Contents lists available at ScienceDirect

Integrative Medicine Research

journal homepage: www.elsevier.com/locate/imr

Education Article

A practical guide to implementing artificial intelligence in traditional East Asian medicine research

Hyojin Bae [a], Sa-Yoon Park [b,c,*], Chang-Eop Kim [c,*]

[a] Department of Physiology, Seoul National University College of Medicine, Seoul, Korea
[b] Department of Physiology, College of Korean Medicine, Wonkwang University, Iksan, Korea
[c] Department of Physiology, College of Korean Medicine, Gachon University, Seongnam, Korea

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning
Traditional East Asian medicine
Traditional Chinese medicine
Medical AI

ABSTRACT

In this paper, we present a comprehensive guide for implementing artificial intelligence (AI) techniques in traditional East Asian medicine (TEAM) research. We cover essential aspects of the AI model development pipeline, including research objective establishment, data collection and preprocessing, model selection, evaluation, and interpretation. The unique considerations in applying AI to TEAM datasets, such as data scarcity, imbalance, and model interpretability, are discussed. We provide practical tips and recommendations based on best practices and our own experience. The potential of large language models in TEAM research is also highlighted. Finally, we discuss the challenges and future directions of AI application in TEAM, emphasizing the need for standardized data collection and sharing platforms.

1. Introduction

Traditional East Asian medicine (TEAM) is characterized by its emphasis on the body as an integrated whole, the dynamic interactions between various bodily systems, and the importance of maintaining harmony and balance within the body and between the body and its environment. One of the major strengths of TEAM lies in its holistic approach, which considers multiple variables simultaneously to understand the complex interactions within the human body. However, objectively and quantitatively analyzing these complex interactions has been a challenging task.

To address these challenges, recent advancements in artificial intelligence (AI) technologies, such as machine learning (ML) and deep learning (DL) methodologies (Table 1), are offering new opportunities to further advance TEAM research. These AI techniques enable objective and quantitative pattern identification in complex, multi-dimensional data, thereby allowing for a more comprehensive understanding of the underlying mechanisms of TEAM theories and practices.

In this paper, we aim to provide guidance for TEAM researchers on how to effectively and appropriately apply AI techniques in their domain, taking into account the unique characteristics of TEAM. We focus on the most critical and potentially error-prone elements of the AI model development pipeline, from data preparation to model evaluation and interpretation. By drawing upon best practices and lessons learned from the broader field of AI in healthcare, we offer practical insights, tips, and recommendations based on the unique characteristics and requirements of TEAM research and our own experience. Our goal is to assist researchers in harnessing the power of AI while ensuring that the application of these techniques aligns with the fundamental principles and practices that define TEAM.

We begin by discussing the research objectives establishment step, followed by data selection/collection and preprocessing steps specific to TEAM datasets (Fig. 1). We then provide guidance on selecting appropriate AI models for specific TEAM research needs, considering factors such as data characteristics and research objectives. Next, we delve into model evaluation strategies and techniques for improving model performance, such as cross-validation (CV), hyperparameter tuning (Table 1), and handling data imbalance. We also emphasize the importance of model interpretation in the context of TEAM. Furthermore, we highlight the need for qualitative assessment of model outputs to ensure their clinical relevance and validity. The potential of large language models (LLMs, Table 1) in TEAM research is also gained attention. Finally, we discuss the challenges and future directions of AI application in TEAM research.

2. Research objectives establishment

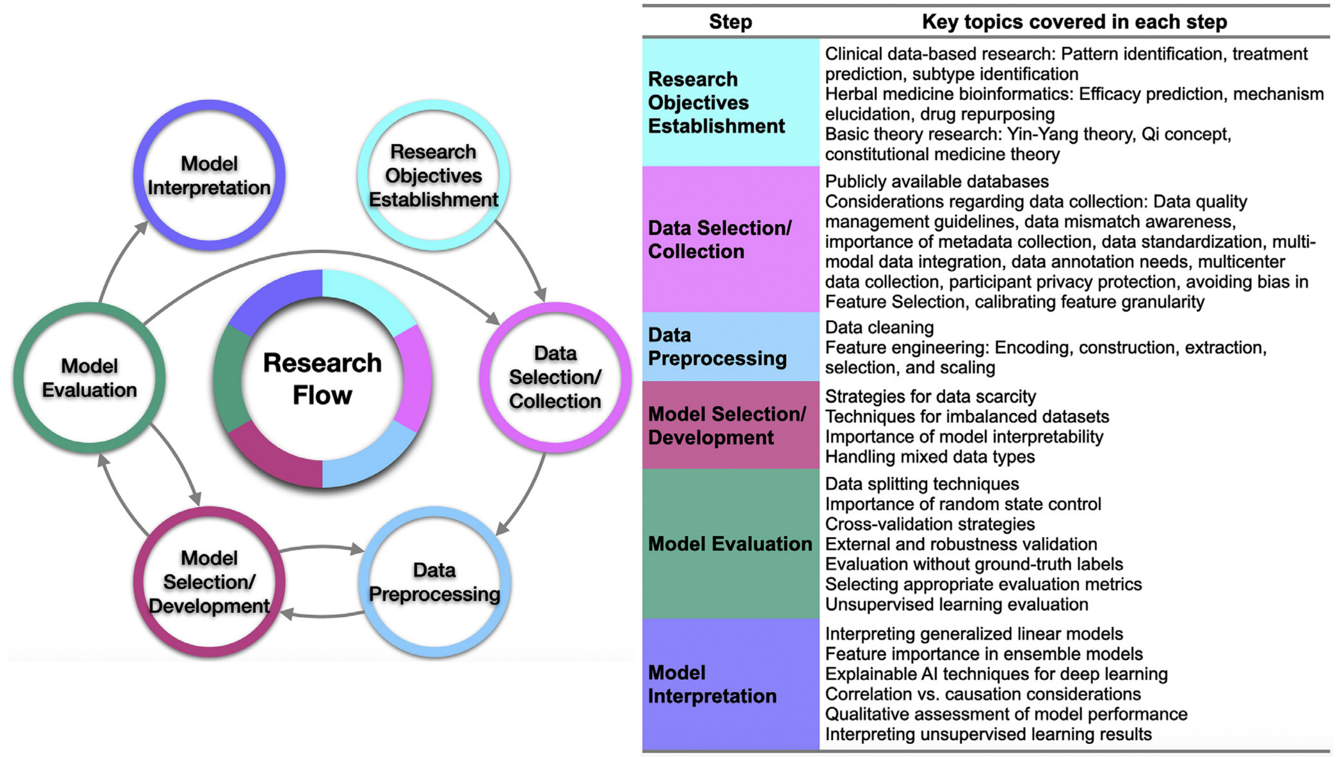
Effective planning is crucial when embarking on AI-integrated research. Clearly define research questions and objectives at the outset. Well-defined goals will guide the data selection/collection process and

[*] Corresponding authors at: Department of Physiology, College of Korean Medicine, Wonkwang University, 460, Iksan-daero, Iksan, Jeonbuk-do 54538, South Korea (S.-Y. Yoon); Department of Physiology, College of Korean Medicine, Gachon University, 1342 Seongnam-daero, Seongnam 13120, South Korea (C.-E. Kim).
E-mail addresses: sayoon8@wku.ac.kr (S.-Y. Park), eopchang@gachon.ac.kr (C.-E. Kim).

https://doi.org/10.1016/j.imr.2024.101067
Received 14 May 2024; Received in revised form 24 June 2024; Accepted 2 July 2024
Available online 7 July 2024

Table 1
Glossary of key terms

Terms	Explanations
Artificial Intelligence (AI)	a broad field focused on designing and developing computer systems that can mimic human intelligence. It encompasses algorithms and technologies that can perform cognitive functions such as learning, reasoning, and problem-solving. AI encompasses various subfields, including ML and DL, which have driven significant advancements in recent years across diverse domains.
Machine learning (ML)	a subset of AI that involves training models to learn patterns and make predictions or decisions from data, without being explicitly programmed. ML algorithms can be categorized into supervised learning (learning from labeled data), unsupervised learning (discovering patterns in unlabeled data), and reinforcement learning (learning through interaction with an environment).
Deep learning (DL)	a subfield of ML that utilizes artificial neural networks with multiple layers to learn hierarchical representations of data. DL models can automatically extract relevant features from raw data, enabling them to learn complex patterns and achieve state-of-the-art performance on tasks such as image classification, natural language processing, and speech recognition.
Large language model (LLM)	a type of DL model trained on vast amounts of text data to understand and generate human language. LLMs, such as GPT (Generative Pre-trained Transformer), can perform various natural language processing tasks, including text generation, translation, summarization, and question answering, by learning the statistical patterns and structures of language from the training data.
Supervised learning	a type of ML where the model is trained on labeled data, meaning that the desired output for each input is provided. The goal is to learn a function that maps input data to the correct output labels. Examples include classification tasks (e.g., identifying spam emails) and regression tasks (e.g., predicting housing prices).
Unsupervised learning	a type of ML where the model is trained on unlabeled data, meaning that no desired output is provided. The goal is to discover hidden patterns or structures in the input data. Examples include clustering (e.g., grouping similar customers) and dimensionality reduction (e.g., compressing high-dimensional data while preserving important information).
Data dimensionality	the number of features or attributes in a dataset. High-dimensional data refers to datasets with a large number of features, which can make it challenging to train models effectively due to the "curse of dimensionality." Feature selection and dimensionality reduction techniques can help address this issue.
Hyperparameters	user-specified values that define the structure or learning process of a model. They are distinct from the weights and biases that the model optimizes during the learning process. Examples include the maximum depth of a decision tree, the number of neighbors in k-nearest neighbors, the learning rate, batch size, and the number of hidden layers in a DL model.
Overfitting	a common problem in ML where a model learns the noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the model is too closely fit to the training data and fails to generalize well to unseen data. Overfitting often occurs when the amount of training data is insufficient to capture generalized patterns or when the model is excessively complex relative to the given problem.
Ensemble models	ML models that combine multiple individual models to improve predictive performance. By combining the predictions of several models, ensemble models can achieve higher accuracy than any single model while improving generalization ability.


Model Interpretation
Research Objectives Establishment
Model Evaluation
Research Flow
Data Selection/ Collection
Model Selection/ Development
Data Preprocessing

Step	Key topics covered in each step
Research Objectives Establishment	Clinical data-based research: Pattern identification, treatment prediction, subtype identification Herbal medicine bioinformatics: Efficacy prediction, mechanism elucidation, drug repurposing Basic theory research: Yin-Yang theory, Qi concept, constitutional medicine theory
Data Selection/ Collection	Publicly available databases Considerations regarding data collection: Data quality management guidelines, data mismatch awareness, importance of metadata collection, data standardization, multi-modal data integration, data annotation needs, multicenter data collection, participant privacy protection, avoiding bias in Feature Selection, calibrating feature granularity
Data Preprocessing	Data cleaning Feature engineering: Encoding, construction, extraction, selection, and scaling
Model Selection/ Development	Strategies for data scarcity Techniques for imbalanced datasets Importance of model interpretability Handling mixed data types
Model Evaluation	Data splitting techniques Importance of random state control Cross-validation strategies External and robustness validation Evaluation without ground-truth labels Selecting appropriate evaluation metrics Unsupervised learning evaluation
Model Interpretation	Interpreting generalized linear models Feature importance in ensemble models Explainable AI techniques for deep learning Correlation vs. causation considerations Qualitative assessment of model performance Interpreting unsupervised learning results

Fig. 1. Research Flow and Key Topics in AI-based Traditional East Asian Medicine Studies.
A circular diagram depicting the research flow, consisting of six interconnected steps: Research Objectives Establishment, Data Selection/Collection, Data Preprocessing, Model Selection/Development, Model Evaluation, and Model Interpretation (left). A table detailing the topics covered in each step of the research process (right).

ensure the gathering of the most relevant information for AI models. To stimulate potential research directions for readers, we will briefly introduce some research topics of AI applications in TEAM.

2.1. Research based on clinical medical data

In research based on clinical medical data, AI techniques have been actively employed to analyze patient data, including symptoms, constitution, diagnosis, and treatment information. The most representative form is to develop predictive models that enhance the objectivity and reproducibility of traditional diagnostic methods, such as pattern identification. Data-driven AI predictions can improve the objectivity and standardization of pattern identification diagnosis and reduce human errors in the diagnostic process.[1–5] Similarly, in the field of constitutional medicine, which emphasizes the importance of diagnosing patients' constitutional types, research on AI-assisted constitution diagnosis is actively being conducted.[6] In treatment prediction or clinical decision support system research, models are constructed to predict the optimal acupuncture and/or herbal medicine treatment method based on patients' pattern types, constitutional types, past treatment records, and clinical indicators.[7,8] This can support TEAM practitioners' clinical decision-making and maximize treatment effectiveness. Furthermore, subtype identification studies within a single disease are actively being conducted, primarily using unsupervised learning (Table 1).[9] This is particularly based on the theory of TEAM that different subtypes exist within a patient group with the same disease.

2.2. Research based on herbal medicine bioinformatics data

Research based on herbal medicine bioinformatics data leverages diverse data sources and analytical methods to predict and optimize the safety, efficacy, combinations, and mechanisms of action of single medicinal herb or herbal formula. Diverse bioinformatics data sources include herbal component information, such as, omics (genomics, transcriptomics, proteomics, and metabolomics) data related to herbal treatments as a result of pharmacological experiments or literature findings. In efficacy prediction, ML is used to predict the pharmacological actions and efficacy of individual herbs or herbal combinations, and to analyze interactions between multiple herbs, synergistic/antagonistic effects, and more.[10–14] In drug repurposing, AI is used to predict new indications for existing herbs and to develop new drugs utilizing the synergistic effects of herbal combinations.[15–17] Finally, in predicting adverse drug reactions, AI is used to predict potential side effects and toxicity of herbs or formulas, and to assess the possibility of adverse reactions due to drug interactions.[18]

2.3. Basic theory research

In addition to the two main categories of research based on usage data, AI techniques are being explored as new approaches to establish the scientific basis for TEAM basic theories. For example, the Yin-Yang theory, which is closely related to the Cold-Heat pattern diagnosis in TEAM, has been investigated using a systems biology approach, revealing that hormones are predominant in the Cold network, immune factors are predominant in the Heat network, and these two networks are connected by neuro-transmitters, providing a molecular basis for the Yin-Yang theory in the context of the neuro-endocrine-immune system.[19] Similarly, the concept of Qi, which is central to TEAM, has been investigated using network pharmacology-based approaches to understand the unique functions of Qi-invigorating herbs.[20] Network pharmacological approach can also be applied to elucidate the essence of Sasang constitutional medicine theory by identifying constitution type-specific compounds and biomarkers.[21,22] Furthermore, text mining and natural language processing techniques are being employed to extract core concepts and relationships of TEAM theories from classical literature.[23–25] In our previous work[26], we modeled pattern identification from a ML perspective. We proposed that interpreting traditional theory through a ML lens offers a novel framework for mathematically understanding the underlying mechanisms of TEAM's theory and practice, while broadening the scope of inquiry.

3. Data selection/collection

The second step in conducting AI-based research is selecting or collecting appropriate data that aligns with the research objectives. In the medical AI field, six major categories of data have been identified for leveraging AI for health: multi-omics, clinical (e.g., medical images, EHR, physiologic data such as EKGs, EEGs), behavioral (e.g., social media, video and conversational data, mobile sensor data), environmental (e.g., air pollution exposures), pharmaceutical research and development (e.g., chemical compounds, clinical trials, spontaneous reports such as information on adverse events), and biomedical literature data.[27] While the data utilized in the field of TEAM is similar to these categories, there are also some aspects specific to TEAM. In TEAM, clinical data often takes a relatively qualitative form due to the nature of examination items, leading to extensive use of questionnaire data for diagnosis and evaluation purposes. Regarding pharmaceutical research and development, herbal medicine-related databases are actively utilized as valuable resources. Moreover, TEAM boasts an extensive repository of classical literature accumulated over its long history, providing a rich source of textual data for mining insights and knowledge extraction through AI techniques.

Representative examples of publicly available databases in Korean Medicine include the following. The Korea Institute of Oriental Medicine (KIOM) provides the Korean Medicine Data Center data, which encompasses a comprehensive collection of clinical information, including survey data, anthropometric measurements, device-based assessments, and biological data. KIOM also offers TM-MC,[28] which provides information about the chemical compounds in medicinal materials from chromatographic articles in PubMed. Additionally, the National Institute of Korean Medicine Development offers extensive databases on various aspects of herbal medicine. These include detailed information on herbal components, in vitro and in vivo pharmacodynamic test results, toxicology study outcomes, and digitized historical medical texts. Furthermore, the field of Traditional Chinese Medicine (TCM) contributes several databases that are valuable for network pharmacology research and herbal medicine-related bioinformatics studies. Notable among these are the Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform (TCMSP[29]) and the Bioinformatics Analysis Tool for Molecular mechANism of Traditional Chinese Medicine (BATMAN-TCM.[30])

While existing databases provide valuable resources, they may not always meet the specific needs of every research project. Given these constraints, researchers often develop custom datasets, where several key considerations come into play to ensure data quality and relevance. To address data quality management, researchers can refer to resources such as the National Information Society Agency of Korea's guidelines for AI training data quality. In the context of clinical data collection, researchers can refer to frameworks such as the Kahn framework,[31] which addresses data quality issues in secondary use of Electronic Health Records, and the DQ4HEALTH framework,[32] which is tailored to the current state of healthcare data in Korea.

When implementing these data quality principles in practice, several key considerations emerge. One critical aspect researcher must address is the potential for data mismatch issues that could affect the generalizability of AI models, such as dataset shift and sample selection bias.[33] To mitigate these, it's crucial to gather comprehensive metadata, including details about the data collection, annotation process, patient demographics and environmental factors. This metadata helps identify biases and adjust for confounding variables. Additionally, ensuring diverse sampling and standardizing data acquisition protocols can further enhance the robustness of AI models across different clinical settings.

Standardization is also essential in data collection, requiring consistent templates, clear data entry rules, and uniform coding schemes. When feasible, incorporating multi-modal data can provide a more comprehensive view, though it increases sample size requirements and analytical complexity. Planning for data annotation and curation needs is crucial. Researchers should determine necessary annotations (e.g., constitutions, diagnoses, prescriptions) and allocate sufficient resources for data labeling, often securing multiple independent labelers to ensure reliability. Collecting data from multiple centers enhances generalizability and robustness, encompassing diverse patient populations and clinical practices. Throughout this process, researchers must prioritize the protection of participants' rights and privacy, ensuring secure storage and anonymization of personal information.

In the context of TEAM research, additional considerations arise. It's vital to include all clinically relevant features without prematurely filtering based on presumed significance, as our previous work[34] has revealed disparities between features TEAM doctors deemed important and those identified as crucial by ML models. Many clinical features in TEAM involve qualitative judgements by doctors, such as strength in a patient's voice and overall color and tone of the skin. When designing data collection protocols for these qualitative assessments, researchers face a trade-off between information granularity and assessment burden. Information granularity refers to the level of detail in measurements, such as using a 3-point versus a 5-point scale for evaluating symptoms. While finer scales provide more detailed information, they also increase the cognitive load on doctors and potentially the time required for assessment. Researchers must carefully balance these factors, establishing clear, objective criteria for each scale point to minimize inter-practitioner variability.

4. Data preprocessing

Data preprocessing can be broadly divided into two main stages: data cleaning and feature engineering, which includes feature encoding, feature construction, feature extraction, feature selection, and feature scaling.

The obtained or collected data often exists in an incomplete state with substantial noise. The data cleaning process involves removing unnecessary information, correcting errors, and standardizing data formats to improve data quality. Handling missing values is crucial as it can affect the accuracy of the research. Interpolation methods or substitution values are used to maintain data integrity, and typically, cases or variables with more than 20 % missing values are removed. This threshold value can vary depending on the research topic and data situation, and researchers should select appropriate criteria. As there can be various scenarios of data missingness, domain knowledge should be utilized to proceed in a direction that does not introduce bias. Common missing value handling methods include zero imputation, mean/median imputation, and multiple imputation.[35]

Next, feature engineering is a critical step that directly impacts model performance. We will examine this process by breaking it down into feature encoding, feature construction and extraction, feature selection, and feature scaling. Feature encoding is the process of converting categorical variables into numerical values that the model can understand. One-hot encoding, which transforms categorical data into binary vectors, and ordinal encoding, which assigns an order to the categories, are commonly used.[36]

Feature construction leverages domain knowledge to generate new informative features from existing ones beyond the data itself. For instance, in clinical medical data, deriving the body mass index feature from height and weight measurements offers a standardized metric for assessing obesity risk. Combining multiple disease codes can generate a 'major disease category' feature (e.g., diabetes and hypertension codes mapped to metabolic/circulatory disorders), facilitating disease-specific investigations. Incorporating such constructed features, informed by domain expertise, plays a crucial role in enriching the dataset, enhancing the predictive power and interpretability of ML models applied to data analysis. Feature extraction techniques, on the other hand, focus on automatically discovering latent features from high-dimensional data without relying on domain knowledge. These methods aim to capture the most informative aspects of the data while reducing its dimensionality. By leveraging these extracted features, ML models can better capture the underlying structure of the data, leading to improved predictive performance and novel insights into disease mechanisms.

Feature selection is employed to identify important features and remove unnecessary ones, thereby reducing model training time and the risk of overfitting (Table 1). Filter and wrapper methods are representative approaches.[37] Filter methods assess and select features based on statistical measures (e.g., correlation coefficient, Chi-square, feature importance). These methods evaluate the relevance of each individual variable to the target variable without considering the interactions or combined effects of variable subsets, resulting in lower computational complexity at the cost of neglecting potential synergistic effects between variables. Wrapper methods, on the other hand, select feature subsets based on model performance (e.g., recursive feature elimination, BorutaShap).[38,39] These methods evaluate the performance of various variable combinations, taking into account the combined effects of variables, but require higher computational resources compared to filter methods. The dataset used for selecting features via filter or wrapper methods should not overlap with the dataset used for evaluating model performance. Otherwise, the selected variables may be overfitted to the evaluation dataset, resulting in inflated performance. It's worth noting that in scenarios where data is limited, the aforementioned feature engineering techniques – particularly feature construction, extraction, and selection – become even more crucial. These methods can effectively augment the available information, create more meaningful representations of the data, and identify the most relevant features, thereby improving model performance despite data constraints.

Finally, before inputting the data into the model, feature scaling processes are applied. In many cases, features in a dataset can have vastly different scales, which can negatively impact the performance of some ML models. To address this issue, standardization (z-score normalization) or min-max scaling are commonly applied before inputting the data into the model. Standardization transforms the feature values to have a mean of 0 and a standard deviation of 1, while min-max scaling adjusts the feature values to a specific range, typically between 0 and 1. Effective feature scaling can significantly improve the convergence speed and model performance for scale-sensitive algorithms, making this process a crucial step in the feature engineering pipeline.

5. Model selection and development

AI model selection is a crucial aspect of data analysis, as no single model is universally optimal for all data and objectives. When selecting a model, several factors should be considered, including the problem type, data characteristics, model interpretability, and computational efficiency. This section presents practical tips for model selection, focusing on the challenges and considerations commonly encountered in medical AI research (For a comprehensive survey of model selection techniques, refer to.[40])

When data scarcity is prevalent, as commonly seen in TEAM research, it necessitates a more meticulous approach in model selection and development. Limited data samples increase the risk of overfitting, where the model learns features specific to the training data that are not generally applicable. Primarily, when available data is limited, employing models with lower complexity can be advantageous. For example, favoring logistic or linear regression over deep neural networks, or opting for shallow neural architectures, can help learn latent data patterns while mitigating the risk of overfitting. Additionally, incorporating regularization terms like the sum of model weights into the cost function can help in simplifying the model. This process forces unimportant feature weights to be zero or close to zero during training, which helps in preventing overfitting. Combining multiple simple models through ensemble techniques (Table 1) can also be an alternative way to effectively solve complex problems while keeping individual model complexity low. Each model in the ensemble focuses on capturing different aspects of the data, collectively capturing a more comprehensive view of the problem space. When labeled data is scarce but related databases or unlabeled data are abundant, transfer learning and self-supervised learning can be used as alternative approaches to tackle the data scarcity problem. Transfer learning involves applying knowledge from a model trained on one task to a related task. For example, TCMBERT[41] is a two-stage model that generates TCM prescriptions by pre-training on TCM books and fine-tuning on a limited number of medical records. GreasyCoatNet[42] is a model for recognizing greasy coating of the tongue, built by fine-tuning the ResNet pre-trained on ImageNet dataset. Self-supervised learning is a ML paradigm that learns representations from unlabeled data by training models to solve pretext tasks. Pretext tasks are designed to capture meaningful patterns and structures in the data, such as predicting missing parts of an input or distinguishing similar samples from dissimilar ones. In the medical domain, self-supervised learning is particularly promising because it can leverage the abundance of unlabeled medical data to learn robust and generalizable representations, overcoming the challenge of limited labeled data due to the high cost and expertise required for manual annotation.[43] For example, training the model to learn to identify similar chest X-ray views of the same patient while distinguishing different patients leads to improved performance for chest X-ray interpretation task.[44]

Another common challenge encountered when training models in the medical field is the presence of significant label imbalance in datasets. For example, certain classes such as rare Sasang constitution (e.g., Tae-Yang constitution) or infrequently prescribed herbal formula[8] may be significantly outnumbered, leading to an imbalanced dataset. This can lead to models that perform well on frequent classes but poorly on rare ones. To address this imbalance problem, ensemble methods are commonly recommended due to their built-in handling for imbalance using weighted loss function, which impose greater penalties for misclassifying infrequent classes during training. Advanced techniques to mitigate data imbalance include oversampling the minority class, which involves creating synthetic examples of underrepresented classes (e.g., SMOTE,[45]) and under-sampling the majority class by randomly removing samples from the majority class. These techniques help the model to learn a more balanced representation of classes.

In addition to addressing data-related challenges, interpretability is another crucial consideration in model selection, where understanding the decision-making process of models can be as important as their predictive accuracy. When the problem at hand is not overly complex and interpretability is paramount, traditional models such as logistic regression and decision trees can be considered. These models offer insights into how predictors influence the outcome, although they usually do not guarantee the highest performance compared to more advanced ML models. Logistic regression quantifies the impact of each predictor with coefficients, and decision trees provide interpretability through a hierarchical structure, visually mapping the paths from features to outcomes, although their clarity may diminish as complexity increases due to extensive branching.

Furthermore, medical data often comprises diverse data types such as patient demographics (categorical) and clinical measurements (numerical). Such mixed data types typically require preprocessing steps including normalization of numerical values and encoding of categorical variables. Tree-based models can naturally handle numeric and categorical data without complex preprocessing, since they segment data by choosing splits based on thresholds for numerical data and grouping for categories in categorical data.

In addition to the supervised learning (Table 1) considerations discussed above, there are scenarios where the goal is to discover novel patterns within the data, such as identifying previously unknown patient subtypes. In these cases, unsupervised learning methods, including dimensionality reduction (Table 1) and clustering, become relevant. Additionally, topological data analysis techniques like the Mapper algorithm[46] offer a means to visualize high-dimensional data in a graphical format, revealing inherent data structures.

Lastly, it is worth addressing a common misconception among beginners when choosing between ML and DL models. It is often mistakenly assumed that DL is always the superior or more advanced approach, overlooking the potential of traditional ML algorithms. DL models excel at handling large datasets, especially image or sequential data, due to their ability to capture complex spatial or temporal dependencies. For example, DL models such as convolutional neural networks and vision transformers have been successfully employed for image recognition involving tongue[47–49] or herb images,[50,51] as well as for feature extraction from unstructured pulse data.[52,53] However, when working with smaller tabular datasets that have high-dimensional features, standard neural networks may not always be the optimal choice. In such cases, support vector machines or tree-based ensemble models can provide more robust results.

6. Model evaluation

Rigorous model evaluation is a critical step in the AI workflow, ensuring the developed models are reliable, generalize well to unseen data, and align with the intended objectives before deployment in real-world applications. However, evaluating models rigorously without data leakage and selecting appropriate metrics can be formidable for beginners in AI research.

Data splitting involves dividing a dataset into subsets for training, validating, and testing a model (e.g., 60 %, 20 %, and 20 %, respectively). The validation set is used for hyperparameter tuning and intermediate evaluation, while the test set assesses the final performance. Since the validation set is used to adjust the model's hyperparameters during the training process, the model may overfit to the validation set. Therefore, by using a completely separate test set to evaluate the final performance of the model, it is possible to more accurately measure how well the model generalizes to unseen data.

In ML, data splitting is often performed using randomization to ensure unbiased distribution of samples across training, validation, and test sets. The randomness in this process is typically controlled by a random seed or random state parameter, which determines the sequence of random numbers generated by the algorithm. By default, this parameter is set to None, resulting in different data splits each time the code is executed. However, when sharing code on platforms like GitHub for other researchers to reproduce results or in tutorials, it may be necessary to set a specific value for the random state to ensure consistent data splitting across multiple runs. Moreover, during the model development process, fixing the random state along with all other hyperparameters, except the one being tuned, can help isolate the effect of the manipulation being performed. Nonetheless, when evaluating the final performance of a model, it is crucial to verify that the results are not significantly influenced by the specific data split used. This can be achieved by iterating over multiple random states or employing techniques such as CV. These methods assess the robustness and generalizability of the model's performance across different data splits, providing a more reliable estimate of its true performance on unseen data.

In an optimal scenario with sufficient data, a validation set would be reserved to evaluate the performance of each hypothesis model. However, when data are scarce, the conventional approach of partitioning the available data into three distinct sets significantly reduces the number of samples available for model training. Moreover, since test set size is also small, test performance can vary significantly depending on the data split. In such scenarios, CV is a viable solution. First, the dataset is partitioned into training and test sets. In k-fold CV, the training set is further divided into k smaller subsets, known as folds (Fig. 2), with a typical choice of k between 5 and 10. For each hyperparameter configuration, the model is then trained on k-1 folds while using the remaining

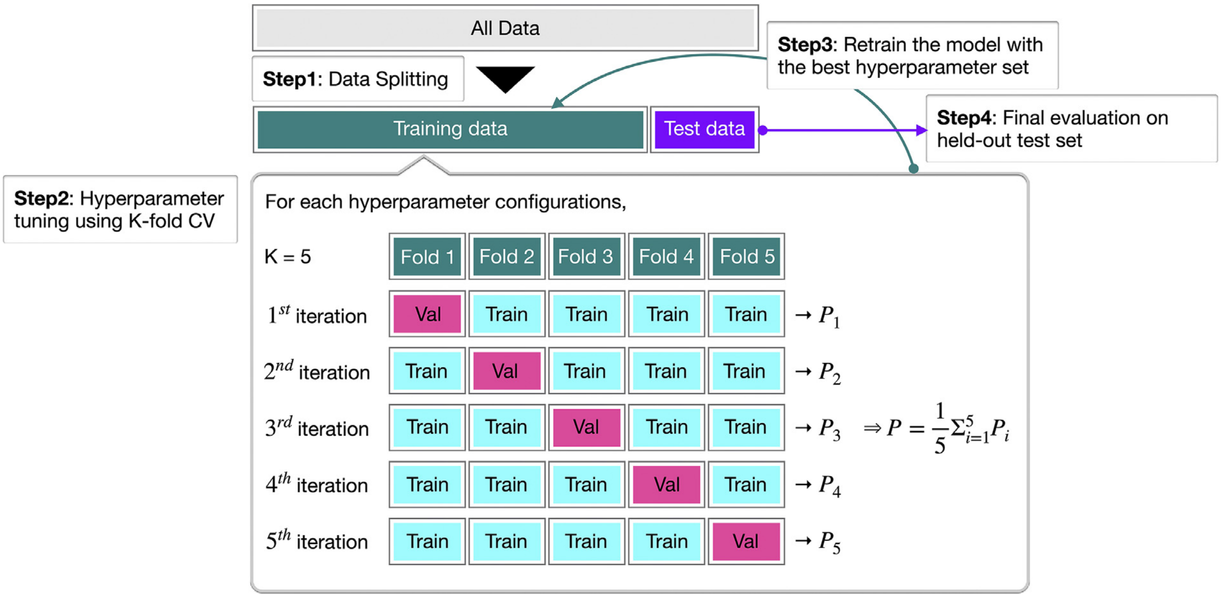


Fig. 2. K-fold cross-validation procedure for model training and hyperparameter tuning.
The dataset is split into training and test sets. The training set undergoes a 5-fold cross-validation for hyperparameter optimization, where each fold serves as the validation set once while the other folds are used for training. This process yields five performance estimates (P_1 to P_5), and their average (P) is used to select the best hyperparameter configuration. The model is then retrained on the entire training set using the optimal hyperparameters and evaluated on the held-out test set to assess its generalization performance.

fold for validation. This process is repeated k times, ensuring that each fold is used exactly once for validation. The optimal hyperparameter set is determined by the best average performance across the k folds. Finally, the model is trained on the entire training set using the optimal hyperparameters and evaluated on the held-out test set to report its final performance. In extremely data-scarce situations, such as when the dataset contains fewer than 100 samples, leave-one-out CV can be employed, wherein each single data point is used as a separate validation set while the model is trained on all remaining data points. Lastly, it is of great importance to note that, during model development, it is essential to inspect the performance of each fold individually, rather than relying solely on summarized metrics. This practice provides valuable insights into the model's behavior and can help identify potential issues or inconsistencies, thus enabling researchers to make informed decisions regarding model development and evaluation.

The process we've discussed thus far pertains to internal validation, which is sufficient when the dataset is large enough and representative of the target population. However, in many cases, these conditions are not met, necessitating external validation, which utilizes a completely independent dataset often sourced from different institutions or the same population but at a different time period.[54] This is especially critical in healthcare, where external validation ensures that predictive models reliably support clinical decisions across diverse patient populations. When conducting external validation or preparing for real-world application, researchers should identify potential sources of data mismatch between development and target populations. Strategies to mitigate the data mismatch include applying importance weighting techniques to adjust for demographic or prevalence differences, and implementing domain adaptation methods if test images are available to align feature distributions across domains.[33] Beyond addressing distribution shifts, AI system evaluation on real-world deployment data often lacks ground-truth labels in clinical settings. This common scenario significantly complicates the assessment of AI model performance in practical applications. To tackle this issue, frameworks like SUDO[55] have been proposed, utilizing pseudo-labels to estimate AI prediction reliability without requiring ground-truth labels. SUDO works by discretizing AI-generated probability scores, assigning temporary labels, and training classifiers to measure discrepancies between pseudo-labeled data and known outcomes, thus identifying unreliable predictions and potential biases. Robustness validation evaluates a model's capability to maintain its performance in the presence of data irregularities such as noise[56], outliers, or missing values.[57] This type of validation is also crucial, ensuring that predictive models can effectively handle the imperfect data encountered in clinical environments.

Selecting an appropriate evaluation metric is a critical step that directly impacts the interpretation and reliability of the model's performance. For classification tasks, accuracy is used when dealing with uniform class distribution. However, accuracy can be misleading when handling imbalanced data. For instance, in the case of rare diseases where positive samples are scarce, a model that predicts all samples as negative can still yield considerably high accuracy. In such scenarios, precision, recall, or F1 score are preferred (Fig. 3a). Precision is used when minimizing false positives is important, while recall is used when minimizing false negatives is crucial. The F1-score is the harmonic mean of precision and recall and is useful when considering the balance between the two metrics. Macro F1 averages the F1-scores of each class, treating all classes equally, while micro F1 calculates metrics globally, giving more weight to larger classes. It is strongly advised to visualize and examine the confusion matrix in addition to calculating the aforementioned metrics, as it offers valuable insights into the models' class-wise performances and error types, enabling targeted improvements. In a binary classification, area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPR) can be used to comprehensively evaluate the model's performance at various thresholds (Fig. 3b). The optimal threshold can be selected based on the domain requirements (e.g., whether minimizing false positives or false negatives is more important). AUROC, plotted based on the true positive rate and false positive rate, represents how well the model distinguishes between positive and negative classes. However, it may not adequately reflect the decrease in recall for the minority class, which is a common concern in highly imbalanced scenarios. In such cases, AUPR, which considers both precision and recall, is preferred. For regression tasks, metrics such as mean squared error (MSE), mean absolute error (MAE), and R-squared (coefficient of determination) are suitable. MSE

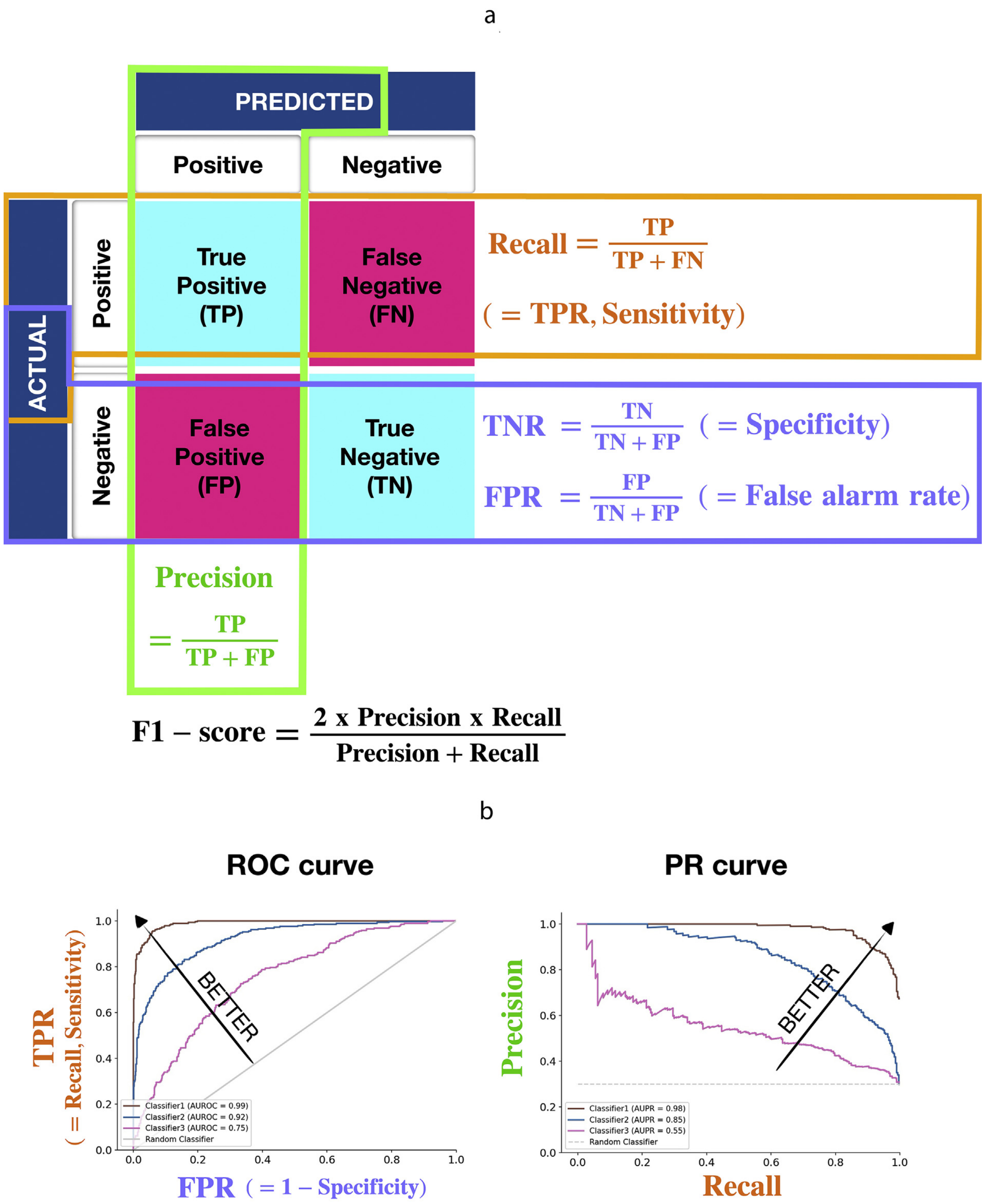


Fig. 3. Evaluation of classifier performance using confusion matrix and ROC/PR curves.
(A) Confusion matrix with key performance metrics: precision, recall, and F1-score. This confusion matrix visualizes the classifier's predictions against the actual classes, highlighting true positives, true negatives, false positives, and false negatives. **(B)** Receiver operating characteristic (ROC) curve (left) and precision-recall (PR) curve (right) for assessing the classifier's discriminative power. The ROC curve shows the trade-off between the true positive rate (TPR) and false positive rate (FPR), while the PR curve illustrates the trade-off between precision and recall. The area under the ROC curve (AUROC) and area under the PR curve (AUPR) quantify the overall effectiveness of the classifier, used to evaluate either different models or variations in threshold settings with the same model.

and MAE measure the error between the predicted and actual values, while R-squared indicates how well the model explains the variability in the data. Specifically, R-squared is defined as 1 minus the ratio of the residual sum of squares to the total sum of squares, representing the unexplained variance. In linear regression, the total variance equals the sum of the explained and unexplained variances, allowing R-squared to be interpreted as the proportion of the total variance explained by the model. However, this equality does not hold for nonlinear regression, making R-squared less interpretable and less suitable for evaluating non-linear models.[58]

Evaluating unsupervised learning models requires a different approach. For clustering, internal evaluation metrics like silhouette score or Davies-Bouldin index can be used to assess the quality of clustering based on the cohesion and separation. Additionally, performing clustering multiple times using a subset of the data (e.g., 80 %) and checking the stability and reproducibility of the results is also a good practice to ensure the robustness of the clustering result. There are various approaches to evaluate the quality of dimensionality reduction results. A common approach is to assess the preservation of local structure (e.g., cluster label) or global structure (e.g., overall cluster arrangement and relative ordering of pairwise sample distances) in the reduced-dimensional space compared to the original high-dimensional space.[59] Another approach to evaluate dimensionality reduction results involves assessing their sensitivity to various choices made during the process, such as parameter settings and preprocessing steps. Additionally, the preservation of information can be measured using techniques like explained variance in principal component analysis or reconstruction error in autoencoders.[60]

7. Model interpretation

While reporting performance metrics is important, interpreting the results is equally crucial. In the medical field, clearly explaining the factors that drive AI model predictions is particularly critical, as it allows clinicians to assess the model's reasoning and ensures transparency and trust in the system's outputs.

Generalized linear models are widely utilized in medical research due to their interpretability. These models provide regression coefficients, which indicate the direction and magnitude of the effect of each predictor variable on the outcome variable. Logistic regression particularly uses odds ratios, obtained by exponentiating the coefficients, to assess the relative influence of predictor variables. An odds ratio represents the change in the odds of the outcome variable being 1 (e.g., the presence of a disease) for a one-unit increase in the predictor variable, while controlling for other variables in the model. While coefficients provide information about the effect size, p-values and confidence intervals are used to assess the statistical significance and precision of these estimates. A p-value indicates the probability of observing an effect as extreme as the one found, assuming the null hypothesis is true, while a confidence interval provides a range of plausible values for the true effect size. Both statistical significance and effect size are crucial for making informed clinical decisions, as they offer complementary insights into the reliability and practical relevance of research findings. Volcano plots are instrumental in visualizing these two essential aspects simultaneously, displaying the magnitude of coefficient values alongside their corresponding p-values. This allows researchers to quickly identify key biological or medical insights that are both statistically significant and have a substantial effect size.

While these interpretation methods are well-established for traditional statistical models, ensemble models provide another valuable approach for model interpretation. These models inherently provide feature importance scores, which assess the impact of each feature on the model's predictions. For instance, in random forest, permutation importance measures the decrease in model performance when a feature is randomly shuffled, while MDI (mean decrease in impurity) quantifies the reduction in impurity (e.g., Gini impurity or entropy) achieved by splitting on a feature, averaged across all decision trees. Features that significantly affect model performance or consistently reduce impurity are considered important. For instance, one study[6] employed extremely randomized trees classifiers to predict Sasang constitution types using a comprehensive clinical dataset. The feature importance scores from the trained model identified the most informative features for classifying Sasang types, revealing that body measurement features were the most crucial, followed by personality, general information, and cold-heat characteristics, with costal angle being the single most important feature.

While these interpretation methods are well-established for traditional statistical models, recent advancements in explainable AI (XAI) have provided techniques to interpret DL models, which are often considered "black boxes" due to their complex architectures. One prominent XAI method is SHapley Additive exPlanations (SHAP,[39]) which assigns importance values to each input feature based on its contribution to the model's output. SHAP values are calculated by considering all possible combinations of features and comparing the model's predictions with and without each feature, providing a robust and individualized measure of feature importance. Another popular XAI technique is Local Interpretable Model-agnostic Explanations (LIME,[61]) which focuses on explaining individual predictions. LIME generates local interpretable models by slightly perturbing the input data around a specific instance and learning a simpler, interpretable model (e.g., linear regression) that mimics the DL model's behavior in the vicinity of that instance. The learned local model's coefficients provide insights into the most important features for that particular prediction. These XAI methods enable researchers to interpret DL models, providing transparency and understanding of their inner workings. For instance, in a study focused on distinguishing between morphologically similar medicinal herbs, a convolutional neural network was employed for classification, and LIME was subsequently applied to identify the crucial morphological features that differentiate between species.[62] Another study utilized ML models to predict quality of life in middle-aged adults. The application of SHAP revealed that stress and sleep quality were the most significant predictors of quality of life in this demographic.[63]

While these interpretation methods provide valuable insights, caution must be exercised to avoid confusing correlation with causation. To address this issue, practitioners can utilize causal inference methods such as causal diagrams to map out assumed relationships and potential confounders.[64] Additionally, implementing randomized controlled trials where feasible and applying sensitivity analyses to observational data can enhance the robustness of causal claims, ensuring more reliable and actionable insights in medical research and practice.

When evaluating model performance, it's essential to go beyond just reporting metrics and provide a qualitative assessment as well. It's important to consider what level of performance improvement is considered meaningful and significant in the specific clinical context. This involves comparing the model's performance to a relevant baseline, such as the performance of a simpler model or the current standard of care. The required performance threshold for a model to be deemed useful may vary depending on the severity of the condition, the consequences of false positives or false negatives, and the available alternative methods.

In cases where unsupervised learning is used for subtype identification, it's crucial to further investigate the clinical relevance and implications of the discovered subtypes. Expert interpretation and assessment of whether the identified subtypes align with known disease categories or represent novel, biologically meaningful distinctions are essential. For instance, one study[65] employed a deep autoencoder-powered pattern identification model using multi-site cross-sectional survey data from patients with sleep disturbances. The unsupervised learning model derived three distinct patient clusters differentiated by changes in sleep quality, dietary habits, and concomitant gastrointestinal symptoms, which were interpreted as corresponding to specific pattern identification types recognized in TEAM. Another study[9] integrated biopsychosocial information from conventional and traditional medicine, as well as quality of life questionnaires. By applying nonlinear dimensionality reduction followed by clustering, four novel functional gastrointestinal disorder subtypes were identified and interpreted as mild, severe, mind-symptom predominance, and body-symptom predominance based on the normalized average scores of the top 50 body and mind-related variables. Furthermore, analyzing the clinical characteristics, outcomes, biomarker profiles, and treatment responses associated with each subtype can help validate their utility and guide further research into tailored interventions.

8. Large language models

LLMs are a revolutionary subset of DL algorithms that leverage neural networks with billions of parameters to process and understand human language with remarkable proficiency. These models are pre-trained on a vast corpus of diverse text, enabling them to capture complex linguistic structures and semantics. Conversational language models, such as ChatGPT, have further propelled their capabilities by enabling interactive, context-aware conversations.

Developing LLMs from scratch by individual researchers or smaller institutions is practically infeasible due to significant resource and time requirements. Instead, leveraging existing models, particularly through fine-tuning smaller open-source models, presents a more viable approach.[66,67] For instance, models like TCM-GPT[68] can be effectively fine-tuned on a large TCM-specific corpus, demonstrating outperforming results on TCM examination and TCM diagnosis. Additionally, retrieval augmented generation provides an alternative method of leveraging LLMs without training.[69] Unlike fine-tuning, which involves modifying the internal parameters of a pre-trained model to adapt it to specific tasks, retrieval augmented generation retrieves relevant information from external sources to enhance the model's responses without altering its foundational structure. These methods bridge the gap between general-purpose models and domain-specific needs, offering a scalable solution for enhancing model accuracy and relevancy in specialized fields like TEAM. (For a comprehensive review of research trends applying LLMs in the medical field, refer to.[70,71])

For users without deep technical expertise, there are more accessible methods for utilizing generative LLMs. Web-based usage is a convenient way to engage with LLMs, allowing users to interact with these models through simple web interfaces without complex setup. In the realm of language models, prompting refers to the process of providing input to steer or influence the model's generated output. By focusing on prompt engineering, users can harness LLMs' capabilities to extract high-quality responses.[72] An example of this is the Medprompt[73] technique, which combines dynamic few-shot selection, self-generated chain-of-thought reasoning, and choice-shuffling ensemble strategies to improve the accuracy and robustness of GPT-4′s responses in medical application. API-based usage offers more flexibility, allowing users to integrate LLMs into custom applications. This option is ideal for those who need to incorporate LLM functionalities into specific software or platforms. With the help of user-friendly libraries and tools, users can create chatbots with minimal coding knowledge, making the development process more accessible. These chatbots can serve various purposes, including providing simulated clinical scenarios for students to practice diagnosing and treating virtual patients[74], facilitating communication between healthcare providers and patients, and offering clinicians relevant information to support clinical decision-making.[75]

Remarkably, the advent of LLMs with advanced coding capabilities has significantly lowered the barrier for researchers seeking to integrate AI into their work. Tools like AI-assisted code generation and automated data preprocessing have made AI more accessible to a broader range of researchers, even those without extensive coding expertise. However, it is crucial to recognize that this increased accessibility does not diminish the importance of understanding the key concepts and potential challenges throughout the entire research process. From data selection and preprocessing to model development, evaluation, and interpretation, researchers must be aware of intricate issues such as data leakage, overfitting, and ethical considerations. As LLMs continue to evolve and shape the landscape of TEAM research, it is essential for researchers to not only leverage these powerful tools but also approach them with a critical understanding of their limitations and implications. By doing so, we can harness the potential of LLMs to advance TEAM research while ensuring the integrity and reliability of our findings.

9. Challenges to overcome and future directions

To effectively utilize AI techniques in TEAM research, several challenges need to be addressed. The most pressing issue is the scale and quality of the collected data. Currently, there is a severe lack of quantitative data in the field of TEAM, which acts as a significant obstacle to the development and validation of AI models. This scarcity of data can be attributed, at least in part, to certain inherent characteristics of TEAM practice. These characteristics, such as the emphasis on individualized treatment, the holistic diagnosis based on patients' self-reported symptoms and observations (inspection, listening, inquiry, and palpation), and the use of complex herbal formulas, pose unique challenges in collecting and standardizing data for computational analysis. As a result, it is challenging to systematically collect high-quality quantitative data in clinical settings. Recognizing this problem, the TEAM community is making multifaceted efforts to collect data. For example, a project aims to establish an infrastructure for collecting, processing, and utilizing various forms of data generated in TEAM clinical practice and to develop AI-based diagnostic and treatment support systems using this infrastructure.[76,77] It is hoped that such initiatives will expand and evolve in the future, leading to the construction of large-scale datasets in the field of TEAM. Instead of relying solely on the collection of real-world data, virtual patient generation offers an alternative approach to address the data scarcity issue in TEAM. This technique involves using generative models, such as generative adversarial networks[78,79] or GPT,[80,81] to create synthetic patient profiles. These models learn the underlying distribution of real patient data and generate virtual samples that capture the essential characteristics and variability of the original data. This augments the existing dataset with a wide range of symptoms, diagnoses, and treatment responses. However, it is crucial to validate the generated virtual patient data to ensure its quality and representativeness of real-world scenarios.

Complementing the quantitative expansion of data acquisition, the recent integration of advanced diagnostic devices into TEAM practice has facilitated the collection of more objective and quantifiable data. For instance, the utilization of diagnostic devices such as ultrasound imaging devices, digital stethoscopes, and electroencephalography devices enable the object collection of biosignals and imaging data. By integrating this information with laboratory results from blood tests, a more comprehensive and multimodal TEAM dataset can be constructed.

To fully harness the potential of the collected TEAM data, it is crucial to establish a robust platform that facilitates the accumulation, integration, and sharing of data from various clinical settings. This system should prioritize data privacy and security, employing stringent anonymization techniques to safeguard patient confidentiality. Simultaneously, the system should promote open access to the anonymized data, allowing researchers to freely explore, collaborate, and generate new hypotheses. This open-source approach will foster innovation, expand the scope of TEAM research, and accelerate the field's progress.

Effective utilization of clinical data hinges on the integration of information from various sources into a consistent format. Common Data Models[82] have long been the go-to strategy for this purpose, but the advent of LLMs is now offering innovative approaches to data standardization. First, as demonstrated by the TEMED-LLM methodology,[83] LLMs can effectively extract structured tabular data from unstructured textual medical reports, enabling the transformation of heterogeneous data sources into a consistent format and facilitating interoperability. Second, LLMs' ability to learn from and process unstructured data directly, such as clinical text summarization[84] and learning from electronic health records,[85] allows for the utilization of clinical data without the need for extensive data preprocessing.

The application of AI technologies in TEAM research offers a promising approach to address long-standing challenges while preserving its unique strengths. TEAM's highly abstract theoretical concepts, while facilitating intuitive reasoning and systemic pattern recognition, have led to significant inter-practitioner variability and difficulties in scientific validation. AI can bridge this gap by translating abstract concepts (low-dimensional, high-level features) into combinations of clinical phenotypes or molecular characteristics (high-dimensional, low-level features).[86] For instance, pattern identification, a unique diagnostic system of TEAM, can be modeled as a dimensionality reduction algorithm in ML.[26,65] This approach allows us to quantify how TEAM doctors combine various clinical symptoms to identify core patterns, similar to how dimensionality reduction algorithms extract key features from complex datasets. By analyzing this process, we can not only objectify theoretical concepts but also identify potentially valuable features that might be overlooked in traditional diagnostics. Moreover, as mentioned earlier, advanced medical technology now allows us to observe symptoms that were previously undetectable by naked sense and collect diverse biological data. By applying AI techniques to these rich, multidimensional datasets, researchers can uncover new subtypes within diseases and identify patterns that were not discernible through traditional methods alone.[87,88] These AI-derived clusters can then be compared with traditional TEAM classifications, potentially validating some aspects of traditional knowledge while also refining and expanding our understanding of disease mechanisms. This integration of AI with TEAM thus offers a pathway to enhance the precision, reliability, and scientific validity of traditional practices, contributing to theoretical advancement and potentially more effective, personalized treatment strategies.

Declaration of competing interest

The authors declare that they have no conflicts of interest.

CRediT authorship contribution statement

Hyojin Bae: Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Sa-Yoon Park:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Chang-Eop Kim:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Funding

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science & ICT (2020M3A9E4104380 and 2022R1F1A1068841) and supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2023-00248152).

Ethical statement

No ethical approval was required as this study did not involve human participants or laboratory animals.

Data availability

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used chatGPT and Claude in order to proofread and polish the English writing. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

1. Hu Q, Yu T, Li J, Yu Q, Zhu L, Gu Y. End-to-End syndrome differentiation of Yin deficiency and Yang deficiency in traditional Chinese medicine. *Comput Methods Programs Biomed*. 2019;174:9–15.
2. Liu G-P, Yan J-J, Wang Y-Q, Zheng W, Zhong T, Lu X, et al. Deep learning based syndrome diagnosis of chronic gastritis. *Comput Math Methods Med*. 2014;2014(1):938350.
3. Yuan L, Yang L, Zhang S, Xu Z, Qin J, Shi Y, et al. Development of a tongue image-based machine learning tool for the diagnosis of gastric cancer: a prospective multicentre clinical cohort study. *EClinicalMedicine*. 2023;57.
4. Zhang Q, Zhou J, Zhang B. Computational traditional Chinese medicine diagnosis: a literature survey. *Comput Biol Med*. 2021;133:104358.
5. Zhang Q, Zhou J, Zhang B. Graph based multichannel feature fusion for wrist pulse diagnosis. *IEEE J Biomed Health Inform*. 2020;25(10):3732–3743.
6. Park S-Y, Park M, Lee W-Y, Lee C-Y, Kim J-H, Lee S, et al. Machine learning-based prediction of Sasang constitution types using comprehensive clinical information and identification of key features for diagnosis. *Integr Med Res*. 2021;10(3):100668.
7. Jung W-M, Park I-S, Lee Y-S, Kim C-E, Lee H, Hahm D-H, et al. Characterization of hidden rules linking symptoms and selection of acupoint using an artificial neural network model. *Front Med*. 2019;13:112–120.
8. Lee W-Y, Lee Y, Lee S, Kim YW, Kim J-H. A machine learning approach for recommending herbal formulae with enhanced interpretability and applicability. *Biomolecules*. 2022;12(11):1604.
9. Park S-Y, Bae H, Jeong H-Y, Lee JY, Kwon Y-K, Kim C-E. Identifying novel subtypes of functional gastrointestinal disorder by analyzing nonlinear structure in integrative biopsychosocial questionnaire data. *J Clin Med*. 2024;13(10):2821.
10. Lee W-Y, Lee C-Y, Kim Y-S, Kim C-E. The methodological trends of traditional herbal medicine employing network pharmacology. *Biomolecules*. 2019;9(8):362.
11. Lee W-Y, Lee C-Y, Kim C-E. Predicting activatory and inhibitory drug–target interactions based on structural compound representations and genetically perturbed transcriptomes. *PLoS One*. 2023;18(4):e0282042.
12. Li D, Hu J, Zhang L, Li L, Yin Q, Shi J, et al. Deep learning and machine intelligence: new computational modeling techniques for discovery of the combination rules and pharmacodynamic characteristics of traditional chinese medicine. *Eur J Pharmacol*. 2022;933:175260.
13. Lin Y, Zhang Y, Wang D, Yang B, Shen Y-Q. Computer especially AI-assisted drug virtual screening and design in traditional Chinese medicine. *Phytomedicine*. 2022;107:154481.
14. Park S-Y, Kim K-S, Lee W-Y, Kim C-E, Lee S. Integrative approach to identifying system-level mechanisms of Chung-Sang-Bo-Ha-Hwan's influence on respiratory tract diseases: A network pharmacological analysis with experimental validation. *Plants*. 2023;12(17):3024.
15. Park S-Y, Lee YY, Kim MH, Kim C-E. Deciphering the systemic impact of herbal medicines on allergic rhinitis: a network pharmacological approach. *Life*. 2024;14(5):553.
16. Qiu Q, Huang Y, Liu X, Huang F, Li X, Cui L, et al. Potential therapeutic effect of traditional Chinese medicine on coronavirus disease 2019: a review. *Front Pharmacol*. 2020;11:570893.
17. He X, Zhao L, Zhong W, Chen H-Y, Shan X, Tang N, et al. Insight into potent leads for alzheimer's disease by using several artificial intelligence algorithms. *Biomed Pharmacother*. 2020;129:110360.
18. Lee W-Y, Lee C-Y, Lee J-S, Kim C-E. Identifying Candidate flavonoids for non-alcoholic fatty liver disease by network-based strategy. *Front Pharmacol*. 2022;13:892559.
19. Li S, Zhang Z, Wu L, Zhang X, Li Y, Wang Y. Understanding ZHENG in traditional Chinese medicine in the context of neuro-endocrine-immune network. *IET Syst Biol*. 2007;1(1):51–60.
20. Tran MN, Kim S, Nguyen QHN, Lee S. Molecular mechanisms underlying qi-invigorating effects in traditional medicine: network pharmacology-based study on the unique functions of qi-invigorating herb group. *Plants*. 2022;11(19):2470.
21. Park S-Y, Kim YW, Song YR, Bak SB, Jang YP, Kim I-K, et al. Compound-level identification of sasang constitution type-specific personalized herbal medicine using data science approach. *Heliyon*. 2023;9(2).
22. Lee W-Y, Lee C-Y, Kim C-E, Kim J-H. Investigating the biomarkers of the sasang constitution via network pharmacology approach. *Evid-Based Complement Alternat Med*. 2021;2021:1–10.
23. Jang D-Y, Oh K-C, Jung E-S, Cho S-J, Lee J-Y, Lee Y-J, et al. Diversity of acupuncture point selections according to the acupuncture styles and their relations to theoretical elements in traditional asian medicine: a data-mining-based literature study. *J Clin Med*. 2021;10(10):2059.
24. Bae H, Kim C, Lee C, Shin S, Kim J. Investigation of the possibility of research on medical classics applying text mining - focusing on the Huangdi's internal classic. *J Korean Med Classics*. 2018;31(4):27–46.
25. Hyojin Bae C-EK. Geometrical analysis of Geum-Won era prescriptions: a novel approach to traditional Asian medicine's basic theory. *J Physiol Pathol Korean Med*. 2023;37(5):129–133.
26. Bae H, Lee S, Lee C-y, Kim C-E. A novel framework for understanding the pattern identification of traditional asian medicine from the machine learning perspective. *Front Med (Lausanne)*. 2022;8:763533.
27. Wang F, Preininger A. AI in health: state of the art, challenges, and future directions. *Yearb Med Inform*. 2019;28(01):016–026.
28. Kim S-K, Lee M-K, Jang H, Lee J-J, Lee S, Jang Y, et al. TM-MC 2.0: an enhanced chemical database of medicinal materials in Northeast Asian traditional medicine. *BMC Complement Med Ther*. 2024;24(1):40.
29. Ru J, Li P, Wang J, Zhou W, Li B, Huang C, et al. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *J Cheminform*. 2014;6:1–6.

30. Liu Z, Guo F, Wang Y, Li C, Zhang X, Li H, et al. BATMAN-TCM: a bioinformatics analysis tool for molecular mechanism of traditional Chinese medicine. *Sci Rep*. 2016;6(1):21146.
31. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *Egems*. 2016;4(1).
32. Kim K-H, Choi W, Ko S-J, Chang D-J, Chung Y-W, Chang S-H, et al. Multi-center healthcare data quality measurement model and assessment using OMOP CDM. *Appl Sci*. 2021;11(19):9188.
33. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. *Nat Commun*. 2020;11(1):3673.
34. Park M, Kim MH, Park S-Y, Choi I, Kim C-E. Individualized diagnosis and prescription in traditional medicine: decision-making process analysis and machine learning-based analysis tool development. *Am J Chin Med (Gard City N Y)*. 2022;50(07):1827–1844.
35. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338.
36. Potdar K, Pardawala TS, Pai CD. A comparative study of categorical variable encoding techniques for neural network classifiers. *Int J Comput Appl*. 2017;175(4):7–9.
37. Wah YB, Ibrahim N, Hamid HA, Abdul-Rahman S, Fong S. Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika J Sci Technol*. 2018;26(1).
38. Kursa MB, Jankowski A, Rudnicki WR. Boruta–a system for feature selection. *Fundam Inform*. 2010;101(4):271–285.
39. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30.
40. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:181112808 2018.
41. Liu Z, Luo C, Fu D, Gui J, Zheng Z, Qi L, et al. A novel transfer learning model for traditional herbal medicine prescription generation from unstructured resources and knowledge. *Artif Intell Med*. 2022;124:102232.
42. Wang X, Wang X, Lou Y, Liu J, Huo S, Pang X, et al. Constructing tongue coating recognition model using deep transfer learning to assist syndrome diagnosis and its potential in noninvasive ethnopharmacological evaluation. *J Ethnopharmacol*. 2022;285:114905.
43. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. *Nat Biomed Eng*. 2022;6(12):1346–1352.
44. Sowrirajan H, Yang J, Ng AY, Rajpurkar P. *Medical Imaging with Deep Learning*. Moco pretraining improves representation and transferability of chest x-ray models. PMLR; 2021.
45. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–357.
46. Singh G, Mémoli F, Carlsson GE. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics*. 2007;2:091–100.
47. Li J, Hu X, Tu L, Cui L, Jiang T, Cui J, et al. *Diabetes Tongue Image Classification Using Machine Learning and Deep Learning*; 2021.
48. Jiang T, Hu X-j, Yao X-h, Tu L-p, Huang J-b, Ma X-x, et al. Tongue image quality assessment based on a deep convolutional neural network. *BMC Med Inform Decis Mak*. 2021;21(1):147.
49. Tang W, Gao Y, Liu L, Xia T, He L, Zhang S, et al. An automatic recognition of tooth-marked tongue based on tongue region detection and tongue landmark detection via deep learning. *IEEE Access*. 2020;8:153470–153478.
50. Sun X, Qian H, Xiong Y, Zhu Y, Huang Z, Yang F. Deep learning-enabled mobile application for efficient and robust herb image recognition. *Sci Rep*. 2022;12(1):6579.
51. Roopashree S, Anitha J. DeepHerb: a vision based system for medicinal plants using xception features. *IEEE Access*. 2021;9:135927–135941.
52. Yan J, Cai X, Chen S, Guo R, Yan H, Wang Y. Ensemble learning-based pulse signal recognition: classification model development study. *JMIR Med Inform*. 2021;9(10):e28039.
53. Hu X, Zhu H, Xu J, Xu D, Dong J. *2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*. Wrist pulse signals analysis based on deep convolutional neural networks; 2014 21-24 May 2014.
54. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2021;14(1):49–58.
55. Kiyasseh D, Cohen A, Jiang C, Altieri N. A framework for evaluating clinical artificial intelligence systems without ground-truth annotations. *Nat Commun*. 2024;15(1):1808.
56. Xue C, Yu L, Chen P, Dou Q, Heng P-A. Robust medical image classification from noisy labeled data with global and local representation guided co-training. *IEEE Trans Med Imaging*. 2022;41(6):1371–1382.
57. Çallı E, Murphy K, Kurstjens S, Samson T, Herpers R, Smits H, et al. Deep learning with robustness to missing data: A novel approach to the detection of COVID-19. *PLoS One*. 2021;16(7):e0255301.
58. Spiess A-N, Neumeyer N. An evaluation of R 2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC Pharmacol*. 2010;10:1–11.
59. Huang H, Wang Y, Rudin C, Browne EP. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Commun Biol*. 2022;5(1):719.
60. Wang Y, Yao H, Zhao S. Auto-encoder based dimensionality reduction. *Neurocomputing*. 2016;184:232–242.
61. Ribeiro MT, Singh S, Guestrin C. Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016.
62. Jung D-H, Kim H-Y, Won JH, Park SH. Development of a classification model for Cynanchum wilfordii and Cynanchum auriculatum using convolutional neural network and local interpretable model-agnostic explanation technology. *Front Plant Sci*. 2023;14:1169709.
63. Kim J, Jeong K, Lee S, Baek Y. Machine-learning model predicting quality of life using multifaceted lifestyles in middle-aged South Korean adults: a cross-sectional study. *BMC Public Health*. 2024;24(1):159.
64. Causality in digital medicine. *Nat Commun*. 2021;12(1):5471.
65. Lee H, Choi Y, Son B, Lim J, Lee S, Kang JW, et al. Deep autoencoder-powered pattern identification of sleep disturbance using multi-site cross-sectional survey data. *Front Med (Lausanne)*. 2022;9:950327.
66. Han T, Adams LC, Papaioannou J-M, Grundmann P, Oberhauser T, Löser A, et al. MedAlpaca–an open-source collection of medical conversational AI models and training data. arXiv preprint arXiv:230408247 2023.
67. Pu H, Mi J, Lu S, He J. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. RoKEPG: RoBERTa and knowledge enhancement for prescription generation of traditional Chinese medicine. IEEE; 2023.
68. Yang G, Shi J, Wang Z, Liu X, Wang G. TCM-GPT: Efficient pre-training of large language models for domain adaptation in traditional Chinese medicine. arXiv preprint arXiv:231101786 2023.
69. Kang B, Kim J, Yun T-R, Kim C-E. Prompt-RAG: pioneering vector embedding-free retrieval-augmented generation in niche domains, exemplified by Korean medicine. arXiv preprint arXiv:240111246 2024.
70. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940.
71. SL Bongsu Kang, Bae Hyojin, Kim Chang-Eop. Current status and direction of generative large language model applications in medicine - focusing on East Asian medicine. *J Physiol Pathol Korean Med*. 2024;38(2):49–58.
72. Jang D, Yun T-R, Lee C-Y, Kwon Y-K, Kim C-E. GPT-4 can pass the Korean National Licensing Examination for Korean Medicine Doctors. *PLOS Digit Health*. 2023;2(12):e0000416.
73. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. arXiv preprint arXiv:231116452 2023.
74. Kim J, Lee H-Y, Kim J-H, Kim C-E, Kim J, Lee H-Y, et al. Pilot development of a'clinical performance examination (CPX) practicing chatbot' utilizing prompt engineering. *J Korean Med*. 2024;45(1):200–212.
75. Kim T-H, Kang JW, Lee MS. AI Chat bot-ChatGPT-4: A new opportunity and challenges in complementary and alternative medicine (CAM). *Integr Med Res*. 2023;12(3):100977.
76. Lee S. *Collection of Clinical Big Data and Construction of Service Platform for Developing Korean Medicine Doctor with Artificial Intelligence*. Seoul, Korea: Korea Institute of Oriental Medicine; 2023.
77. Jang B-H. *Development of Smart EMR System for Data Collection with the Purpose of Building of Korean Medicine CDSS*. Dajeon, Korea: Kyung Hee University; 2021.
78. Li J, Cairns BJ, Li J, Zhu T. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *NPJ Digit Med*. 2023;6(1):98.
79. Jadon A, Kumar S. *2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*. Leveraging generative ai models for synthetic data generation in healthcare: Balancing research and privacy. IEEE; 2023.
80. Pang C, Jiang X, Pavinkurve NP, Kalluri KS, Minto EL, Patterson J, et al. CEHR-GPT: generating electronic health records with chronological patient timelines. arXiv preprint arXiv:240204400 2024.
81. Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, Yeung JA, et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit Health*. 2024;6(4):e281–ee90.
82. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Informat Assoc*. 2015;22(3):553–564.
83. Bisercic A, Nikolic M, van der Schaar M, Delibasic B, Lio P, Petrovic A. Interpretable medical diagnostics with structured data extraction by large language models. arXiv preprint arXiv:230605052 2023.
84. Van Veen D, Van Uden C, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med*. 2024:1–9.
85. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5(1):194.
86. Hyojin Bae C-EK. The unexpected role of deep learning in Traditional Korean Medicine (TKM): the science of deep learning can help better understand TKM. *Korean J Sub-health Med*. 2023;4(1):44–50.
87. Guo F, Tang X, Zhang W, Wei J, Tang S, Wu H, et al. Exploration of the mechanism of traditional Chinese medicine by AI approach using unsupervised machine learning for cellular functional similarity of compounds in heterogeneous networks, XiaoErFuPi granules as an example. *Pharmacol Res*. 2020;160:105077.
88. Jafari M, Wang Y, Amiryousefi A, Tang J. Unsupervised learning and multipartite network models: a promising approach for understanding traditional medicine. *Front Pharmacol*. 2020;11:563852.