Check for updates

Original Research Article

Health Informatics Journal

Health Informatics Journal
1–20
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14604582241290474
journals.sagepub.com/home/jhi
Sage

Multimodal representation learning for medical analytics - a systematic literature review

Emil Riis Hansen and Tomer Sagi
Department of Computer Science, Aalborg University, Aalborg, Denmark

Katja Hose
Department of Computer Science, Aalborg University, Aalborg, Denmark; Department of Informatics, TU Wien, Wien, Austria

Abstract

Objectives: Machine learning-based analytics over uni-modal medical data has shown considerable promise and is now routinely deployed in diagnostic procedures. However, patient data consists of diverse types of data. By exploiting such data, multimodal approaches promise to revolutionize our ability to provide personalized care. Attempts to combine two modalities in a single diagnostic task have utilized the evolving field of multimodal representation learning (MRL), which learns a shared latent space between related modality samples. This new space can be used to improve the performance of machine-learning-based analytics. So far, however, our understanding of how modalities have been applied in MRL-based medical applications and which modalities are best suited for specific medical tasks is still unclear, as previous reviews have not addressed the medical analytics domain and its unique challenges and opportunities. Instead, this work aims to review the landscape of MRL for medical tasks to highlight opportunities for advancing medical applications. **Methods:** This paper presents a framework for positioning MRL techniques and medical modalities. More than 1000 papers related to medical analytics were reviewed, positioned, and classified using the proposed framework in the most extensive review to date. The paper further provides an online tool for researchers and developers of medical analytics to dive into the rapidly changing landscape of MRL for medical applications. **Results:** The main finding is that work in the domain has been sparse: only a few medical informatics tasks have been the target of much MRL-based work, with the overwhelming majority of tasks being diagnostic rather than prognostic.

Corresponding author:
Tomer Sagi, Department of Computer Science, Aalborg University, Selma Lagerløfs Vej 300, Aalborg East, Aalborg 9220, Denmark.
Email: tsagi@cs.aau.dk

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Similarly, numerous potentially compatible information modality combinations are unexplored or under-explored for most medical tasks. **Conclusions:** There is much to gain from using MRL in many unexplored combinations of medical tasks and modalities. This work can guide researchers working on a specific medical application to identify under-explored modality combinations and identify novel and emerging MRL techniques that can be adapted to the task at hand.

Keywords
correlation, embedding, fusion, machine learning, medical analytics, multi-modality, similarity

Introduction

Background

The world is inherently multimodal. Entities, from patients to proteins, can be described in various ways called modalities. The onset of diseases and conditions can be measured in a medical setting through a measurable change in biomarker modalities, such as blood-pressure, heart-rate, and x-ray findings. As an example, the progression of Alzheimer's Disease (AD) has shown a correlation with modalities such as Magnetic Resonance Imaging (MRI),[1] Positron Emission Tomography (PET),[2] and protein measures of Cerebrospinal Fluid (CSF).[3] MRI provides a means of detecting atrophied brain regions, PET can reveal hypometabolism,[4] and protein measures of CSF can detect the presence of beta-amyloid (Aβ42) and tau (τ) proteins characteristic of AD.[5] Each modality provides unique information, which, combined, could improve AD progression classification.

Medical machine learning (ML)-based analytics attempt to improve the quality and speed of previously manual tasks and have featured predominantly uni-modal approaches. Combining multiple information modalities, similar to a physician considering multiple sources of information, can enhance the performance of complex predictive ML-based analytics.

Multimodal representation learning (MRL)[6] is a theoretical and practical framework for combining multiple information modalities to improve the effectiveness of ML. MRL has recently expanded into the medical analytics domain, where it has been used to combine multiple medical modalities for diagnosis and prognosis tasks.[7] However, no comprehensive survey of MRL in the medical domain has been performed, leaving researchers to piece together which modality combinations have been attempted for various medical analytics using MRL techniques.

Furthermore, various medical information modalities exist in the medical space, such as omics data, medical images, textual medical records, electronic health records (EHR), computerized clinical practice guidelines, and biomedical knowledge graphs. It has become daunting to sift through these options with medical analytics in mind and identify which are relevant, have been used previously, and in what combination.

This work reviews the use of MRL as a computational technique for utilizing multiple sources of information to improve the performance of ML-based medical analytics. The review describes and classifies the techniques and provides a hierarchy of medical information modalities over which one could attempt MRL. The review represents a comprehensive, structured survey of publications utilizing MRL for ML-based medical analytics, positioning them in the following classification spaces. The MRL classification space describes the specific MRL technique used. The medical

information modality classification space describes the modalities being integrated. The medical application classification space describes the clinical motivation, and the utility classification space describes the intended use of the analytics.

The review thereby addresses the following questions: What has been done in the medical analytics field with MRL? Which dimensions of analysis can be defined and used to identify unexplored techniques and application areas of MRL in the medical domain? This paper's contribution to answering these questions can be summarized as follows: it provides a comprehensive review of MRL technologies, improving upon previous surveys by proposing a novel and updated MRL classification space, including neural network technologies as a top-level category. This category is further expanded with recent technologies, such as attention techniques, convolution neural networks, and autoencoders. The paper provides a novel taxonomic hierarchy for structuring medical information modalities into three levels, starting from structured and unstructured data. Furthermore, an explorable online analysis is provided for diving into MRL-based techniques for medical applications, opening up the potential for researchers to investigate the current state of the art and novel ideas for medical MRL.

Related work

Previous surveys of MRL over general-purpose application domains have focused on the type of MRL technique. While surveys reviewing multimodal deep-learning[8–10] review new technologies and narrow the scope of the MRL technique employed. In this work, we do not limit ourselves to specific branches of MRL techniques. Additional surveys contextualize MRL in the general challenge of multimodal ML,[6] with a focus on fusion-based MRL[11] or its mathematical-theoretical foundations,[12] disregarding the application domain. We, however, provide a comprehensive systematic review of the medical analytics domain.

To the best of our knowledge, this paper presents the first attempt to review MRL for medical applications and provide a classification space in which modality combinations from the literature can be placed and future medical analytics designed. Furthermore, this work is the first to comprehensively review more than 1400 papers for MRL in multimodal medical applications while classifying the literature into four dimensions, i.e., *utility*, *medical*, *modality*, and *MRL*.

Methods

The following section describes the review methodology employed in this work. It begins by describing the classification space used throughout this paper, followed by the details of the section *Review Methodology.*

Classification space

The following classification space is used throughout this work to structure our survey of previous work utilizing MRL for ML-based medical analytics. The classification space comprises three orthogonal dimensions of classification. The *utility* dimension describes the intended use of the medical analytics task. The *information modality* dimension describes the types of medical information modalities incorporated in the MRL approach. Finally, the *MRL approach* dimension describes the MRL technique used. We begin by defining a medical analytics task.

Definition 1. Medical Analytics Task

A medical entity *is an item of interest for medical purposes. Entities can be patients, diseases, tumors, viruses, blood samples, etc. A* medical analytics task *provides information on a medical entity in an automated manner using an algorithm or a learned model over some input data.*

Medical analytics utility. It is common to classify *analytics* by utility.[13] *Descriptive* analytics describes the given input. This type of analytics is the most common. It contains methods, such as classification (this is an ultrasound image) and object detection (the image contains three lesions in these coordinates). *Diagnostic* analytics methods attempt to identify the root cause of the observed phenomena and are used to diagnose diseases and sub-types of diseases with similar symptoms but slightly different causes. *Predictive* analytics, also often described in the medical domain as *prognostic*, attempt to predict the occurrence of a future event or state from the current state or the sequence of states given as input. A medical example could be *sepsis mortality prediction*.[14] *Prescriptive* analytics provide one or more recommended actions to take in response to the given input. While in the general domain, this form of analytics is sometimes employed as autonomous agents, e.g., as recommendation systems, in the medical domain, they are used for decision support, e.g., treatment recommendation.[15]

Medical information modalities. A *medical information modality* is a data representation used to present information about a medical entity and used by medical analytics. In data management and ML,[16] it is common to distinguish between structured and unstructured data. Structured information is discretized into records, each containing fields that are assigned values describing some medical entity. For example, a relational database, where tables contain records sharing a fixed schema for describing the status of a patient.

Figure 1 presents a partial view of the proposed hierarchy, showing levels one and two in full and examples from the third level.

The top level of the hierarchy separates structured and unstructured modalities. On the second level, multiple medical modalities are grouped using common groupings in ML literature, such as *image*, *text*, and *timeseries*. The third level represents specific medical information modalities used in MRL analytics. Furthermore, level three concepts are mapped to SNOMED taxonomy concepts to the extent possible. Hence, the SNOMED sub-classes can serve as the fourth and onward levels of the hierarchy. Furthermore, through these concepts, our third-level concepts can be connected to other terminologies and translated into other languages. For example, our level three concept *Computed Tomography* is mapped to the SNOMED concept *Computed tomography (procedure)*. Using the Bioportal SNOMED ontology[1] the SNOMED concept can be mapped to other taxonomies like MedDRA[2] with the concept *CT scan* and BIM[3] with *Computed_Tomography*. The medical modality hierarchy is complete with respect to the set of publications surveyed in this work and available online.[4]

Multimodal representation learning. To perform a structured analysis of existing MRL approaches, they are organized in a hierarchical structure. Let us first define MRL.

Definition 2. Multimodal Representation Learning

Given two datasets x and y of disparate but correlated information modalities, where $x_i \in x$ *and* $y_i \in y$ *represent samples describing the same real-world entity, then* multimodal representation learning (MRL) *is defined as the challenge of finding a latent space where uni-modal modalities can coexist.*

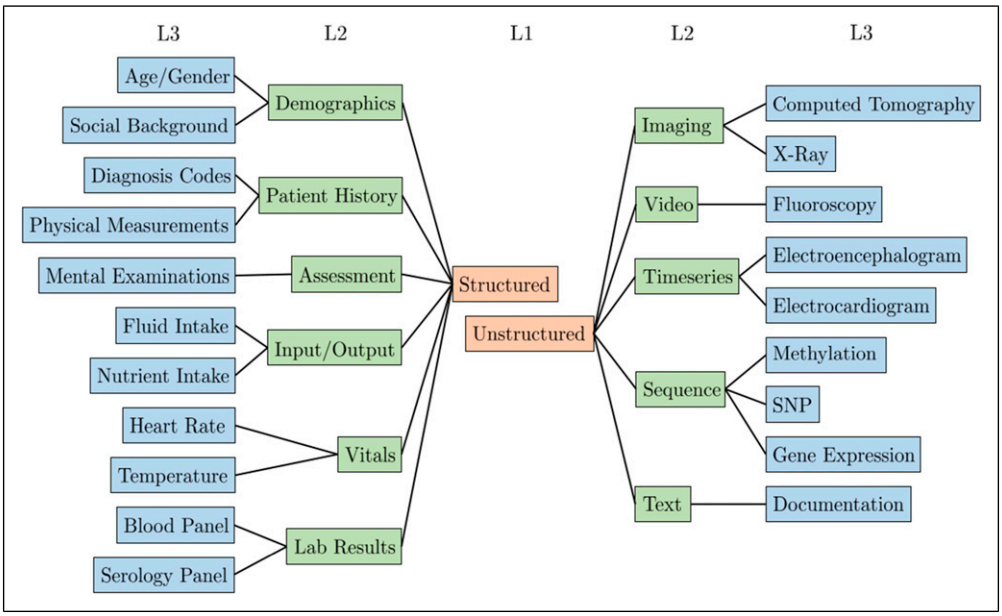

Figure 1. Partial hierarchy for structuring medical modalities. The full hierarchy can be found at https://tabsoft.co/40aAECd.

Thus, the latent space contains information from both medical modalities and hence should enable improved subsequent medical analytics compared to uni-modal approaches. Figure 2 exemplifies MRL for Alzheimer's Disease (AD) classification.

MRL techniques can be broadly classified into *alignment*, *fusion*, and *neural*, as illustrated in Figure 3 (adapted from[12]). Generally, *alignment* techniques find a feature space where modalities can coexist, *fusion* techniques combine uni-modal features into a new latent representation, and *neural* techniques jointly learn a latent representation combining uni-modalities and learn a model for solving a medical analytics task. The remainder of this section presents these subcategories.

Alignment MRL (AMRL). AMRL learns a representation space in which uni-modal modalities x and y can coexist, with the goal that similar samples should be closer together in the learned space than dissimilar samples. Mathematically, this can be formulated as $f(x_1) \sim g(y_i)$, where f and g are modality-specific projection functions that map individual samples x_i and y_i into a multimodal space, and $\sim$ indicates that some distance measure aligns the new space, as illustrated in Figure 3(a). AMRL can be subdivided into *correlation* and *similarity* techniques.

Similarity-aligned representation learning learns an aligned space between x and y by optimizing a distance function for positive and negative modality samples.[6] Shared for all similarity-based methods is the idea of learning transformation matrices f and g by minimizing a distance metric, such as the dot-product similarity or hinge rank loss, often by utilizing stochastic gradient descent (SGD) (Figure 4).

One of the earliest examples is *general similarity learning (GSL)*.[17] GSL creates an aligned space between pairs of images and textual annotations by learning projection functions to map the modalities into a shared space using the weighted approximate-rank pairwise loss. In the resulting coordinated space, similar samples of images and textual annotations have a smaller cosine distance from each other.

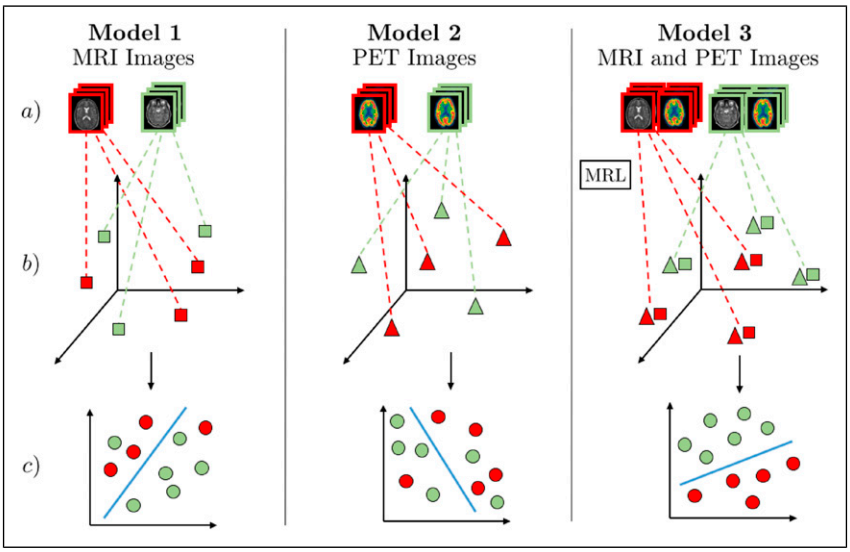

Figure 2. MRL for discriminating Alzheimer's disease (AD) patients from healthy subjects∼(HS). The three models (1 and 2 - uni-modal, 3 - multimodal) consist of three steps: (a) Receive multimodal samples, (b) map samples to their individual representation spaces, or, in the case of Model 3, use MRL to map modalities to a shared semantic space, (c) classification of AD/HS. Red represents AD-positive samples, and green represents AD-negative samples. Models 1 and 2 use MRI and PET images for uni-modal AD classification. In Model 3, MRL is used to find a shared semantic space combining MRI and PET images, capturing the underlying semantic correlation between these modalities. As illustrated in step c) of Model 3, the combined discriminative information from a shared semantic space between MRI and PET can be used for superior medical analytics such as AD classification.[39]

Whereas GSL is limited by the choice of initial uni-modal embeddings, *deep similarity learning* (DSL)[18] jointly learns initial uni-modal feature representations and subsequent transformation matrices f and g in an end-to-end framework. This can be achieved by adding layers of trainable fully-connected neural networks (NN) to step *a* in Figure 4.

Hence, the initial embeddings and subsequent aligned representation space can be jointly learned. Extensions of DSL include using different combinations of loss functions[19] and neural network architectures.

Correlation employs statistical methods for finding the correlation between two sets of variables. One of the most popular techniques is canonical correlation analysis (CCA). CCA was first introduced in 1936 by H. Hotelling.[20]

Given two sets of variables x and y, CCA (Figure 5) finds the linear projections f and g that maximize the correlation between variables from the projected space of $f(x)$ and $g(y)$ as $arg\max_{f,g} corr(f(x), g(y))$. Finding the transformations resulting in a maximally correlated space can be solved by generalized eigendecomposition. CCA is thus able to find the linear transformations f and g, that maximize the correlation between variables of the transformed modalities. Hence, the original CCA technique is linear with respect to the projection matrices f and g.

Various non-linear extensions to the classical CCA have been proposed, such as Deep CCA (DCCA)[21] using Fully Connected Neural Networks (FCNNs) for initial feature learning and Kernel CCA (KCCA)[22] utilizing kernels for non-linear feature transformation. Furthermore, extensions to multiple sets of variables have also been proposed, such as Multi CCA (MCCA),[23] which learns a shared space between multiple sets of variables. The many CCA variants are reviewed in ref.[24]

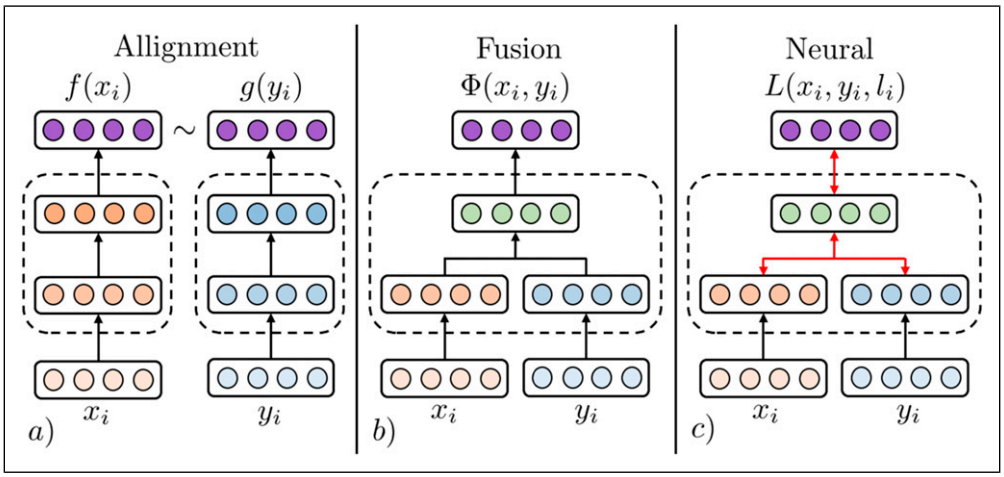

Figure 3. An illustration of fusion and coordination categories of MRL. The MRL step of Model 3 from Figure 2 can be substituted by techniques from all three categories of MRL. x_i and y_i are disparate but correlated uni-modal samples describing the same real-world entity. Arrows represent data transformations, dashed lines are optional transformation steps, and colored dots represent features of. x_i and y_i. (a) illustrates alignment MRL, where . x_i and y_i are aligned though the coordination operator $\sim$ on $f(x_i)$ and $g(y_i)$. (b) illustrates fusion MRL, where uni-modal features from x_i and y_i are fused through a vector combination technique ϕ. (c) illustrates neural MRL, where neural network technologies combined with a loss function L are used to simultaneously learn uni-modal latent representations, a shared latent representation and a medical analytic based on it. Red arrows indicate representation updates using backpropagation.

Fusion MRL (FMRL). Mathematically, FMRL can be formulated as $z = \phi(x_i, y_i)$, where ϕ is a function that combines uni-modal data samples x_i and y_i, and z is the combined multimodal representation. Fusion techniques are usually used to increase the accuracy of classification problems where multiple modalities have distinct discriminative properties.[10] FMRL techniques are further divided into *joining*, *kernels*, and *graphical models*, with complexities varying from linear feature concatenation to complex kernel combinations.

Joining combines modalities by concatenating early, intermediate, or late modality-specific features. *Early Joining (EJ)*[25] combines modality features using concatenation functions before any data transformations have been applied to individual modalities (Figure 6(a)). While EJ is simple and efficient in combining multimodal data, problems arise when modalities have varying sampling rates. For example, when combining MRI images and EEG signals. To alleviate such problems, *Intermediate Joining* can be used, where uni-modalities are transformed before latent features are concatenated (Figure 6(b)); however, manual engineering of modality-specific feature transformations is time-consuming and requires extensive domain knowledge.

Decision Joining (DJ) combines the results of multiple uni-modal analytics, either by majority vote, weighted linear combinations, or more complex techniques (Figure 6(c)). DJ is sometimes preferred in tasks involving low-correlated modalities as the technique is modality-independent, and errors from individual analytics tend to be uncorrelated.[11]

Kernels project linearly inseparable data into higher dimensional but linearly separable representation spaces using a non-linear kernel transformation. *Multiple Kernel Learning (MKL)* is a sub-type utilizing multiple such kernels. Well-known kernel techniques include support vector machines, the kernel-fisher discriminant, and regularized AdaBoost[26]). Multimodal representation

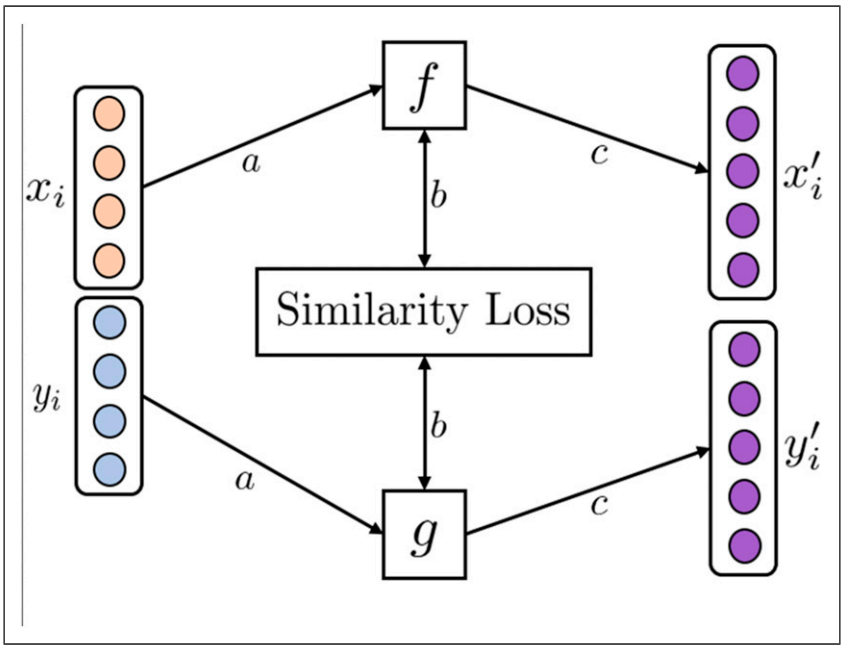

Figure 4. GSL technique of AMRL. (a) Modality samples x_i and y_i are transformed by modality-specific transformations f and g. (b) Using a similarity loss between $f(x_i)$ and $g(y_i)$, SGD iteratively updates f and g. (c) When learning has finished, f and g transform entities x_i and y_i into the coordinated space x_i' and y_i'.

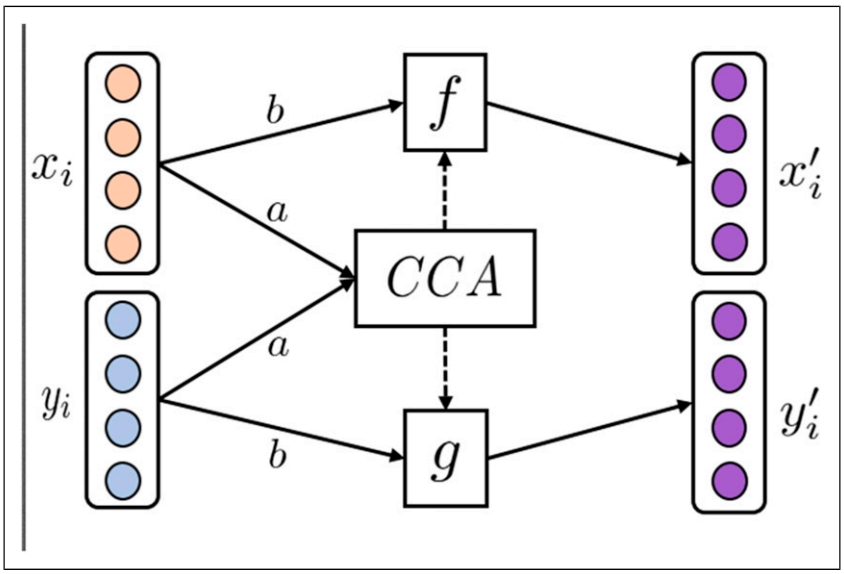

Figure 5. CCA technique of AMRL. (a) CCA finds linear transformations, f and g for uni-modal samples x_i and y_i that maximize their projected correlation. (b) The linear transformations f and g are used to project uni-modal samples into the new correlation-optimized aligned space.

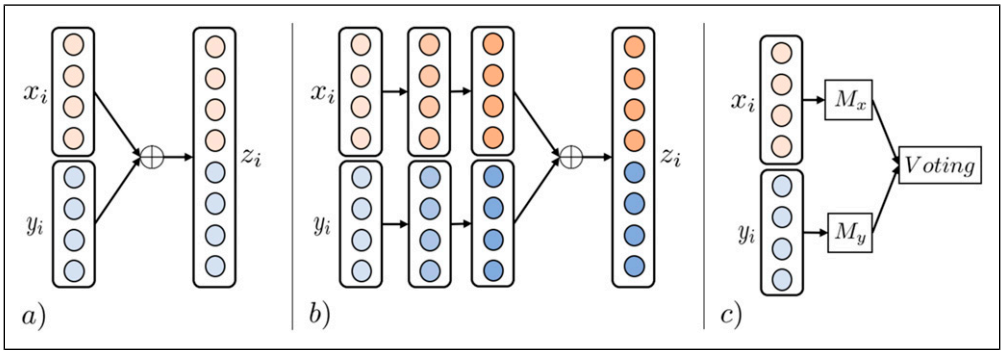

Figure 6. Joining MRL techniques. (a) Illustrates EJ FMRL. Features of uni-modal samples x_i and y_i are concatenated through $\oplus$ to form the fused representation z_i. (b) Illustrates IJ FMRL, where uni-modal modalities x_i and y_i are first processed individually. Later these are concatenated through $\oplus$. (c) Illustrates DJ FMRL. Uni-modal modalities x_i and y_i are processed through disparate models M_x and M_y. Finally, a voting mechanism is applied to the outputs of the individual models.

learning can be achieved by linear, non-linear, or weighted combinations of the resulting modality-specific kernel transformations.

Graphical Models are a class of probabilistic ML techniques used to discover latent factors explaining the data distribution. Among the most common graphical models for MRL is the *multimodal deep Boltzmann machine (MDBM)*.[27] An MDBM stacks layers of fully connected restricted Boltzmann machines to form a multi-layer network structure for each modality, which are subsequently joined by an output layer.

Neural MRL (NMRL). *Neural Architectures* aim to learn to join representation spaces for multimodal data in supervised, semi-supervised, or unsupervised ways. An idea shared among all architectures is learning layers of non-linear transformations for fusing uni-modal representations into a multimodal representation space guided by optimizing a loss function.[28] The basis of neural network architectures is the perceptron. The perceptron contains a learnable transformation matrix to linearly transform incoming data modalities into a new representation space, subsequently exercising non-linearity by applying an activation function such as sigmoid.

Concatenation (Concat) is the most straightforward neural architecture for multimodal data fusion.[28] Multiple layers of fully connected perceptrons are used to ultimately fuse uni-modal representations in either early, intermediate, or late layers of the network structure (Figure 7(a)). In ref,[29] a fully connected neural network structure predicts patient diagnosis codes by concatenating patient medication prescription history with demographic information.

An *Auto Encoder (AE)* is an unsupervised architecture that utilizes a reconstruction loss to learn low-dimensional entity representations that capture most of the original modality information.[30] Multimodal AE architectures have three stages (Figure 7(b)). Modality-specific networks transforming uni-modal modalities are initiated and then joined by an intermediary layer that acts as the fused modality representation. The last stage splits the intermediary layer into uni-modal networks trained on a reconstruction loss between the final representations x_i' and y_i' and the initial representations x_i and y_i.

A *Convolutional Neural Network (CNN)* is a technique for learning representations of imaging modalities. Due to the essential domain-specific information that images contain, CNNs are often

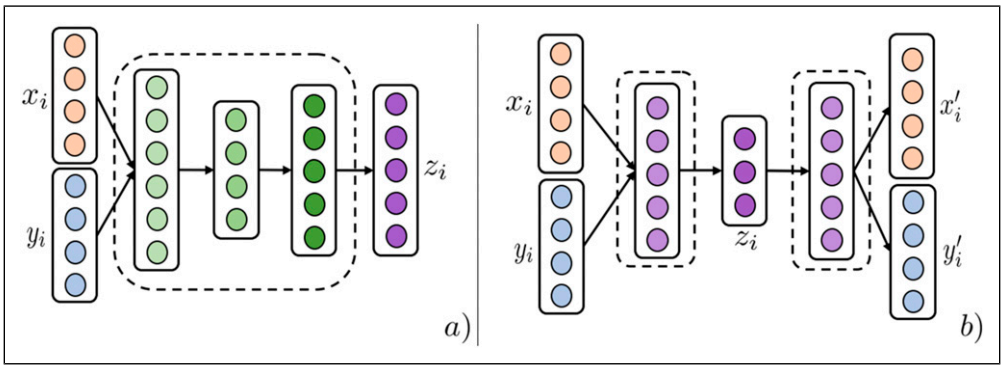

Figure 7. Neural MRL techniques 1. (a) Illustrates the Concat technique of NMRL. Uni-modalities are fused by layers of non-linear transformations between input modalities x_i and y_i and the fused output space z_i. Arrows between two neural layers represent the existence of a connection between each neuron/input from a layer to each neuron/output of the next layer. This is true for all neural architectures. (b) Illustration of the AE neural network structure. Uni-modal features x_i and y_i are transformed through multiple fully connected neural layers. The AEs middle layer learns a low-dimensional fused representation of uni-modalities z_i by training the neurons using a reconstruction loss between the original modalities x_i and y_i and their corresponding reconstructed representations x_i' and y_i'.

used to learn low-dimensional image representations in end-to-end architectures (Figure 8). CNNs apply layers of convolution matrices and pooling operations to condense images to their essential discriminative features. Due to their properties, they are often used as an intermediary step of imaging processing with subsequent fully connected layers fusing uni-modal entities.[8]

Transformers (TF) specialize in learning to represent sequence data. Utilizing a powerful component called self-attention, the model learns the relationships between different parts of the input sequence. This allows the model to attend to specific parts of the input sequence while learning a latent representation for each part of the sequence. This technique can be extended to multimodal networks by using the learned self-attention weights from one modality in the self-attention mechanism of other modalities. In ref,[31] skin lesion diagnosis is performed using an end-to-end transformer neural network to learn a latent representation between images and clinical features. In ref,[32] hospitalization length of stay was predicted based on the first 24 h of observations modelled as event sequences, effectively combining multiple clinical modalities into the same model.

Review methodology

A systematic search was done using the PubMed search engine[5], searching for MRL articles targeting medical analytics tasks while following the PRISMA guidelines[33] for structured surveys, as summarized in Figure 9. Our search terms were *("joint" OR "fusion" OR "coordinated" OR "alignment") AND ("multimodal" OR "multi-view") AND ("machine learning" OR "deep learning" OR "representation learning")* and *("different modalities" OR "multiple modalities") AND ("machine learning" OR "deep learning" OR "representation learning")*. Studies were excluded on criteria as summarized in the screening and eligibility steps of the PRISMA guidelines (Figure 9), excluding inaccessible records, papers that did not employ MRL or did not attempt to perform a medical analytic task, and survey papers. Eventually, 146 eligible publications were

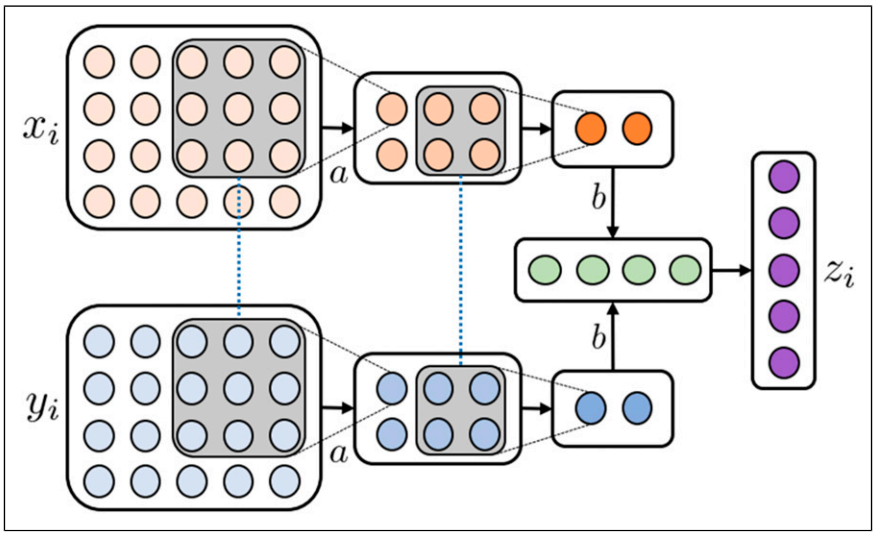

Figure 8. Illustration of the NMRL CNN technique. (a) A 3 × 3 convolution matrix with shared weights (as indicated by dotted blue lines) slide over the two input modalities x_i and y_i. (b) - When sufficiently condensed, features from both modalities can be appended for further processing.

identified. Since the nature of this paper is a literature review, it is not registered in a medical review registry.

The analysis employs four dimensions, *Utility*, *Medical*, *Modality*, and *MRL,* to better structure the analysis and put the many surveyed papers into a medical and algorithmic context. The utility dimension characterizes the analytic utility or purpose of the developed model. The medical dimension uses the ICD-10[34] diagnosis classification hierarchy to describe the analytics task's medical domain. The modality dimension uses our medical information modality hierarchy as introduced above. The MRL dimension (Figure 10) describes the MRL technique employed. Our primary measure of interest is the number of papers in a specific intersection of dimension values, such as how many papers have used joining MRL for combining structured patient data and MRI. A single reviewer (one of the authors) performed the entire review, consulting the other authors on reviewed papers they deemed difficult to classify. The authors decided the inclusion and exclusion criteria jointly based on the pilot survey examples.

Based on the structured survey and our four classification dimensions, an explorable online analysis tool is provided together with this work as an electronic supplement.[6] The analysis was compiled on the Tableau public platform and shared using the same platform as an online tool. The tool can be used to investigate the publications included in this review on our four classification dimensions and provide visual representations of findings.

Results

In this section, the findings are presented. The first section explores the pairs of modalities observed. We then examine the prevalence of MRL techniques in disparate medical fields and for different modality types. The last section examines the results from the perspective of medical analytics tasks.

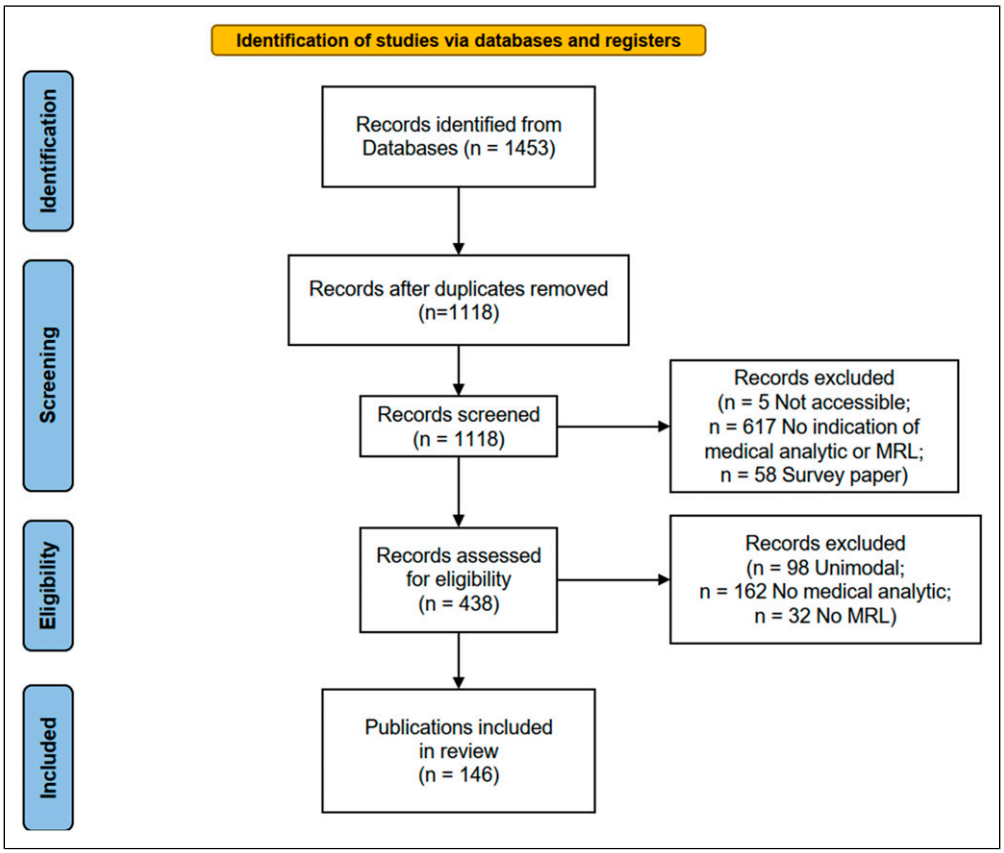

Figure 9. PRISMA flow chart for reporting systematic reviews.

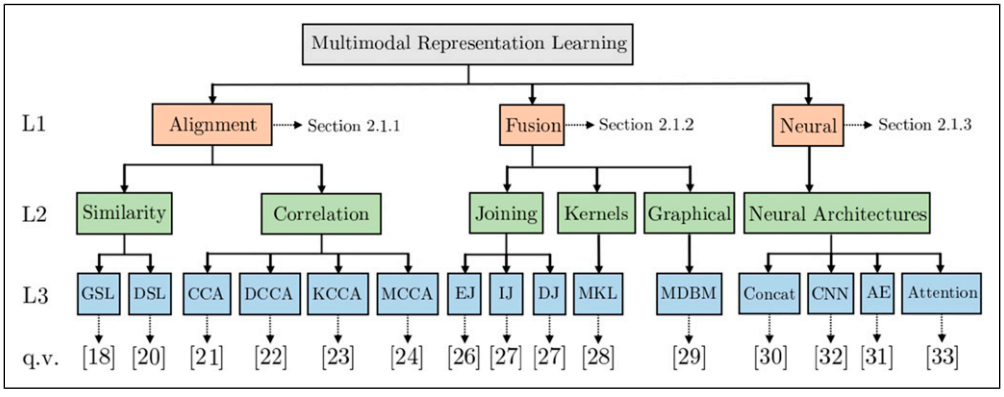

Figure 10. Classification space of reviewed MRL techniques used in medical analytics.

Modality pairings

Figure 11 presents the frequency of MRL applications utilizing level 2 (L2) MRL modalities (Figure 10) combined with level 1 (L1) modalities (Figure 1). As illustrated, imaging modalities are often combined with other unstructured modalities, specifically other imaging modalities. This tendency is primarily due to medical brain imaging applications, such as AD classification, utilizing the distinct discriminative properties of disparate medical imaging technologies. This insight is verified when drilling down into the darkest box (representing imaging-imaging pairings) in Figure 11 using the Level 3 (L3) modality level. One can see (Figure 12) that many of these pairs involve PET and MRI scans often utilized in brain studies. A more direct verification can be achieved by adding the medical dimension to this diagram (Figure 13), where one can see that an overwhelming majority of MRI and PET modalities are used as part of a mental or nervous system analytics task.

MRL techniques

A hierarchical analysis of the MRL techniques used in literature Figure 14 shows that the majority (46.7%) uses neural MRL. Further drill-down into L3 is available in the explorable online analysis.[7]

A few results stand out when comparing modalities and MRL techniques (Figure 15). While neural architectures and joining techniques are evenly used, joining techniques are more prevalent in time-series data and lab results. For time series, this amounts to 75% of the papers, while in lab results, over 70% of the papers utilize MRL joining techniques. Frequently, medical time series data has an immense sampling frequency, leading to hundreds of observations per second. Although data transformations can be learned directly for raw time-series data using deep learning techniques, such as the artificial recurrent neural network (RNN), the significant sampling rate of modalities like electroencephalography (EEG) can pose algorithmic problems, e.g., processing time, due to the sheer extent of raw data. This could explain why most time-series data is processed uni-modally and then fused with other modalities using joining.

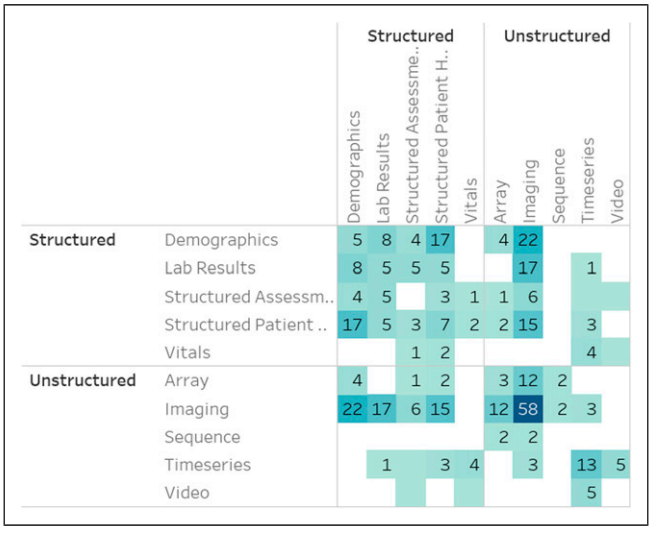

Figure 11. Number of papers by level 2 modality combinations used.

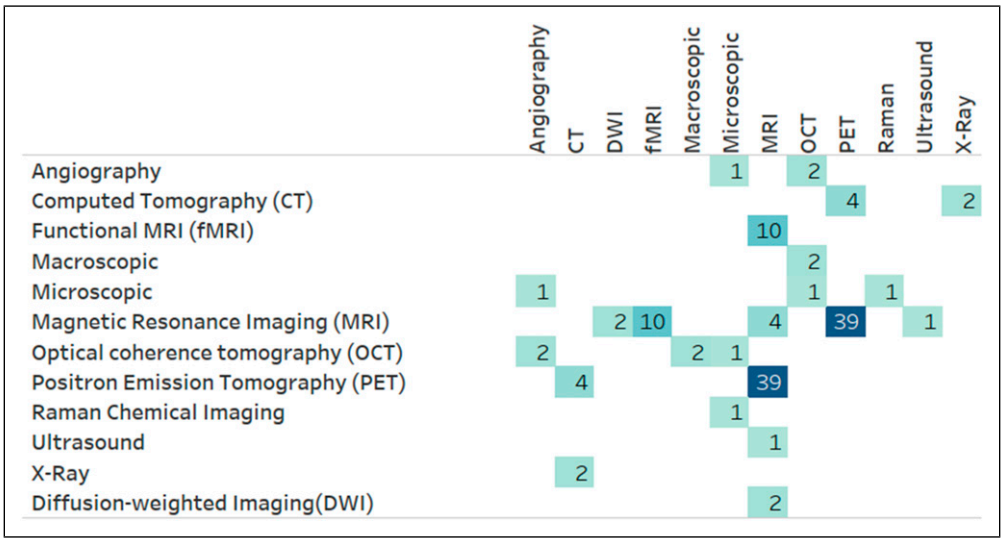

Figure 12. Number of papers by level 3 modality combinations, limited to imaging modalities.

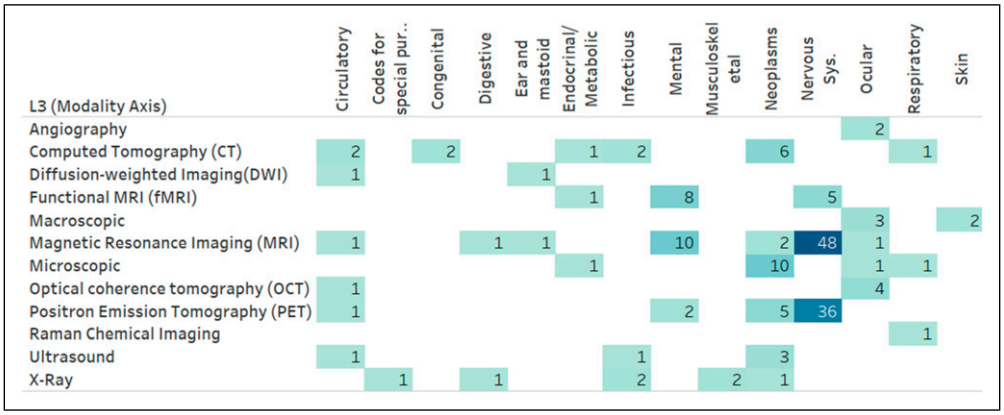

Figure 13. Number of papers by level 3 modality and level 1 ICD-10 category, limited to imaging modalities.

Medical analytics tasks

Table 1 lists the number of publications by medical task (ICD 10 code level 1) in descending order. Category names were shortened for brevity. Thus, *diseases of the eye and adnexa* became *Ocular*. Most publications attempt to identify conditions in the nervous system, most commonly the brain itself, as evidenced by 75 of 146 papers being of the nervous system or mental disease categories. Of the remaining categories, neoplasms receive most of the attention, which is an expected result, given that most of the MRL papers center around imaging modalities. It is somewhat surprising that circulatory system diseases are not commonly addressed, as it is a significant focus of medical AI research, specifically using imaging modalities.[35]

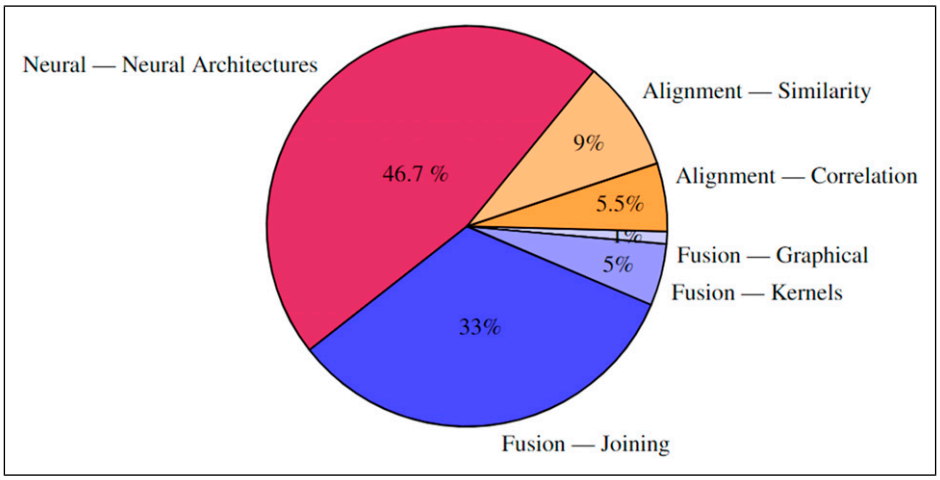

Figure 14. Percentage of papers by level-two MRL technique. Techniques in red/orange employ NMRL and CMRL respectively. The rest of the techniques employ FMRL. Level one followed by level two MRL class is shown for each group of papers.

		MRL L2			
L1 (Modality Axis)	L2 (Modality Axis)	Graphical Models	Joining	Kernels	Neural Architectures
Structured	Demographics		43%		57%
	Lab Results		71%	6%	24%
	Structured Patient ..		43%		57%
Unstructured	Array		25%		75%
	Imaging	3%	35%	12%	51%
	Timeseries		75%		25%

Figure 15. Percent of papers utilizing a level 2 MRL technique by level 2 modality. Results are limited to modalities with over 15 papers and to fusion MRL techniques.

Analysis of the type of analytical tasks derived from the MRL revealed that the only two types used were predictive[17] and descriptive (128), as illustrated in Figure 16.

Discussion

Across all dimensions, it is clear that the use of MRL is clustered around specific use cases and techniques. Numerous unexplored areas are available for future research. These under-explored areas can be identified and analysed in more detail using the online analysis tool provided by this paper. In the following, some dimension-specific findings are further discussed.

Table 1. Number of papers by medical task (ICD10 top-level category). Categories with no papers are omitted.

ICD-10 Cat	#Papers
Nervous sys	55
Neoplasms	28
Mental	20
Circulatory	8
Ocular	6
Other	4
Endocrinal/metabolic	4
Infectious	3
Musculoskeletal	3
Digestive	3
Respiratory	3
Congenital	2
Skin	2
Injury	2
Ear and mastoid	1
Special purpose	2

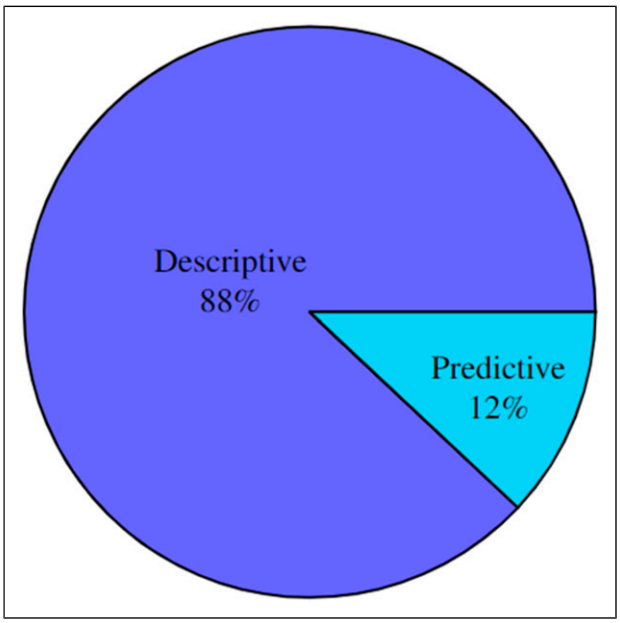

Figure 16. Papers by analytic type.

With respect to the information dimension (Figure 11), notice the significantly limited utilization of *structured data* together with imaging and other unstructured modalities such as Video, sequences, and time-series. This could indicate opportunities for future research, as simple structured data, such as demographics, diagnosis codes, and prescriptions, have been shown to increase the discriminative power in multiple medical MRL tasks.[36,37]

Concerning the medical task dimension, results show the untapped potential of MRL, especially for the less-investigated disease categories of ICD-10, such as dermatological and circulatory diseases.

The results in the MRL techniques dimension show that *Neural methods* have recently received increased interest. These techniques can learn models directly from labelled data instead of human-engineered feature extraction and modelling techniques. However, the amount of labelled data needed for training medical analytics in an end-to-end practice exceeds what is readily available for many medical analytical questions. While some medical analytics tasks have large datasets accessible for immediate consumption in model creation, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[8], researchers are mostly faced with a dearth of annotated datasets[38]). In the absence of large amounts of data, or conversely, when needing to process high-frequency sampled data, the more traditional, join-based techniques seem to be the tool of choice for the foreseeable future.

Limitations

As a time-bound review, this work is limited to the papers reviewed prior to submission. However, the authors hope our classification approach and online tool can be reused in future reviews to provide a clear picture of the current state of the art. The use of keyword-based retrieval for populating the initial list of papers is a limited technique, as we may have overlooked semantically similar keywords or papers using terminology not picked up by this method but still performing MRL in the medical domain.

Conclusion

This work comprehensively reviews the use of Multimodal Representation Learning (MRL) in medical analytics. A novel hierarchical taxonomy of medical information modalities is provided and linked to the SNOMED concept hierarchy and a hierarchy of related MRL techniques. Subsequently, a literature review of more than 1400 papers, following the PRISMA guidelines for structured surveys, is performed, and the eligible papers are inserted into four orthogonal classification dimensions: *utility*, *medical*, *modality*, and *MRL*. Using these classifications, a free and publicly available explorable online analysis is made available to investigate what modalities have been used together, for which medical analytics, and which MRL techniques have been successful in which combinations. In addition, as an electronic supplement to this paper, the complete list of works reviewed with exclusion reasons for excluded papers and classifications for included papers is provided.

Few ICD-10 top-level category disease codes were found to be the primary target for multimodal medical analytics. Many ICD-10 classes had few or no cases of medical analytics using multimodal data. This result could be due to the scarcity of openly available labelled training data for medical analytics, forcing MRL research to progress where such data is readily available.

While some medical information modalities can be integrated using most MRL techniques, modalities like time-series need special attention when utilizing end-to-end learning using neural

architectures to mitigate high sampling rates. Furthermore, investigations of the utility dimension show that most medical applications have been developed for descriptive analytics and only a few for predictive analytics. This finding suggests that we are still in the early phase of adopting ML for medical analytics and opens the door for future work in developing prescriptive utility analytics.

Acknowledgements

The authors wish to acknowledge Kashif Rabbani, Theis Jendal, and the rest of the Data, Knowledge, and Web research group at Aalborg University for their advice and support.

Author contributions

Hansen: Writing - Original Draft, Investigation, Visualization; Hose: Supervision, Writing - Review and Editing, Funding Acquisition; Sagi: Conceptualization, Visualization, Supervision, Writing - Review and Editing.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by the Poul Due Jensen Foundation.

ORCID iDs

Emil Riis Hansen https://orcid.org/0000-0003-4103-1244
Tomer Sagi https://orcid.org/0000-0002-8916-0128
Katja Hose https://orcid.org/0000-0001-7025-8099

Data availability statement

All data collected during this work, including the complete list of surveyed papers and all analyses are available in the online analysis tool at https://tabsoft.co/40aAECd.

Supplemental Material

Supplemental material for this article is available online. The complete list of papers is provided as an electronic supplement. All analyses are available at https://tabsoft.co/40aAECd.

Notes

1. https://bioportal.bioontology.org/ontologies/SNOMEDCT.
2. https://www.meddra.org/ - MedDRA® trademark is registered by ICH.
3. https://bioportal.bioontology.org/ontologies/BIM.
4. https://tabsoft.co/40aAECd.
5. https://pubmed.ncbi.nlm.nih.gov/.
6. https://tabsoft.co/40aAECd.
7. https://tabsoft.co/40aAECd.
8. https://www.loni.ucla.edu/ADNI.

References

1. Killiany RJ, Gomez-Isla T, Moss M, et al. Use of structural magnetic resonance imaging to predict who will get Alzheimer's disease. *Ann Neurol* 2000; 47(4): 430–439.
2. Coleman RE. Positron emission tomography diagnosis of Alzheimer's disease. *Pet Clin* 2007; 2(1): 25–34.
3. Blennow K. Cerebrospinal fluid protein biomarkers for Alzheimer's disease. *NeuroRx* 2004; 1(2): 213–225.
4. Pillai PS and Leong TY. Fusing heterogeneous data for Alzheimer's disease classification. In MEDINFO 2015: eHealth-enabled Health - Proceedings of the 15th World Congress on Health and Biomedical Informatics. São Paulo, Brazil, 19–23 August 2015, Vol 216, pp. 731–735. (Studies in Health Technology and Informatics).
5. Zhang D, Wang Y, Zhou L, et al. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* 2011; 55(3): 856–867.
6. Baltrusaitis T, Ahuja C and Morency LP. Multimodal machine learning: a survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell* 2019; 41(2): 423–443.
7. Cai Q, Wang H, Li Z, et al. A survey on multimodal data-driven smart healthcare systems: approaches and applications. *IEEE Access* 2019; 7: 133583–133599.
8. Guo W, Wang J and Wang S. Deep multimodal representation learning: a survey. *IEEE Access* 2019; 7: 63373–63394.
9. Wang W, Arora R, Livescu K, et al. On deep multi-view representation learning. In: JMLR Workshop and Conference Proceedings ICML. JMLR.org, New York, NY, June 24, 2016, Vol. 37, pp. 1083–1092.
10. Zhang SF, Zhai JH, Xie BJ, et al. Multimodal representation learning: advances, trends and challenges. In 2019 International Conference on Machine Learning and Cybernetics (ICMLC), Kobe, 07-10 July 2019, pp. 1–6. IEEE.
11. Ramachandram D and Taylor GW. Deep multimodal learning: a survey on recent advances and trends. *IEEE Signal Process Mag* 2017; 34(6): 96–108.
12. Li Y, Yang M and Zhang Z. A survey of multi-view representation learning. *IEEE Trans Knowl Data Eng* 2018; 31(10): 1863–1883.
13. Pospieszny P. Software estimation: towards prescriptive analytics. In: Staron M and Meding W (eds). Proceedings of the 27th International Workshop on Software Measurement and 12th International Conference on Software Process and Product Measurement, IWSM-Mensura 2017. Gothenburg, Sweden, October 25 - 27, 2017, pp. 221–226.
14. Moreno RP, Metnitz B, Adler L, et al. Sepsis mortality prediction based on predisposition, infection and response. *Intensive Care Med* 2008; 34(3): 496–504.
15. Wang L, Zhang W, He X, et al. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018. pp. 2447–2456.
16. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med* 2019; 25(1): 24–29.
17. Weston J, Bengio S and Usunier N. WSABIE: scaling up to large vocabulary image annotation. In The International Joint Conferences on Artificial Intelligence, Inc. (IJCAI), Barcelona, Spain, July 16 - 22, 2011, pp. 2764–2770.
18. Frome A, Corrado GS, Shlens J, et al. DeViSE: a deep visual-semantic embedding model. In NIPS, Lake Tahoe, December 5 - 10, 2013, pp. 2121–2129.
19. Kiros R, Salakhutdinov R and Zemel RS. *Unifying visual-semantic embeddings with multimodal neural language models*. arXiv preprint arXiv:14112539, 2014.

20. Hotelling H. Relations between two sets of variates. In *Breakthroughs in statistics*. Berlin: Springer, 1992, pp. 162–190.
21. Andrew G, Arora R, Bilmes JA, et al. Deep canonical correlation analysis. In: JMLR Workshop and Conference Proceedings ICML (3). JMLR.org, New York, NY, 2013, Vol. 28, pp. 1247–1255.
22. Hardoon DR, Szedmak S and Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput* 2004; 16(12): 2639–2664.
23. Kettenring JR. Canonical analysis of several sets of variables. *Biometrika* 1971; 58(3): 433–451.
24. Zhuang X, Yang Z and Cordes D. A technical review of canonical correlation analysis for neuroscience applications. *Hum Brain Mapp* 2020; 41(13): 3807–3833.
25. Atrey PK, Hossain MA, El Saddik A, et al. Multimodal fusion for multimedia analysis: a survey. *Multimed Syst* 2010; 16(6): 345–379.
26. Gönen M and Alpaydn E. Multiple kernel learning algorithms. *J Mach Learn Res* 2011; 12: 2211–2268.
27. Srivastava N and Salakhutdinov R. Multimodal learning with deep Boltzmann machines. In: Bartlett PL, Pereira FCN, Burges CJC, et al. (eds). Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012 Proceedings of a Meeting Held December 3-6, 2012. Lake Tahoe, Nevada, United States, 2012, pp. 2231–2239.
28. Gao J, Li P, Chen Z, et al. A survey on deep learning for multimodal data fusion. *Neural Comput* 2020; 32(5): 829–864.
29. Hansen ER, Sagi T, Hose K, Lip GYH, Larsen TB and Skjøth F. Assigning diagnosis codes using medication history. *Artif Intell Med*. 2022; 128: 102307. doi:10.1016/j.artmed.2022.102307.
30. Jaques N, Taylor S, Sano A, et al. Multimodal autoencoder: a deep learning approach to filling in missing sensor data and enabling better mood prediction. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, 2017, pp. 202–208.
31. Ou C, Zhou S, Yang R, et al. A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata. *Front Surg* 2022; 9: 1029991.
32. Hansen ER, Nielsen TD, Mulvad T, et al. Patient event sequences for predicting hospitalization length of stay. In: Juarez JM, Marcos M, Stiglic G, et al. (eds). *Artificial intelligence in medicine*. Cham: Springer Nature Switzerland, 2023, pp. 51–56.
33. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int J Surg* 2021; 88: 105906.
34. WHO. *The international statistical classification of diseases and health related problems ICD-10: tenth revision. Volume 1: Tabular List. Vol. 1*. Geneva: World Health Organization, 2004.
35. Briganti G and Le Moine O. Artificial intelligence in medicine: today and tomorrow. *Front Med* 2020; 7: 27.
36. Cheerla A and Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* 2019; 35(14): i446–i454.
37. Jin M, Bahadori MT, Colak A, et al. Improving hospital mortality prediction with medical named entities and multimodal learning. CoRR 2018: 12276, abs/1811. ML4H: Machine Learning for Health https://ml4h.cc/2018/pages/papers.html
38. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019; 25(9): 1337–1340.
39. Song J, Zheng J, Li P, et al. An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis. *Front Digit Health* 2021; 3: 19.