribution. The KL Divergence is then interpreted as the average difference of the number of bits required for encoding samples of P using a code optimized for Q rather than one optimized for P .

Definition II.3 (KL Divergence).

For two probability distributions P and Q over the same space

$$D_{\text{KL}}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}. \quad (3)$$

These are the properties of the KL Divergence:

- $D_{\text{KL}}(P || Q) \geq 0$, with equality iff $P = Q$,
- Asymmetric: $D_{\text{KL}}(P || Q) \neq D_{\text{KL}}(Q || P)$,
- Not a true distance metric but a divergence.

Clinical applications of KL Divergence:

- Population shift detection: Comparing current patient distribution to the historical baseline,
- Model validation: Detecting when clinical prediction models become unreliable,
- Multi-center trial monitoring: Identifying sites deviating from protocol,
- Diagnostic test evaluation: Quantifying how much a test shifts diagnostic probabilities.

The following fundamental relationships hold between mutual information and entropy, as well as mutual information and KL divergence:

$$I(X; Y) = H(X) - H(X|Y), \quad (4)$$

$$I(X; Y) = H(Y) - H(Y|X), \quad (5)$$

$$I(X; Y) = H(X) + H(Y) - H(X, Y), \quad (6)$$

$$I(X; Y) = D_{\text{KL}}(P(X, Y)||P(X)P(Y)). \quad (7)$$

In the above equations, $H(X, Y)$ is the joint entropy of the random variables X and Y , $H(A | B)$ is the conditional entropy of A when B is given, and $P(X, Y)$ is their joint probability distribution.

Equation (4) is particularly important clinically, since it shows that mutual information is exactly the reduction in uncertainty about X gained by knowing Y , which is the information gain from a diagnostic test. KL divergence remains fundamental in ensuring that fused multimodal representations preserve meaningful information while maintaining statistical consistency across different medical imaging modalities and clinical data sources.

D. Rate-Distortion Function

Rate-Distortion (R-D) theory, pioneered by Claude Shannon, addresses a key question in lossy data compression. Given a source signal X , it explores what is the minimum number of bits per symbol R required to represent X such that the average distortion between X and its reconstructed version \hat{X} does not exceed a given threshold D . This fundamental limit is called *rate-distortion function*, and is denoted $R(D)$.

Definition II.4 (Rate-Distortion function).

The Rate Distortion function $R(D)$ is defined as the minimum of the mutual information $I(X; \hat{X})$ over all conditional probability distributions $p(\hat{x}|x)$ that satisfy the distortion constraint $E[d(X, \hat{X})] \leq D$:

$$R(D) = \min_{p(\hat{x}|x): E[d(X, \hat{X})] \leq D} I(X; \hat{X}). \quad (8)$$

Here, $d(X, \hat{X})$ is some distortion metric (e.g., Mean Squared Error (MSE) for continuous signals or Hamming distance for discrete data). $R(D)$ is a convex, non-increasing function, signifying that to achieve lower distortion, a higher bit rate is necessary. The curve effectively defines the boundary of what is theoretically achievable. Applying R-D theory to fused medical data is more complex than applying it to a single modality. The goal shifts from compressing a single signal to compressing a set of correlated signals (modalities) in a way that preserves the clinically relevant information contained in their joint structure. The theoretical insights from R-D analysis can inform the development of practical compression algorithms for medical data fusion. Neural networks, particularly autoencoders with a rate-distortion loss function, can be trained to learn an optimal, non-linear transformation of the multi-modal data. The loss function is designed to minimize a weighted sum of the rate and the task-specific fusion distortion, effectively learning to approximate the $R(D)$ function for the given data and clinical task. In summary, Rate-Distortion theory provides an indispensable, rigorous foundation for designing efficient and clinically effective medical data fusion systems.

E. Directed Information

While conventional information-theoretic measures such as mutual information capture symmetric statistical dependencies between variables, they are fundamentally limited in their ability to distinguish causal relationships. In the context of medical data fusion, where the goal is often to understand the directional influence between modalities or from modalities to clinical outcomes, a more precise framework is required. Directed information, originally developed by Massey [1990], provides a framework to quantify the causal influence from one stochastic process to another, respecting the temporal or logical flow of information.

Definition II.5 (Directed information).

Directed information measures the amount of information that flows from a source process $\mathbf{X} = (X_1, X_2, \dots, X_n)$ to a target process $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ in a causal manner. It is formally defined as

$$I(\mathbf{X} \rightarrow \mathbf{Y}) = \sum_{i=1}^n I(\mathbf{X}^i; Y_i | \mathbf{Y}^{i-1}), \quad (9)$$

where $\mathbf{X}^i = (X_1, X_2, \dots, X_i)$ represents the past and present of the source up to time i , and $\mathbf{Y}^{i-1} = (Y_1, Y_2, \dots, Y_{i-1})$ represents the past of the target before time i . The key distinction from mutual information is the conditioning on the past of the target, which ensures that only predictive information that is not already contained in the target's own history is attributed as directed information.

This formulation captures the intuitive notion of causality: \mathbf{X} causes \mathbf{Y} if knowing the past of \mathbf{X} helps predict the next value of \mathbf{Y} better than knowing only the past of \mathbf{Y} alone. The relationship between directed information and mutual information is given by

$$I(\mathbf{X} \rightarrow \mathbf{Y}) + I(\mathbf{X} \leftarrow \mathbf{Y}) = I(\mathbf{X}; \mathbf{Y}), \quad (10)$$

where $I(\mathbf{X} \leftarrow \mathbf{Y})$ represents the information flow in the opposite direction. This decomposition reveals that mutual information can be viewed as the sum of information flowing in both directions. The application of directed information to medical data fusion addresses several critical challenges that arise when integrating heterogeneous clinical data sources. Modern healthcare generates extensive longitudinal data, including continuous physiological monitoring, repeated imaging studies, and sequential laboratory measurements. Directed information provides a principled framework for understanding how these temporal processes influence one another. For example, in critical care settings, one might ask how changes in electrocardiogram (ECG) patterns direct information to subsequent blood pressure measurements, or how medication administration times influence physiological responses.

III. THEORETICAL FRAMEWORKS FOR MULTIMODAL DATA FUSION

In this section, we describe three frameworks that are motivated by the mathematical foundations of information theory and are proven to be useful in multimodal data fusion in general, hence in the medical applications that rely on multimodality.

A. Directed Information

The Information Bottleneck (IB) principle provides a theoretical foundation for learning efficient representations. The information bottleneck method is a technique designed for finding the best tradeoff between accuracy and complexity (compression) when summarizing a random variable X , given a joint probability distribution $P(X, Y)$ between X and an observed relevant variable Y . It was introduced by Tishby et al. [2000]. Applications include distributional clustering and dimension reduction, and more recently, it has been suggested as a theoretical foundation for deep learning. It generalized the classical notion of minimal sufficient statistics from parametric statistics to arbitrary distributions, not necessarily of exponential form. It does so by relaxing the sufficiency condition to capture some fraction of the mutual information with the relevant variable Y . The information bottleneck can also be viewed as a rate distortion problem, with a distortion function that measures how well Y is predicted from a compressed representation T compared to its direct prediction from X . This interpretation provides a general iterative algorithm for solving the information bottleneck trade-off and calculating the information curve from the distribution $P(X, Y)$. Let the compressed representation be given by a random variable Z_θ , which is a learned representation of the input X . The algorithm minimizes the following functional with respect to the conditional distribution $\theta = P(Z_\theta | X)$. The information bottleneck is defined as

$$\min_{\theta} (I(X; Z_\theta) - \lambda I(Z_\theta; Y)), \quad (11)$$

where λ is a Lagrange multiplier that controls the trade-off between compression and prediction.

The following clinical applications were identified for the use of information bottleneck: learning robust features invariant to artefacts and patient variations, filtering out irrelevant information while preserving diagnostically relevant content, and improving generalization across different hospitals and populations.

B. Transfer Entropy

Transfer entropy (TE) measures directional information flow between physiological signals. It was proposed by Schreiber [2000] and is a powerful concept that adds a dynamic, directional layer to the information-weighted sensor fusion we discussed earlier. While fusion focuses on combining data at the same time for a reliable snapshot, transfer entropy helps us understand the flow of informa-

tion over time between different signals. It is a statistical measure that quantifies the directed (asymmetric) flow of information from one time series to another. In simpler terms, it tells you how much the past values of a source variable (e.g., heart rate) help predict the future values of a target variable (e.g., blood pressure), over and above what the target's own past could predict. If including the source's past reduces uncertainty about the target's future, we say information has flowed from source to target. This makes transfer entropy particularly useful for understanding complex, interconnected systems. Let's for $i < j$, denote any process $(A_i, A_{i+1}, \dots, A_j)$ by the vector \mathbf{A}_i^j . Also, denote the process $(B_{-\infty}, \dots, B_j)$ by the vector \mathbf{B}^j . The most common general definition of transfer entropy from a process \mathbf{X}^t to a process \mathbf{Y}^t is

$$T_{\mathbf{X} \rightarrow \mathbf{Y}} = I(Y_t; \mathbf{X}_{t-1}^{t-L} | \mathbf{Y}_{t-1}^{t-K}), \quad (12)$$

where $I(C;D|F)$ is the conditional mutual information between C and D given F , Y_t is the future state of the target variable at time t , \mathbf{X}_{t-1}^{t-L} represents the past states of the source variable (from $t-1$ to $t-L$), and \mathbf{Y}_{t-1}^{t-K} represents the past states of the target variable (from $t-1$ back to $t-K$), which we are conditioning on. In essence, this definition reads that the transfer entropy from \mathbf{X} to \mathbf{Y} is the information shared between the future of \mathbf{Y} and the past of \mathbf{X} , given that we already know the past of \mathbf{Y} . To make this practical for calculation, transfer entropy is usually expressed using Shannon Entropy H . If we simplify by assuming the processes are first-order Markov chains (i.e., we only need to consider the immediate past $t-1$, to predict the next state), the equation becomes

$$T_{\mathbf{X} \rightarrow \mathbf{Y}} = H(Y_t | Y_{t-1}) - H(Y_t | Y_{t-1}, X_{t-1}), \quad (13)$$

where $H(Y_t | Y_{t-1})$ is the entropy (or uncertainty) of current Y_t given its own immediate past, $H(Y_t | Y_{t-1}, X_{t-1})$ is the entropy of the future of current Y_t given both its own past and the immediate past of \mathbf{X} .

By subtracting the second term from the first, we see exactly how much the uncertainty of \mathbf{Y} is reduced by knowing the past of \mathbf{X} . This reduction is, in fact, the transfer entropy. Transfer entropy can also be defined as the Kullback-Leibler divergence between two probability distributions

$$T_{\mathbf{X} \rightarrow \mathbf{Y}} = D_{\text{KL}}(P(Y_{t+1}, |Y_t, X_t) || P(Y_{t+1} | Y_t)). \quad (14)$$

The fundamental goal of TE is to detect causality or directional influence. It is an asymmetric measure, meaning the information flow from \mathbf{X} to \mathbf{Y} is different from the flow from \mathbf{Y} to \mathbf{X} . This allows us to map the directional influences within a network, which is crucial for identifying which variable is driving changes in another. Transfer entropy is actively being used to solve practical problems such as the prediction of patient deterioration in intensive care units (ICUs).

C. Partial Information Decomposition (PID)

The Partial Information Decomposition (PID) framework was introduced by Williams and Beer [2010] as an extension of information theory that aims to generalize the pairwise relations described by information theory to the interaction of multiple variables. It has been shown that PID cannot follow the core properties and axioms of classical information theory. While mutual information is able to capture non-linear statistical relationships between two random variables, the PID framework may quantify multivariate statistical dependencies in a more general way. For two source variables X_1, X_2 and a target variable Y , the mutual information $I(X_1, X_2; Y)$ can be decomposed into four components:

$$I(X_1, X_2; Y) = U(X_1; Y \setminus X_2) + U(X_2; Y \setminus X_1) + S(X_1, X_2; Y) + R(X_1, X_2; Y), \quad (15)$$

where $U(X_1; Y \setminus X_2)$ is the unique information from X_1 to Y which is not in X_2 , $U(X_2; Y \setminus X_1)$ is the unique information from X_2 to Y which is not in X_1 , $S(X_1, X_2; Y)$ is the synergistic information that is in the interaction of X_1 and X_2 about Y , and $R(X_1, X_2; Y)$ is the redundant information present in both sources. PID has been applied to di-

verse fields. Liang et al. [2023] applied PID to information fusion for medical data.

Clinical Implications of PID Components may be summarized in the following examples, where these measures could be used:

- **High redundancy (R):** Multiple tests provide overlapping information. In that case we consider eliminating redundant tests to reduce cost and patient burden.
- **High uniqueness ($U(X_1; Y \setminus X_2)$, $U(X_2; Y \setminus X_1)$):** Each modality contributes distinct information, hence both are needed for comprehensive assessment.
- **High synergy (S):** Information emerges only from combination, it requires sophisticated fusion approaches (early or intermediate fusion).

Zhang et al. [2025] conducted the first comprehensive evaluation of PID metrics across diverse biomedical data using seven different modality pairs. The data modalities studied in this paper include clinical radiology, pathology, and genomic data. Future Research Directions may involve hybrid frameworks combining PID with deep learning for end-to-end interpretable fusion and causal PID extending decomposition to capture causal relationships between modalities.

IV. CLINICAL APPLICATIONS OF INFORMATION-THEORETIC FUSION

There is an enormous potential for the use of information theory for the purpose of medical data fusion. Next, we summarize the medical applications in which information-theoretic fusion, may improve the outcome of the patient examination.

Neurology and Neuroimaging are promising use cases, as Alzheimer's disease diagnosis and multi-center clinical trial monitoring for brain functioning. Multimodal fusion combining MRI, PET, Cerebrospinal fluid (CSF) biomarkers, and cognitive tests has transformed Alzheimer's diagnosis. Main findings by [Zhang et al., 2020] conclude that amyloid PET and CSF provide redundant information, tau PET provides unique prognostic information, and cognitive tests provide information almost independent of biomarkers. Speaking about multi-center clinical trial monitoring, traditional methods for analyzing brain connectivity often rely on correlations across a group of subjects, which averages out individual variability. For a single patient's PET scan, it compares the statistical distribution of signals in every pair of brain regions using the symmetric Kullback-Leibler divergence defined as $\text{DKL}(P||Q) + \text{DKL}(Q||P)$. By collecting data from different sites and measuring some of their features, the authors conclude that some particular sites have patients with more advanced disease. It may allow early detection and may enable corrective action.

Cardiovascular Medicine is another important area where the information-theoretic approach may improve the diagnostics and treatment. It is used in heart failure prognosis and multi-sensor intensive care unit (ICU) monitoring. Wu et al. [2025] reviewed multimodal deep learning for heart failure prognosis using: admission notes (unstructured text), clinical tabular data (labs, vitals), electrocardiogram (ECG) waveforms, and echocardiography images. Integration improved prognostic precision, with moderate redundancy. By fusing information with an understanding of context and signal quality, it can filter noise, reduce cognitive load, and alert clinicians to the most critical events with higher confidence. Intelligent systems assess the quality and context of each signal to create a fused, more accurate estimate. The improved monitoring comes from the enhanced signal reliability, context-aware alarms, multi-dimensional anomaly detection, efficient mobility assessment, and comprehensive respiratory monitoring.

Other areas where information theory may be used are oncology and sepsis prediction. There are several emerging applications that include digital twins for personalized medicine and wearables for continuous monitoring. Artificial Intelligence (AI) powered aging digital twins are proposed to combine genetic data, medical history, continuous physiological monitoring (e.g., from wearable sensors or

other standard consumer electronics), lifestyle factors, and similar. Information theory guides which data sources are most informative for individual health trajectories. The integration of consumer wearables with clinical monitoring systems enables seamless transition between in-hospital and ambulatory monitoring, early detection of deterioration after discharge, longitudinal tracking of physiological baselines, and personalized alert thresholds based on individual patterns.

V. CHALLENGES AND FUTURE DIRECTIONS

Current challenges may be summarized into two groups: theoretical and clinical challenges. Theoretical challenges are connected to parameter sensitivity, scalability, interpretability, and missing modalities. Some of the metrics may heavily depend on the preprocessing, while information-theoretic calculations may become computationally expensive. In addition, real-world clinical data are often incomplete. Potential solutions involve sensitivity analysis, approximation methods, hierarchical fusion, and generative models. The clinical challenges are related to the regulatory hurdles across various jurisdictions. Most studies remain retrospective rather than prospective. Many clinicians require interpretable outputs for trust and adoption, and the new proposed methodologies must complement rather than disrupt existing practices. Lastly, one has to ensure that the novel models perform equally across diverse populations. There are multiple emerging technologies and opportunities. Generative AI offers powerful tools for medical data fusion, including the synthesis of the missing modalities when appropriate, augmenting limited datasets for rare conditions, creating counterfactual examples for model explanation, and generating realistic synthetic patients for simulation and training. Large Language Models promise to unlock information in unstructured clinical narratives by extracting structured information from physician notes, integrating textual data with structured Electronic Health Records, providing natural language explanations of fusion results, and enabling conversational AI for clinical decision support. Self Supervised Learning approaches may reduce dependence on labeled data by learning representations from unlabeled multimodal data, pre-training on large-scale public datasets, and fine-tuning on specific clinical tasks with limited labels. Federated learning enables multi-institutional collaboration without sharing patient data, exploring the models trained across hospitals while data remains local. It may also enable rare disease research through data aggregation. Future research directions may improve the estimation by using robust methods with uncertainty quantification. It is also interesting to explore the causal information theory by distinguishing correlation from causation in fusion. Another potential improvement may be applied with the use of dynamic information theory by tracking information flow over time. From a research point of view, a promising approach may be the use of Riemannian geometry of statistical manifolds.

VI. CONCLUSION

Information theory provides a rigorous mathematical framework that is transforming medical data fusion across the healthcare continuum. This comprehensive review has synthesized the foundational concepts, theoretical frameworks, clinical applications, and future directions of this rapidly evolving field. We have provided a systematic presentation of the main information-theoretical quantities with clear clinical interpretations and practical examples. The PID framework offers theoretical guidance on fundamental fusion questions: whether to fuse, how to fuse, and what to expect from fusion. Information-theoretic fusion has demonstrated impact across major medical domains: oncology, neurology, cardiology, and critical care. The future of IT medical data fusion lies at the intersection of these converging trends: theoretical advances, methodological innovations, clinical translation, and technological convergence. Information theory has fundamentally altered how we think about medical data fusion, moving from an ad-hoc combination of data sources to principled quantification of information content, redundancy, and synergy. The field has matured from theoretical foundations to practical applications that are already improving

patient care through earlier detection, to more accurate diagnosis, and personalized treatment. Yet significant challenges remain. The parameter sensitivity of PID metrics, the gap between theoretical recommendations and empirical performance, and the substantial barriers to clinical translation remind us that this field is still in development. The most exciting advances may lie ahead, as emerging technologies like generative AI and large language models create new opportunities for IT analysis of increasingly complex, multimodal medical data. For clinicians, researchers, and developers working at this intersection, the message is clear: information theory provides powerful tools for understanding and optimizing medical data fusion, but these tools must be applied thoughtfully, validated rigorously, and interpreted in the context of real-world clinical constraints. When used appropriately, they can unlock the full potential of multimodal patient data to improve diagnostic accuracy, prognostic precision, and ultimately, improve the overall health of patients.

REFERENCES

- [1] G. I. Csiszar. I-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.
- [2] J. Duan, J. Xiong, Y. Li, and W. Ding. Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion*, 2024. doi: 10.1016/j.inffus.2024.102536.
- [3] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.
- [4] P. P. Liang, Y. Cheng, X. Fan, C. K. Ling, S. Nie, R. Chen, Z. Deng, N. Allen, R. Auerbach, F. Mahmood, R. Salakhutdinov, and L.-P. Morency. Quantifying & modeling multimodal interactions: An information decomposition framework, 2023.
- [5] J. Massey. Causality, feedback and directed information. In *Proceedings 1990 International Symposium on Information Theory and its Applications (ISITA)*. Waikiki, Hawaii, Nov. 1990.
- [6] M. Salvi, H. W. Loh, S. Seoni, P. D. Barua, S. Garcia, F. Molinari, and U. R. Acharya. Multi-modality approaches for medical support systems: A systematic review of the last decade. *Information Fusion*, 103, 2024. doi: 10.1016/j.inffus.2023.102134.
- [7] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000. doi: 10.1103/PhysRevLett.85.461.
- [8] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–423, 1948.
- [9] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method, 2000. URL <https://arxiv.org/abs/physics/0004057>.
- [10] P. L. Williams and R. D. Beer. Nonnegative decomposition of multivariate information. 2010. URL <https://arxiv.org/abs/1004.2515>.
- [11] J. Wu, K. He, R. Mao, X. Shang, and E. Cambria. Harnessing the potential of multimodal EHR data: A comprehensive survey of clinical predictive modelling for intelligent healthcare. *Inf. Fusion*, 123, 2025.
- [12] T. Zhang, R. Ding, K.-D. Luong, and W. Hsu. Evaluating an information theoretic approach for selecting multimodal data fusion methods. *Journal of Biomedical Informatics*, 167, 2025.
- [13] Y.-D. Zhang, Z. Dong, S.-H. Wang, X. Yu, X. Yao, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez, et al. Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. *Information Fusion*, 64:149–187, 2020. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955.