Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(5): 8708–8726.
DOI: 10.3934/mbe.2023382
Received: 30 September 2022
Revised: 15 February 2023
Accepted: 15 February 2023
Published: 06 March 2023

Review

A review on multimodal machine learning in medical diagnostics

Keyue Yan[1,*], Tengyue Li[1], João Alexandre Lobo Marques[2], Juntao Gao[3] and Simon James Fong[1,4,*]

[1] Department of Computer and Information Science, University of Macau, Macau SAR, China
[2] Laboratory of Applied Neurosciences, University of Saint Joseph, Macau SAR, China
[3] Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
[4] Institute of Artificial Intelligence, Chongqing Technology and Business University, Chongqing, China

* **Correspondence:** Email: yc17928@um.edu.mo, ccfong@um.edu.mo.

Abstract: Nowadays, the increasing number of medical diagnostic data and clinical data provide more complementary references for doctors to make diagnosis to patients. For example, with medical data, such as electrocardiography (ECG), machine learning algorithms can be used to identify and diagnose heart disease to reduce the workload of doctors. However, ECG data is always exposed to various kinds of noise and interference in reality, and medical diagnostics only based on one-dimensional ECG data is not trustable enough. By extracting new features from other types of medical data, we can implement enhanced recognition methods, called multimodal learning. Multimodal learning helps models to process data from a range of different sources, eliminate the requirement for training each single learning modality, and improve the robustness of models with the diversity of data. Growing number of articles in recent years have been devoted to investigating how to extract data from different sources and build accurate multimodal machine learning models, or deep learning models for medical diagnostics. This paper reviews and summarizes several recent papers that dealing with multimodal machine learning in disease detection, and identify topics for future research.

Keywords: multimodal learning; machine learning; deep learning; medical data

1. Introduction

The rapid development of science and technology are changing everyone's life all over the world. Large amounts of new inventions have been making a huge impact on our research in the 21st century [1]. Among these inventions, machine learning techniques are the most famous and widely applied.

However, machine learning is not the latest technology since it is the intersection of computing and statistics [2]. In recent years, the proliferation of data from a wide range of industries has provided the opportunity for machine learning to implement widely in education, finance, economics, smart cities and medical areas. Further, the swarm intelligence algorithms are widely used and applied to resolve different optimization problems in machine learning [3]. Also, machine learning is a scientific system encompassing a variety of different classes of techniques, and it can be trained and refined to make accurate predictions based on the information and data in its environment. When application scenarios and data sources have changed, machine learning can be retrained and applied again.

In the medical area, machine learning can be applied to healthcare and medical diagnostics [4]. Since people prefer the performance of ther own accurate results from diagnostics models, the quality of the data is particularly important for training and testing models in medical diagnostics. Except for traditional machine learning models, multimodal machine learning models are dominating by collecting data from different aspects of the patient with the development of science and technology. When fuzing multiple data, multimodal machine learning models make medical diagnostics more accurate, predictable and interpretable. During the medical examination in reality, a patient (sample) has a variety of clinical tests, which will always generate a wide range of data, including basic personal information, such as testing number, name, gender, age, height, weight, etc., as well as some numerical medical testing results and records, such as blood type, blood pressure, BMI, microelements in the body and so on. Specialized imaging data, such as electrocardiography (ECG), magnetic resonance imaging (MRI), facial expressions and body postures also plays irreplaceable roles in the diagnosis. Doctors and health workers often make medical diagnostics based on different types of data from the patient. However, this process is often biased towards the doctors or health workers' experience and subjectivity, while the large number of patients may put a lot of pressure on diagnostic efficiency. Nowadays, all aspects of data listed above are used as the inputs of multimodal machine learning models, which are developing and extending well. A number of papers have been devoted to the application of multimodal learning to the diagnosis of cardiovascular disease, Alzheimer's Disease (AD) and stress detection. We will review the research in detail.

In this review, all papers use multimodal machine learning or deep learning methods for medical diagnostics. The remainder of this review can be summarized as follows: Section 2 briefly describes the data preparation, preprocessing and feature extraction for different medical data in reality, which is the key to multimodal machine learning models. Section 3 describes the machine learning classification algorithms used in these papers. In Section 4, we present the modeling process and framework outlines for multimodal machine learning. Also, a discussion about multimodal machine learning in real life application is given in Section 5. Finally, we conclude the review and give out future research direction.

2. Data preparation and preprocessing

Datasets are the source of all machine learning and deep learning models, where models with complete and perfect data are likely to have better results in testing situation. In this review, we present the datasets for different medical diagnoses, such as Alzheimer's Disease (AD), Parkinson's disease, cardiovascular disease and stress tests.

Although some medical diagnostics research about machine learning or deep learning is now based on single data, multimodal learning that combines clinical data with strong connection and correlation

also have excellent results. Since the data in different modalities are in different formats, we need to pre-process and select features from them before substituting all features into machine learning models. In the following parts, we will describe several types of data that machine learning models will use for different diseases, and how the datasets are pre-processed. A summary table of the dataset of research papers in this review is shown in Table 1.

Datasets in current medical diagnostics research can be divided into three main categories: clinical characteristics, time series and image data. Detailed descriptions and applications of these three data types are discussed in this section. Clinical characteristics contain basic patient data, and the data structure is usually in the form of structured data. In Section 2.1, we present information on individual cases and specialist questionnaires that are commonly used as data sources for different pathology studies. The time series data is predominantly of various waveform types. In this review, we select medical diagnostics research based on ECG signals as a dataset for machine learning models, which can either be used for predictive classification using sequence-based models or transformed into images as input for machine learning. In Section 2.2, we present the pre-processing and prediction process of ECG time series data in detail, which includes the noise reduction decomposition of ECG time series, the identification of QRS waves, and a brief introduction to using ECG as an image input to CNN. Image datasets can be divided into radiological images, pathological images and camera images. The most common MRI images are described with more details in Section 2.3.

2.1. Clinical characteristics

Clinical characteristics include data that describe basic personal information, measurements of vital signs of human bodies, physiological tests, disease rating scales and so on. These datasets are often expressed as numerical or categorical variables after the features have been extracted from the raw data. Typical variables of clinical data are listed as follows:

2.1.1. Personal information

- age [7, 13]
- gender [7, 13]
- education level [7, 10]
- number of subject visits, injury history, surgery history [10, 13]
- socioeconomic status [10]
- general health information, such as height, weight, Body Mass Index (BMI) and so on [7, 13]
- complete blood count, comprehensive metabolic examination, hepatitis C co-infection. [7]

2.1.2. External information of human bodies

When testing participants' stress, some researchers try to collect data of face expressions and body posture. For face expressions, a camera has been used to record the face and upper body of the patients. Researchers use a software called *FaceReader* to present the data, and this software analyzes facial expressions in real time, and face expression details are provided. Ultimately, *FaceReader* provides data on more than 30 corresponding expressions, such as head orientation, facial expressions, action units and emotions [9].

Table 1. Overview of data source, type and capture method.

Paper	Source	Type	Capture method
Aziz et al. [5]	stride interval	time series	open resource
Hussain et al. [6]	ECG	time series	open resource
Xu et al. [7]	clinical data, MRI images	numerical, image	MRI scanner et al.
Naik et al. [8]	/	/	/
Walambe et al. [9]	interval data, facial images	time series, image	camera, sensor et al.
Battineni et al. [10]	clinical data, MP-RAGE images	numerical, image	vision scanner et al.
Anand et al. [11]	MRI images	image	open resource
Khan et al. [12]	MRI images	image	open resource
Tiulpin et al. [13]	clinical data, Knee X-ray images	numerical, image	Osteoarthritis et al.
Prashanth et al. [14]	clinical data,		
questionnaire	numerical	PPMI database	
Ieracitano et al. [15]	EEG	time series	sensor
Zhao et al. [16]	ECG, facial video	time series, video	open resource, video
Ma et al. [17]	ECG	time series	open resource
Ramkumar et al. [18]	ECG	time series	open resource
Arteaga-Falconi et al. [19]	fingerprint, ECG	image, time series	open resource
Ahmad et al. [20]	ECG images	image	open resource
Irfan et al. [21]	ECG	time series	open resource
Zeng et al. [22]	ECG	time series	hospital database
Song et al. [23]	ECG	time series	CCDD database
Su et al. [24]	finger vein, ECG	image, time series	veinPolyU, ECG-ID
El-Rahiem et al. [25]	finger vein, ECG	image, time series	veinPolyU, MWM-HIT
Hammad et al. [26]	finger vein, ECG	image, time series	PTB, CYBHi
Bugdol et al. [27]	ECG	time series	sensor
Ketu et al. [28]	ECG	time series	open resource
Alkeem et al. [29]	ECG, facial image et al.	image, time series	open resource
Rahul et al. [30]	ECG	time series	open resource

The dataset of body posture contains over 90 features in an Excel file, including the coordinates necessary to determine angles between upper-body joints and bones, and upper-body bone orientations by fitting the Kinect skeletal model [9].

2.1.3. Rating scales

Different authoritative questionnaires are administered to assist in medical diagnostics. In this review, we can find surveys on scales for Alzheimer's Disease (AD), Parkinson's Disease and other diseases. The data composed of these scales can also be used as inputs for multimodal machine learning to improve the accuracy of medical diagnostics.

Dementia status in AD is assessed by the Clinical Dementia Rating (CDR) scale, which rates the patient's level of impedance in each of six domains: memory, orientation, judgment and critical

thinking, community work, home and hobbies, and individual care. The CDR score is added from a single number rating for each domain. A CDR of 0 indicates no dementia, and CDR of 3 means severe dementia [10].

To identify Parkinson's Disease, the 40-item University of Pennsylvania Smell Identification Test (UPSIT) is widely used. It is a 40-pages booklet, and each page contains a different theodor in a plastic microcapsule. Each theodor is identified by marking the option that describes the theodor truly from the four options. The more theodors that are correctly identified, the higher the scores will be attained for the participants [14]. In addition, the REM sleep Behavior Disorder Screening Questionnaire (RBDSQ) is developed to assess the most salient clinical features of RBD. Researchers have studied the utility of the RBDSQ, and have observed that it has a high sensitivity and reasonable specificity with questionnaire options answered in a choice of "yes" or "no". The higher the scores, the more likely they are to have Parkinson's disease [14].

2.2. Electrocardiography (ECG) data

ECG, also called EKG in Dutch and German, is a media to show the electrical current flowing through people's heart. In detail, ECG clearly shows how depolarization flows in each heartbeat, where depolarization is a wave of positive electrodes. Now, ECG is used to classify cardiac arrhythmias and diagnose heart disease. However, it takes a lot of time for professional doctors to make medical diagnostics accurately. Many research groups have great interest in exploring and studying the important information contained in the ECG. With the development of computer science, using machine learning algorithms to detect and classify heart disease, not only can provide convenience to doctors, but can also give diagnostic results for patients quickly.

2.2.1. Preprocessing and feature selection

The raw data of ECG collected in real hospitals or clinics are easily disturbed by noise, which not only affects doctors to make medical diagnostics, but also decrease the accuracy and effectiveness of machine learning algorithms for classification. Therefore, many papers pre-process the ECG signal to make it clear and clean, and the feature selection extraction part is also presented. Feature selection is a basic and widely used technique for data processing before modeling. The increasing number of features may indicate that the data has more information, but it may consume more computing power, memory and time for machine learning. Useless features in the dataset can also reduce the performance of machine learning algorithms, and cause overfitting problems. Feature selection is the mapping of a larger dimensional sets of features to a smaller dimensional sets of features. In the following parts, we begin to describe Discrete Wavelet Transform (DWT), a QRS feature selection used in processing ECG signals.

DWT is a method of analyzing transforms to evaluate the position of a signal in time, space and frequency, and refine it over time by using an extension and translation process. A subdivision of high frequency time and low frequency time is achieved in the end. The time-frequency signal analysis will be automatically adapted to the user's needs. However, the ECG signal and the noise are often combined. First, a basis wavelet function is chosen to decompose the ECG signal with noise since decomposition generates wavelet coefficients. The wavelet coefficients with larger range are useful after decomposing, while the wavelet coefficients with moderate range are the noise in ECG signals.

In preprocessing, wavelet coefficients (which are smaller than the threshold value) are processed using threshold processing or threshold functions. After DWT, the low frequency coefficients and high frequency coefficients are processed to regenerate the ECG signal [17, 23, 30].

QRS complex waves are the main component of the ECG and represent ventricular depolarization. The amplitude, duration and shape of the QRS region can be used to determine the presence of arrhythmias. Thus, QRS detection is a subject being popularly studied. Normal QRS waves are sharp and narrow, which is shown in Figure 1 [24], while QRS waves with heart disease or cardiac arrhythmias may be wider or narrower. The types of features extracted from the QRS waves are RR intervals, means, variances, percentiles, maximum values, minimum values, kurtosis, skewness and other statistical variables. These features can be used as input variables for machine learning to facilitate the prediction of heart disease [17, 23–25, 27, 30].

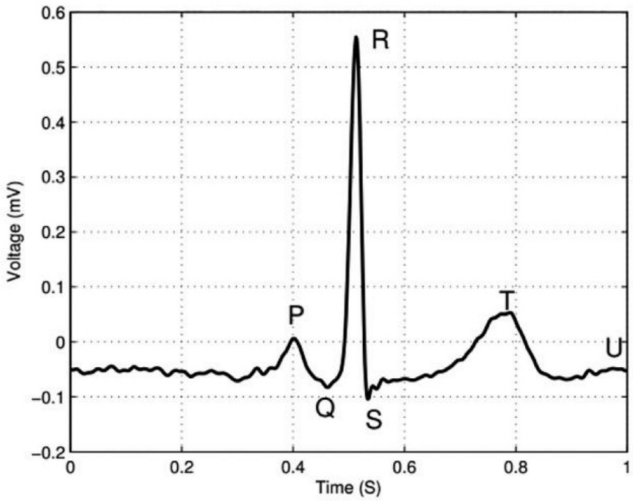

Figure 1. Normal ECG signal.

2.2.2. Images preprocessing

ECG signals can be transformed to ECG images for visual representation. When using ECG image data as variables, these papers do not remove the noise from the raw ECG data and convert the 1D ECG signals into 2D ECG images. Some authors in the literature have suggested that using noise elimination algorithms or segmentation on the original ECG algorithm might lose some key information in the data. For ECG data converted to images, a feature selection of ECG images is performed using a Convolutional Neural Network (CNN) model. 2D convolutional and pooling layers in a CNN are more suitable for filtering the locality of ECG images. Many CNN models have been widely developed for feature selection, such as Caffe-Net, Alex-Net, VGG-Net and Res-Net [26, 29]. In addition to using CNN algorithms directly on image data, other works have used ECG images formed by Gramian Angular Field (GAF), Recurrence Plot (RP) and Markov Transition Field (MTF). The use of these three different transcoding techniques to generate new images as variables to be input into the CNN has also generated good results in multimodal machine learning for classification [20].

2.3. Magnetic resonance imaging (MRI)

MRI is a technique using the principle of nuclear magnetic resonance to determine the location and type of nuclei, and draw images of the structure based on the attenuation of the energy emitted in different structural environments within the material. Structural MRI examines anatomical irregularities of the brain caused by traumatic events, while functional MRI is used to obtain images of the whole brain based on blood flow and oxygen levels, which are convenient to collect data that correlate with the usage of oxygen [8]. After acquiring the images from magnetic resonance imaging instruments, researchers should pre-process and make it possible to classify the diseases or problems inside images in a more specific way. Noise data is the most common type of problem that needs to be solved before MRI classification modelling, as these wide range of image artefacts can be removed by using a number of different kinds of image filters, such as a geometric mean filter [10, 11]. Furthermore, in order to obtain the important feature parts of images in the task of brain tumour identification, fuzzy c-means algorithms are often applied for segmenting the image into smaller parts. The segmentation facilitates the identification or classification of regions for different tasks. Also, the general linear correspondence model (GLCM) algorithm is used to extract features, such as contrast, correlation, entropy, and homogeneity from the photographs [10].

3. Multimodal machine learning algorithms

When datasets from different sources are used to train multimodal machine learning models, they should be transformed and fused into some unified pattern. Among multimodal machine learning models, both traditional machine learning and deep learning models are widely used, and when datasets are in the form of small dimension and amounts, the evaluation of traditional machine learning may be better than deep learning. Meanwhile, traditional machine learning can also avoid overfitting problems and achieve better classification results, effectively. In the following parts, we introduce the machine learning and deep learning that are widely used in multimodal learning. A summary of the model and sample size of research papers in this review is shown in Table 2.

3.1. Support vector machine (SVM)

Before the popularity of deep learning, there was a wide range of applications by the Support Vector Machine (SVM) algorithm, a powerful non-linear ML classifier that is often used to solve linear and non-linear classification and regression problems on ordinary datasets. SVM attempts to find a hyperplane to separate data, and also classifies features into a multidimensional space based on datasets distribution. The boundary on which the data is divided is known as the decision boundary of the hyperplane. The hyperplane is important to improve the performance of the SVM model to classify the datasets. Now, SVM is still widely used in multimodal learning tasks for medical diagnostics [5, 9, 11, 12].

3.2. Logistic regression (LR)

Logistic regression (LR) is a generalized linear model, and has almost the same formulations as multiple linear regression. However, LR is not a regression algorithm, but a classification algorithm. In predictive classification, the problem can be either or multi-categorical. The most commonly used

of LR is binary problem in practice. Before modelling, we assume that the training datasets follow continuous probability distributions, and the modeling process uses the maximum likelihood estimation (MLE) to estimate parameters of LR. Thus, LR requires a high probability distribution of the data, and the results achieved in practice are not always better than other machine learning algorithms [10, 13, 15, 28].

3.3. Bayesian classifiers

Bayesian classifiers is a general term for a class of classification algorithms based on Bayes' theorem. The Naïve Bayes algorithm is a simple probability classifier, assuming that each feature in the dataset is independent of each other. Other practical classifiers include the Gaussian Bayesian algorithm, the polynomial Bayesian algorithm and the Bernoulli Bayesian algorithm. More professional, the Gaussian Bayesian algorithm is suitable for datasets with continuous features, when the Bernoulli Bayesian algorithm is suitable for datasets with discrete features [5, 6, 10, 12, 14, 25, 30].

3.4. Decision tree (DT)

Decision tree is a well-known machine learning model for solving classification and regression problems. From a set of features, it essentially makes decisions . Decision tree in the algorithm starts at the root node and has a number of child nodes between each leaf node, which identify the data input values and move to different child nodes depending on the features until they reach the leaf node. Each input sample goes through this process and obtains its target value at the corresponding leaf node. The predicted value is the mean values of targets associated with leaf nodes [5, 6, 28, 30].

3.5. Ensemble learning

Ensemble learning algorithms improve the overall classifier accuracy by training multiple simple models, such as decision tree and SVM to make predictions. Ensemble learning can be broadly divided into two types, one is bagging, in which every simple model is independent and trained individually, and the other is boosting, in which each simple model has connections. The ensemble learning algorithms are used in the following multimodal machine learning articles. Random Forest (RF) is the most famous bagging algorithm, which operates by constructing multiple decision trees. Unlike building a single decision tree, Random Forest finds the best features to divide the sub-nodes, and selects features randomly, rather than all features in the nodes to build a decision tree. Further, the classification results of the algorithm are achieved by averaging the output of all decision trees [5, 10, 14, 23, 25, 28, 30]. For Gradient Boosting, one of the boosting algorithms, it is also based on decision trees and uses gradients to boost the structure [10, 13]. XGBoost algorithm was proposed by Chen and Guestrin in 2016. It was modified from Gradient Boosting, which combines a weak basic model with a stronger learning model through iterations [5]. In Adaboost, simple model is highly correlated with each other, and misclassified samples made by the previous sample model are used to train the next simple model. Although the simple model used in the Adaboost algorithm make classification inaccurately, it can improve the final model results as long as it is better than classifying randomly [10, 11, 28].

3.6. Deep learning

With the abundance of data forms, new types of data, such as sound, images and video records are becoming popular. Deep learning models are widely used for feature extraction and processing among these kinds of data. In this review, we find that Neural Network (NN) and Convolutional Neural Network (CNN) are used to extract features among these data, which are fuzed with other common features, and then substituted into new models for medical diagnostics.

NN is one of the most widely known deep learning models, which combines the knowledge of biological neural networks with mathematical statistical models. NN consists of neurons that are interconnected with each other. Each neuron represents a specific output function called the activation function. Each connection between two neurons represents a weighted value for the signal passing through that connection called the weight. The connection weights reflect the strength of the connections between units, and the representation and processing of information is reflected in the connection relationships. In a neural network, there are three types of processing units: input units, output units and hidden units. The input units receive data, while the output units implement regression or classification of results, and the hidden units (which are between the input and output units) cannot be directly observed outside [5, 9, 11, 15, 22, 23, 25, 27]. CNN emerged as a solution to the problem of neural networks' parameters being overloaded when classifying image data, the difficulty of training and the failure of using the information of image data. CNN consists of several types of layers: input layer, convolutional layer, ReLu layer, pooling layer and fully connected layer. In practice, the convolutional layer and the ReLu layer are called convolutional layers, so the convolutional layer has both a convolutional operation and the activation function. The parameters in the convolutional layers and normal layers are trained with gradient descent so that the classification labels computed by the convolutional neural network correspond to the labels of each image in the training set [13, 16–18, 20].

4. Modelling process of multimodal machine learning

According to another review paper [31], the modeling process of multimodal machine learning can be broadly divided into three categories, called early fusion, intermediate fusion and late fusion. A summary of the fusion styles and machine learning tasks in this review is shown in Table 3.

Table 2. Overview of model and sample size.

Paper	Year	Model	Sample size
Aziz et al. [5]	2020	SVM, Bayesian, DT, Ensemble learning, NN	10
Hussain et al. [6]	2020	Bayesian, DT	72
Xu et al. [7]	2020	Lasso	101
Naik et al. [8]	2020	/	/
Walambe et al. [9]	2021	SVM, NN	25
Battineni et al. [10]	2021	LR, Bayesian, Ensemble learning	150
Anand et al. [11]	2022	SVM, Ensemble learning, NN	100
Khan et al. [12]	2020	Bayesian, SVM	/
Tiulpin et al. [13]	2019	LR, Ensemble learning, CNN	4840
Prashanth et al. [14]	2015	Bayesian, Ensemble learning	584
Ieracitano et al. [15]	2019	SVM, LR, NN	189
Zhao et al. [16]	2022	CNN	49
Ma et al. [17]	2022	CNN	47
Ramkumar et al. [18]	2022	CNN	47
Arteaga-Falconi et al. [19]	2017	SVM	73
Ahmad et al. [20]	2021	CNN	47
Irfan et al. [21]	2022	CNN	452
Zeng et al. [22]	2022	NN	1046
Song et al. [23]	2022	SVM, Ensemble learning, NN, CNN	140,000
Su et al. [24]	2018	CNN	70
El-Rahiem et al. [25]	2020	SVM, Bayesian, Ensemble learning, NN	70
Hammad et al. [26]	2018	CNN	290
Bugdol et al. [27]	2013	SVM, NN	30
Ketu et al. [28]	2020	LR, DT, Ensemble learning	74,501
Alkeem et al. [29]	2021	CNN	150
Rahul et al. [30]	2020	Bayesian, DT, Ensemble learning	47

The framework of early fusion is shown in Figure 2. Early fusion techniques refer to the usage of all input features stitching together for machine learning or deep learning model, directly. Researchers usually clean, filter and construct variables different sources of data. Since multimodal data are simply spliced for training and prediction, researchers should make a lot of effort in feature engineering during the data preparation phase. For example, the height, weight and gender of a patient or normal person are extracted to construct a BMI feature, or the distance and frequency between waves, the mean, standard deviation and percentile of the waves are generated from ECG signals. For early fusion of multimodal data, it is important that the form of the data is uniform or standard. Normalization and standardization are always used to unify the distribution of all data and improve the quality of features for machine learning models [5–9].

Table 3. Overview of fusion style and tasks.

Paper	Fusion style	Tasks
Aziz et al. [5]	early fusion	classification of walking style
Hussain et al. [6]	early fusion	detection of heart failure
Xu et al. [7]	early fusion	classification of neurocognitive impairment
Naik et al. [8]	early fusion	classification of Alzheimer's disease
Walambe et al. [9]	early fusion, late fusion	detection of stress
Battineni et al. [10]	early fusion	classification of Alzheimer's disease
Anand et al. [11]	early fusion	classification of brain tumor
Khan et al. [12]	intermediate fusion	classification of brain tumor
Tiulpin et al. [13]	intermediate fusion	knee osteoarthritis progression prediction
Prashanth et al. [14]	early fusion	detection of early parkinson's disease
Ieracitano et al. [15]	early fusion	classification of EEG recordings in dementia
Zhao et al. [16]	intermediate fusion	detection of learning fatigue
Ma et al. [17]	intermediate fusion	identification and classification of arrhythmia
Ramkumar et al. [18]	/	classification of arrhythmia
Arteaga-Falconi et al. [19]	/	human authentication
Ahmad et al. [20]	early fusion	classification of heartbeat
Irfan et al. [21]	intermediate fusion	classification of heartbeat, detection of arrhythmia
Zeng et al. [22]	early fusion	identification of left ventricular dysfunction
Song et al. [23]	early fusion	detection of cardiovascular disease
Su et al. [24]	early fusion	human identification
El-Rahiem et al. [25]	early fusion, late fusion	biometric authentication
Hammad et al. [26]	intermediate fusion	biometric authentication
Bugdol et al. [27]	early fusion	biometric authentication
Ketu et al. [28]	early fusion	detection of heart disease
Alkeem et al. [29]	/	human identification
Rahul et al. [30]	early fusion	classification of cardiac arrhythmia

The framework of intermediate fusion is shown in Figure 3. The case of an intermediate is divided into the following stages: First, deep learning is used to construct features from a portion of the original data, and the extracted data is fed into a new machine learning or deep learning algorithm for training along with the remaining data. This type of modelling process is suitable for particularly large dimensional data, such as ECG images and MRI images. We can use Neural Networks (NN) or Convolutional Neural Networks (CNN) for feature reduction and selection on images data, and use new features as inputs, which can achieve higher classification accuracy and save training time [12, 13, 16, 17, 21, 26, 29].

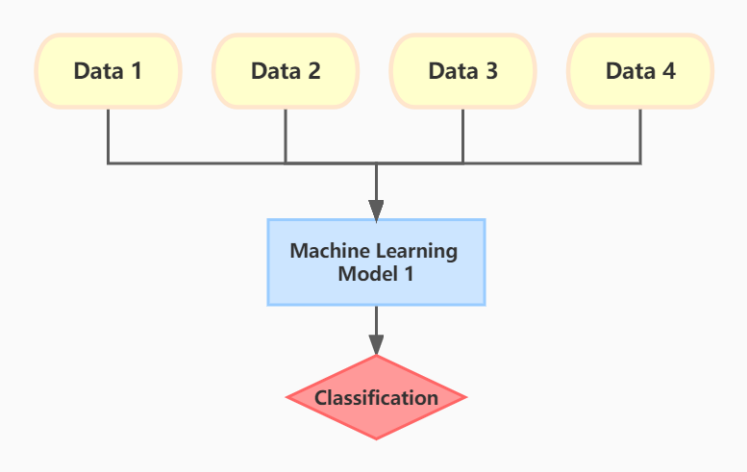

Figure 2. Framework of early fusion.

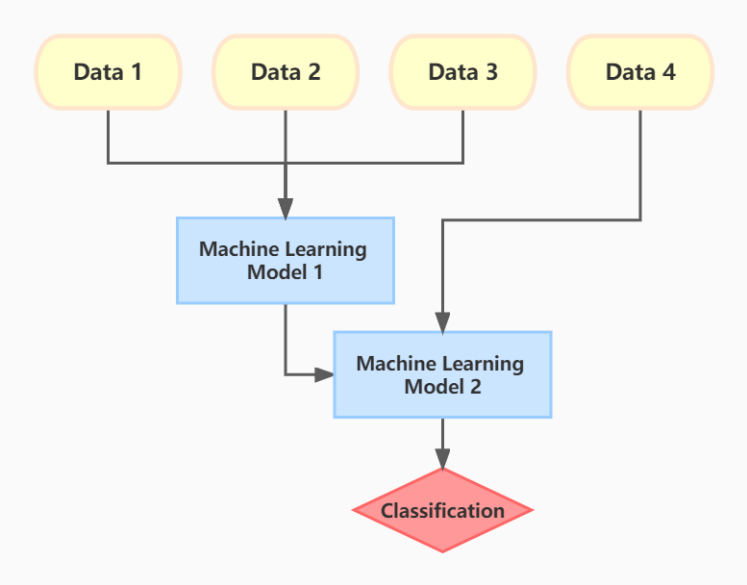

Figure 3. Framework of intermediate fusion.

Similar to Ensemble Learning in machine learning, the principle of late fusion is to use different types of data on different models for training and classification. The framework is shown in Figure 4. Classification results of the different models are scaled and assigned to obtain the final classification results. Different types of machine learning or deep learning algorithms are suitable for handling different types of data for multimodal data, and predictions are made after matching the algorithm to the data one by one. In contrast, the base model in ensemble learning is mostly the same machine learning model, and the base model does not change flexibly for the type of data [9, 25].

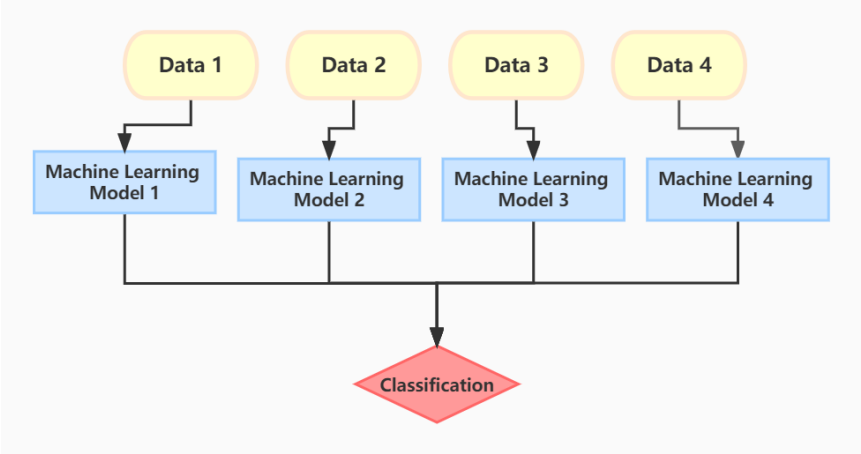

Figure 4. Framework of late fusion.

5. Discussion

In the above sections, we have reviewed and summarized the types of datasets, machine learning models and modelling process for multimodal machine learning in medical diagnostics. Datasets often involved include descriptive data from patient samples commonly used in diagnostics, time series datasets, such as electrical ECG signals, are collected by using devices, and image datasets like MRI are captured by using special devices. Most studies essentially used classical machine learning classification models and deep learning classification models, and the modelling process contains early fusion, intermediate fusion and late fusion. However, there are still many issues that need to be addressed in the study of multimodal machine learning-based medical diagnosis in real-life classification tasks. In Sections 2, 3 and 4 above, we present and analyze the current datasets, machine learning models and modelling processes in multimodal machine learning for medical diagnosis, respectively. In the following discussion, we present some problems that occurr in multimodal machine learning tasks in terms of data quality, comparison of multimodal models with single models, and the explainability of machine learning models.

One of the most critical and fundamental aspects of the multimodal machine learning task is the usage of data. In other fields, such as financial stock price prediction [32], they have huge number of samples and different fields of data to train different deep learning models with many parameters. However, collecting a large sample of test subjects or patient data is time consuming and expensive in the field of medical diagnostics and study. Table 2 shows the number of samples in detail. In Hussain et al.'s study [6], 72 people that consisted of 35 males and 37 females, were collected as samples to detect congenital heart failure. In contrast, in the stress detection made by Walambe et al. [9], only 17 males and 8 females were recruited for the experiment. A larger samples of MRI images collected from 150 patients was used by Battineni et al. [10] to detect Alzheimer's disease by training models and doing data analysis. In our future studies, when low number of samples are obtained, we can consider additional methods, such as the SMOTE algorithm [21], which uses linear interpolation to generate new samples between two minority classes of samples to solve the problem of sample imbalance in the dataset. Alternatively, Generative Adversarial Networks (GAN) in deep learning can generate some new samples, which is also a new way to address the small amount problem in datasets.

For the comparison between multimodal machine learning models and unique machine learning models, different studies show that multimodal machine learning models perform better than models using the uni-modal for medical diagnostics and identification tasks [12, 13, 15, 16, 29]. Xu et al. [7] trained multimodal models consisting of clinical features and MRI imaging and achieved more than 80% accuracy in predicting HIV-infected patients' neurocognitive impairment. However, when clinical features and MRI imaging, were used as input for models independently, the accuracy dropped to 65% and 72%. Zeng et al. [22] used ECG and PCG for prediction in the task of left ventricular dysfunction (LVD) identification, respectively, and the accuracy was around 90%. After using multimodal fusion of ECG and PCG signals, the deep neural network performed recognition with an accuracy of 93%, which is only a 3% improvement. On the other hand, multimodal models fail to have good performance in some tasks. Hammad et al. [26] compared two forms of multimodal data, such as *concat* fusion and addition fusion, with normal data without fusion. They found that multimodal performed better than uni-modal with only 2% more accuracy in the identity recognition problem. Aziz et al. [5] used 9 machine learning models for classification, in which the Decision Tree CART model performed the worst, with an accuracy of 50%, while the Random Forest and Xgboost models had 100% accuracy. With the same multimodal datasets, such a large difference in accuracy indicated that the models may be unfitted. Similarly, Anand et al. [11] showed that SVM model's accuracy was 8–10% higher than the Adaboost model in identifying brain tumor by segmenting MRI images. Moreover, the performance of models using different fusion modals can be different. Taking the study by Walambe et al. [9] as an example, by collecting features fused from facial expression, pose and heart rate as model inputs, different fusion modals have significant difference in the task of detecting stress. The model using early fusion yielded a more informative accuracy with 96%, which is much higher than the model using late fusion with 90% accuracy. Therefore, although multimodal machine learning tends to benefit classification performance, the choice of multimodality should take into account the generalization ability of the model, the amount and quality of data, and the specific problem to be solved. It remains a problem that needs to be discussed and explored by our research team.

In the whole task of medical diagnostics, apart from training and testing models with high accuracy, another key challenge is the explainability of the models. Doctors and health workers using models for diagnostics prefer to know which features in the data are more informative and contribute more to the classification results. However, the black-box operation of deep learning makes the explainability of models very weak. From the papers reviewed, a number of methods were used to measure the contribution of features to the model. Aziz et al. [5] and Prashanth et al. [14] used traditional statistical techniques, such as a hi-square test to remove some less significant features and variables in their studies. Hussain et al. [6] used ROC values to rank all features, and highlight their importance. In addition to these methods in the reviewed papers, some tree based models such as DT, RF and XgBoost can also rank features and, thus, select those that are more important to the classification results.

Recently, in biological sciences, a trending field of analytical study [33] that systematically combines multiple "omics", such as genomics, metabolomics, proteomics, and transcriptomics, emerges as a new type of multi-modal machine learning.

Multi-omics analysis is getting increasingly important as it reveals rich insights of the relationships between different facets of a single cell, creating opportunities for simultaneously discovering phenotype and genotype changes in cancer research. Thus, researchers can have a better understanding of the mechanisms of drug resistance and other therapies in a single cell micro-environment [34–36]. Some success examples include that scientists using multi-omics have advanced knowledge of solid tumors such as melanoma, about its non-genetic drug resistance [37], and, in acute myeloid leukemia, the discoveries of mechanisms of resistance to certain types of

therapies [38].

Multi-omics analysis holds a good future hope for designing the next generation of life-saving cancer drugs and precision therapy. Multi-modal machine learning is anticipated to be an enabler, especially in the aspect of data fusion from different omics modalities, such as imaging and cell profiling [39]. As such, multi-omics share similar data challenges for multi-modal machine learning similar to those described above, for example, data imbalance, curse of high dimensionality and data heterogeneity, etc. As future works, we would be seeing research outcomes in terms of semi- or un-supervised hierarchical machine learning models that allow scientists to select subsets of features across different omics layers, expounding into graph embedding models from which temporal cause-effect relationships could be observed [40]. In addition to the existing fusion models, which were reviewed earlier on, multi-omics requires novel fusion methods that must be robust and biologically interpretable. It is anticipated to see a new breed of multi-modal machine learning models, which will possibly be graph oriented, scalable and explainable in the near future.

It is observed that, in fields of bioinformatics and medicine research, multi-modal machine learning plays an increasingly important role. Deep learning models are being upgraded with capability in fusing data of multiple levels ranging from cell phenotypes, proteins and genomics. Data from singular modal or source is no longer sufficient to power a deep learning model with reasonable accuracy [41]. Multi-modal fusion is a must, rather than an option, due to the natural interrelations of the sources of the data [42]. For example, peptide–protein interaction prediction requires data from multi-levels involving biological and chemical properties and reactions, when it comes to drug design. For another example, designing new treatment plans and new drugs for subsiding tumor progression, requires studying the interactions among the data from physiological samples and cytosolic enzymes, namely, carcinoma cells and Choline Kinase Alpha in particular [43]. Multi-modal machine learning is the core technology in bridging these data that is being demanded for, in tumor treatment research.

6. Conclusions

This review provides a general summary of some papers in recent years that have used multimodal machine learning methods in medical diagnostics or other medical areas. We describe the various kinds of medical data available, how to pre-process different kinds of data with traditional statistical methods or machine learning methods, different types of machine learning models and several different multimodal frameworks. From the current research results, it is obvious that multimodal machine learning has better performance than traditional or uni-modal machine learning, especially when the features are well processed and constructed in a proper modal. We hope that this review paper of multimodal machine learning will facilitate new approaches to multimodal learning in medicine, and also hope that multimodal machine learning will play an important role in medical diagnostics in the future.

Acknowledgments

The authors are thankful for the financial supports of the following grants: Grant No. 2021GH10, Grant No: 2020GH10, and Grant No: EF003/FST-FSJ/2019/GSTIC by Guangzhou Development Zone Science and Technology; Grant No. 0032/2022/A and 0091/2020/A2, by Macau FDCT, and Grant No. MYRG2022-00271-FST and Collaborative Research Grant (MYRG-CRG) – CRG2021-00002-ICI, by University of Macau.

Conflict of interest

The authors declare that there are no personal or organizational conflicts of interest with this work.

References

1. J. Smith, *Science and Technology for Development*, Bloomsbury publishing, 2009.
2. J. Carbonell, R. Michalski, T. Mitchell, An overview of machine learning, *Mach. Learn.*, **5** (1983), 3–23. https://doi.org/10.1016/B978-0-08-051054-5.50005-4
3. J. Tang, G. Liu, Q. Pan, A review on representative swarm intelligence algorithms for solving optimization problems: Applications and trends, *IEEE/CAA J. Autom. Sin.*, **8** (2021), 1627–1643. https://doi.org/10.1109/JAS.2021.1004129
4. A. Triantafyllidis, A. Tsanas, Applications of machine learning in real-life digital health interventions: review of the literature, *J. Med. Internet Res.*, **21** (2019), e12286. https://doi.org/10.2196/12286
5. W. Aziz, L. Hussain, I. Khan, J. Alowibdi, M. Alkinani, Machine learning based classification of normal, slow and fast walking by extracting multimodal features from stride interval time series, *Math. Biosci. Eng.*, **18** (2021), 495–517. http://doi.org/10.3934/mbe.2021027
6. L. Hussain, W. Aziz, I. Khan, M. Alkinani, J. Alowibdi, Machine learning based congestive heart failure detection using feature importance ranking of multimodal features, *Math. Biosci. Eng.*, **18** (2021), 69–91. http://doi.org/10.3934/mbe.2021004
7. Y. Xu, Y. Lin, R. Bell, S. Towe, J. Pearson, T. Nadeem, et al., Machine learning prediction of neurocognitive impairment among people with hiv using clinical and multimodal magnetic resonance imaging data, *J. Neurovirol.*, **27** (2021), 1–11. https://doi.org/10.1007/s13365-020-00930-4
8. B. Naik, A. Mehta, M. Shah, Denouements of machine learning and multimodal diagnostic classification of alzheimer's disease, *Visual Comput. Ind. Biomed. Art*, **3** (2020), 1–18. https://doi.org/10.1186/s42492-020-00062-w
9. R. Walambe, P. Nayak, A. Bhardwaj, K. Kotecha, Employing multimodal machine learning for stress detection, *J. Healthcare Eng.*, **2021** (2021). https://doi.org/10.1155/2021/9356452
10. G. Battineni, M. Hossain, N. Chintalapudi, E. Traini, V. Dhulipalla , M. Ramasamy, et al., Improved alzheimer's disease detection by mri using multimodal machine learning algorithms, *Diagnostics*, **11** (2021), 2103. https://doi.org/10.3390/diagnostics11112103
11. L. Anand, K. Rane, L. Bewoor, J. Bangare , J. Surve, M. Raghunath, et al., Development of machine learning and medical enabled multimodal for segmentation and classification of brain tumor using MRI images, *Comput. Intell. Neurosci.*, **2022** (2022). https://doi.org/10.1155/2022/7797094
12. M. Khan, I. Ashraf, M. Alhaisoni, R. Damaševičius, R. Scherer, A. Rehman, et al., Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists, *Diagnostics*, **10** (2020), 565. https://doi.org/10.3390/diagnostics10080565
13. A. Tiulpin, S. Klein, S. Bierma-Zeinstra, J. Thevenot, E. Rahtu, J. Meurs, et al., Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data, *Sci. Rep.*, **9** (2019), 1–11. https://doi.org/10.1038/s41598-019-56527-3

14. R. Prashanth, S. Roy, P. Mandal, S. Ghosh, High-accuracy detection of early parkinson's disease through multimodal features and machine learning, *Int. J. Med. Inf.*, **90** (2016), 13–21. https://doi.org/10.1016/j.ijmedinf.2016.03.001
15. C. Ieracitano, N. Mammone, A. Hussain, F. Morabito, A novel multi-modal machine learning based approach for automatic classification of eeg recordings in dementia, *Neural Networks*, **123** (2020), 176–190. https://doi.org/10.1016/j.neunet.2019.12.006
16. L. Zhao, M. Li, Z. He, S. Ye, H. Qin, X. Zhu, et al., Data-driven learning fatigue detection system: A multimodal fusion approach of ECG (electrocardiogram) and video signals, *Measurement*, **201** (2022), 111648. https://doi.org/10.1016/j.measurement.2022.111648
17. S. Ma, J. Cui, W. Xiao, L. Liu, Deep learning-based data augmentation and model fusion for automatic arrhythmia identification and classification algorithms, *Comput. Intell. Neurosci.*, **2022** (2022). https://doi.org/10.1155/2022/1577778
18. M. Ramkumar, R. Sarath Kumar, A. Manjunathan, M. Mathankumar, J. Pauliah, Auto-encoder and bidirectional long short-term memory based automated arrhythmia classification for ECG signal, *Biomed. Signal Process. Control*, **77** (2022), 103826. https://doi.org/10.1016/j.bspc.2022.103826
19. J. Arteaga-Falconi, H. Al Osman, A. El Saddik, ECG and fingerprint bimodal authentication, *Sustainable Cities Soc.*, **40** (2018), 274–283. https://doi.org/10.1016/j.scs.2017.12.023
20. Z. Ahmad, A. Tabassum, L. Guan, N. Khan, ECG heartbeat classification using multimodal fusion, *IEEE Access*, **9** (2021), 100615–100626. https://doi.org/10.1109/ACCESS.2021.3097614
21. S. Irfan, N. Anjum, T. Althobaiti, A. Alotaibi, A. Siddiqui, N. Ramzan, Heartbeat classification and arrhythmia detection using a multi-model deep-learning technique, *Sensors*, **22** (2022), 5606. https://doi.org/10.3390/s22155606
22. Y. Zeng, S. Yang, X. Yu, W. Lin, W. Wang, J. Tong, et al., A multimodal parallel method for left ventricular dysfunction identification based on phonocardiogram and electrocardiogram signals synchronous analysis, *Math. Biosci. Eng.*, **19** (2022), 9612–9635. https://doi.org/10.3934/mbe.2022447
23. G. Song, J. Zhang, D. Mao, G. Chen, C. Pang, A multimodel fusion method for cardiovascular disease detection using ECG, *Emerg. Med. Int.*, **2022** (2022). https://doi.org/10.1155/2022/3561147
24. K. Su, G. Yang, B. Wu, L. Yang, D. Li, P. Su, et al., Human identification using finger vein and ecg signals, *Neurocomputing*, **332** (2019), 111–118. https://doi.org/10.1016/j.neucom.2018.12.015
25. B. El-Rahiem, F. El-Samie, M. Amin, Multimodal biometric authentication based on deep fusion of electrocardiogram (ECG) and finger vein, *Multimedia Syst.*, **28** (2022), 1325–1337. https://doi.org/10.1007/s00530-021-00810-9
26. M. Hammad, Y. Liu, K. Wang, Multimodal biometric authentication systems using convolution neural network based on different level fusion of ECG and fingerprint, *IEEE Access*, **7** (2018) , 26527–26542. https://doi.org/10.1109/ACCESS.2018.2886573
27. M. Bugdol, A. Mitas, Multimodal biometric system combining ecg and sound signals, *Pattern Recognit. Lett.*, **38** (2014), 107–112. https://doi.org/10.1016/j.patrec.2013.11.014
28. S. Ketu, P. Mishra, Empirical analysis of machine learning algorithms on imbalance electrocardiogram based arrhythmia dataset for heart disease detection, *Arabian J. Sci. Eng.*, **47** (2022), 1447–1469. https://doi.org/10.1007/s13369-021-05972-2

29. E. Al Alkeem, C. Yeun, J. Yun, P. Yoo, M. Chae, A. Rahman, et al., Robust deep identification using ecg and multimodal biometrics for industrial internet of things, *Ad Hoc Networks*, **121** (2021), 102581. https://doi.org/10.1016/j.adhoc.2021.102581
30. J. Rahul, M. Sora, L. Sharma, V. Bohat, An improved cardiac arrhythmia classification using an rr interval-based approach, *Biocybern. Biomed. Eng.*, **41** (2021), 656–666. https://doi.org/10.1016/j.bbe.2021.04.004
31. A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, et al., Multimodal machine learning in precision health: A scoping review, *npj Digital Med.*, **5** (2022), 1–14. https://doi.org/10.1038/s41746-022-00712-8
32. Y. Wang, K. Yan, Prediction of significant bitcoin price changes based on deep learning, in *2022 5th International Conference on Data Science and Information Technology (DSIT)*, (2022), 1–5. https://doi.org/10.1109/DSIT55514.2022.9943971
33. C. Bock, M. Farlik, N. Sheffield, Multi-omics of single cells: strategies and applications, *Trends Biotechnol.*, **34** (2016), 605–608. https://doi.org/10.1016/j.tibtech.2016.04.004
34. H. Jung, Y. Sung, H. Kim, Omics and computational modeling approaches for the effective treatment of drug-resistant cancer cells, *Front. Genet.*, **12** (2021), 742902. https://doi.org/10.3389/fgene.2021.742902
35. Z. Yuan, Q. Zhou, L. Cai, L. Pan, W. Sun, S. Qumu, et al., Seam is a spatial single nuclear metabolomics method for dissecting tissue microenvironment, *Nat. Methods*, **18** (2021), 1223–1232. https://doi.org/10.1038/s41592-021-01276-3
36. H. Qiao , F. Wang, R. Xu, J. Sun, R. Zhu, D. Mao, et al., An efficient and multiple target transgenic rnai technique with low toxicity in drosophila, *Nat. Commun.*, **9** (2018), 4160. https://doi.org/10.1038/s41467-018-06537-y
37. F. Valenti, I. Falcone, S. Ungania, F. Desiderio, P. Giacomini, C. Bazzichetto, et al., Precision medicine and melanoma: multi-omics approaches to monitoring the immunotherapy response, *Int. J. Mol. Sci.*, **22** (2021), 3837. https://doi.org/10.3390/ijms22083837
38. A. Wojtuszkiewicz, I. van der Werf, S. Hutter, W. Walter, C. Baer, W. Kern, et al., Maturation state-specific alternative splicing in FLT3-ITD and NPM1 mutated AML, *Cancers*, **13** (2021), 3929. https://doi.org/10.3390/cancers13163929
39. S. Stahlschmidt, B. Ulfenborg, J. Synnergren, Multimodal deep learning for biomedical data fusion: a review, *Briefings Bioinf.*, **23** (2022). https://doi.org/10.1093/bib/bbab569
40. Z. Cao, G. Gao, Multi-omics single-cell data integration and regulatory inference with graph-linked embedding, *Nat. Biotechnol.*, **40** (2022), 1458–1466. https://doi.org/10.1038/s41587-022-01284-4
41. Y. Lei, S. Li, Z. Liu, F. Wan, T. Tian, S. Li, et al., A deep-learning framework for multi-level peptide–protein interaction prediction, *Nat. Commun.*, **12** (2021), 5465. https://doi.org/10.1038/s41467-021-25772-4
42. W. Zhou, K. Yang, J. Zeng, X. Lai, X. Wang, C. Ji, et al., FordNet: Recommending traditional Chinese medicine formula via deep neural network integrating phenotype and molecule, *Pharmacol. Res.*, **173** (2021), 105752. https://doi.org/10.1016/j.phrs.2021.105752

43. X. Lin, L. Hu, J. Gu, R. Wang, L. Li, J. Tang, et al., Choline kinase α mediates interactions between the epidermal growth factor receptor and mechanistic target of rapamycin complex 2 in hepatocellular carcinoma cells to promote drug resistance and xenograft tumor progression, *Gastroenterology*, **152** (2017), 1187–1202. https://doi.org/10.1053/j.gastro.2016.12.033

AIMS Press