Am J Transl Res 2025;17(11):8742-8756
www.ajtr.org /ISSN:1943-8141/AJTR0168918

Original Article

Comprehensive evaluation and clinical implications of kernel extreme learning machine long short term memory transformer framework

Wentao Zheng[1*], Yang Pan[2*], Ying Wang[3], Ling Zhu[4]

[1]The First Affiliated Hospital of Dali University, Dali 671000, Yunnan, China; [2]Internal Medicine, Qujing Third People's Hospital, Qujing 655000, Yunnan, China; [3]The First Affiliated Hospital of Dali University, Dali 671000, Yunnan, China; [4]Department of Basic Medicine, Qujing University of Medicine & Health Sciences, Qujing 655011, Yunnan, China. [*]Equal contributors.

Received September 20, 2025; Accepted October 30, 2025; Epub November 15, 2025; Published November 30, 2025

Abstract: Objectives: To develop and validate a hybrid deep learning model to enhance diagnostic and predictive accuracy of Alzheimer's disease (AD) using readily available clinical data. Methods: A triple-architecture joint model was constructed, integrating a Kernel Extreme Learning Machine (KELM), a Long Short-Term Memory (LSTM) network, and a Transformer. This framework was designed to capture nonlinear associations, temporal dynamics, and global feature dependencies. The model was trained and validated on 2,149 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and externally tested on an independent cohort of 1,012 subjects. Results: In internal validation, the model achieved state-of-the-art performance with 95.42% accuracy, 95.63% recall, and an area under the curve (AUC) of 0.981. It also demonstrated strong generalizability in the external cohort, achieving 93.81% accuracy and an AUC of 97.25%. In a longitudinal sub-analysis, the model accurately predicted the 3-year conversion from mild cognitive impairment to AD with 92.78% accuracy. Ablation analyses confirmed the essential contribution of each model component. Conclusions: The proposed KELM-LSTM-Transformer model provides a powerful and robust framework for AD prediction. Its high accuracy and strong generalizability suggest potential as an effective and accessible tool for early risk stratification, supporting timely clinical interventions.

Keywords: Kernel extreme learning machine, long short term memory network, transformer, Alzheimer's disease

Introduction

Alzheimer's disease (AD), the most prevalent neurodegenerative disorder worldwide, imposes a substantial burden on patients, families, and healthcare systems due to its irreversible cognitive decline. With global population aging, the incidence of AD continues to rise, making it a critical public health challenge [1, 2]. The disease progresses insidiously, with key pathological changes - such as amyloid-β (Aβ) deposition and tau protein hyperphosphorylation - occurring years or even decades before clinical symptoms become apparent [3, 4]. Traditional diagnostic approaches rely heavily on clinical manifestations and neuropsychological assessments, leading to diagnoses typically made in the middle or late stages [5]. By then, neuronal damage is largely irreversible, leaving a narrow window for intervention. Consequently, early and accurate risk prediction and preclinical identification have become the central goals and major challenges of current AD research.

Biomarker-based subclinical testing is widely regarded as a crucial strategy for overcoming this limitation [6, 7]. However, biomarker data are high-dimensional and heterogeneous, with complex nonlinear relationships and significant interindividual variability. These characteristics demand advanced analytical tools capable of uncovering deep correlations and enhancing predictive performance.

Machine learning (ML) has revolutionized AD research due to its capacity to handle large-scale, high-dimensional datasets and to identify intricate nonlinear patterns [8-10]. It enables

https://doi.org/10.62347/ELRA9110

efficient multimodal data integration and the construction of comprehensive predictive frameworks that transcend the limitations of single indicators. ML can markedly improve the early recognition of AD and the prediction of mild cognitive impairment (MCI) conversion while extending the prediction window for asymptomatic individuals [11, 12]. Deep learning models - such as convolutional (CNN) and recurrent neural networks (RNN) - can automatically extract subtle atrophic or metabolic abnormalities from MRI or PET images, capturing early pathological signals that may escape human observation [13-15]. Ensemble methods like random forest and XGBoost effectively integrate heterogeneous data from multiple modalities and quantify individualized risk probabilities [16, 17]. These models can enhance clinical decision-making, optimize patient screening, and provide evidence for personalized prevention strategies, thereby driving a paradigm shift in AD management toward a predictive, preventive, and personalized model.

In recent years, the application of ML algorithms for AD and prodromal-stage prediction has become a research focus, leading to diverse methodological developments. Support vector machines (SVMs) are frequently used to analyze structural MRI features for early AD classification [18, 19]. Convolutional neural networks (CNNs) show excellent performance in processing PET or hippocampal images to identify subtle pathological alterations [20, 21]. Random forest models are favored for integrating heterogeneous multimodal data and performing feature selection [22]. Graph convolutional networks (GCNs) are particularly suited for modeling brain functional and structural connectivity networks, capturing coordinated abnormalities across AD-related regions [23, 24]. Long short-term memory (LSTM) networks effectively model longitudinal follow-up data, tracking dynamic trajectories of cognitive or structural decline to predict conversion risk [25].

Although these methods have achieved progress within their respective domains, the complexity and multimodal interdependence of AD pathophysiology make it difficult for any single model to comprehensively capture cross-modal information, spatiotemporal dynamics, and long-range dependencies [26, 27]. To address these limitations, this study proposes a novel Kernel Extreme Learning Machine (KELM)-LSTM-Transformer hybrid model that integrates the efficient feature extraction and nonlinear mapping capabilities of the KELM, the temporal modeling strength of LSTM, and the global contextual learning power of the Transformer. This end-to-end framework is designed to comprehensively fuse complementary multimodal information and uncover complex interaction patterns across time and feature dimensions, thereby enhancing the accuracy and robustness of early AD identification and progression prediction.

Materials and data sources

The data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI, launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner, MD, aims to determine whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org.

Data collection

The training and internal validation dataset included 2,149 participants from the ADNI-1, ADNI-GO, and ADNI-2 phases. All participants were diagnosed according to the NINCDS-ADRDA (National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association) criteria. The cohort consisted of three clinical groups: cognitively normal (CN) controls, patients with MCI, and patients with probable AD. The external validation cohort, comprising 1,012 independent participants, was drawn from the ADNI-3 phase to ensure a rigorous assessment of the model's generalizability. Although longitudinal data within ADNI span over a decade for many participants, our predictive models were developed using baseline and early-stage data. The demographic composition of the combined cohort was predominantly Caucasian (approximately 85%), with smaller representations of African Ame-

Table 1. The various indicators of the dataset

Category	Variables	Description
Patient Identification	Patient ID	Unique identifier for each patient.
Demographics	Age, Gender, Ethnicity, Education Level	Age in years, binary Gender, coded Ethnicity, coded Education level.
Lifestyle Factors	BMI, Smoking, Alcohol Consumption, Physical Activity, Diet Quality, Sleep Quality	Body Mass Index, Smoking status, Weekly alcohol units, Weekly activity hours, Diet score, Sleep score.
Medical History	Family History Alzheimers, Cardiovascular Disease, Diabetes, Depression, Head Injury, Hypertension	Binary indicators for presence of specific conditions.
Clinical Measurements	Systolic BP, Diastolic BP, Cholesterol Total, Cholesterol low-density lipoprotein, Cholesterol high-density lipoprotein, Cholesterol Triglycerides	Blood pressure readings and detailed cholesterol levels.
Cognitive Assessments	Mini-Mental State Examination, Functional Assessment, Memory Complaints, Behavioral Problems, activities of daily living	Cognitive test score, Functional ability score, Binary memory/behavior complaints, Daily living score.
Symptoms	Confusion, Disorientation, Personality Changes, Difficulty Completing Tasks, Forgetfulness	Binary indicators for presence of specific Alzheimer's-related symptoms.
Diagnosis	Diagnosis	Binary outcome: 1. Alzheimer's Disease, 2. No Alzheimer's Disease.

rican (7%), Hispanic (5%), and other ethnicities (3%).

Variables included in the model

The ADNI dataset provides extensive clinical and demographic information, including age, sex, race, education level, and lifestyle factors such as body mass index (BMI), smoking status, and physical activity. A family history of neurological and metabolic diseases was also recorded. Key clinical and cognitive assessments related to AD evaluation were incorporated (see **Table 1**).

Among the clinical measurements, several blood biomarkers were integrated into our model due to their established associations with cardiovascular and cognitive health. These included total cholesterol, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, and triglycerides. These widely available parameters provide valuable systemic health insights that contribute to the model's comprehensive risk stratification.

This structured dataset enables the analysis of continuous variables such as BMI, alcohol intake, physical activity, and clinical scores. Categorical variables were encoded in a binary format (e.g., gender, smoking status, medical history, and diagnosis), where a diagnostic value of 1 indicated the presence of AD and 2 indicated its absence. This structure supports robust risk factor analysis, predictive modeling, and statistical evaluation of AD progression and related health outcomes.

KELM: efficient nonlinear classification with kernel methods

The KELM is a ML algorithm derived from the Extreme Learning Machine (ELM) framework, enhanced by the integration of kernel methods. The network architecture of KELM is illustrated in **Figure 1**.

Its key principle lies in randomly initializing and fixing the hidden-layer parameters of a Single Hidden Layer Feedforward Neural Network, while employing kernel techniques to map input data nonlinearly into a high-dimensional feature space. This approach mitigates the instability often caused by random hidden node generation in traditional ELMs and significantly enhances the model's capability to capture complex nonlinear relationships [28].

KELM implicitly defines a nonlinear mapping from the input space to a reproducing kernel Hilbert space via a Mercer kernel. In this high-dimensional space, data becomes more linearly separable. The model then solves for the output weight matrix directly, without iterative optimization of hidden parameters. This is achieved by constructing a kernel matrix through pairwise inner products of training samples in the feature space, effectively replacing the hidden-layer output matrix used in conventional ELMs.

Subsequently, the regularized least squares method is applied to derive a closed-form solution for the output weights. The inclusion of a

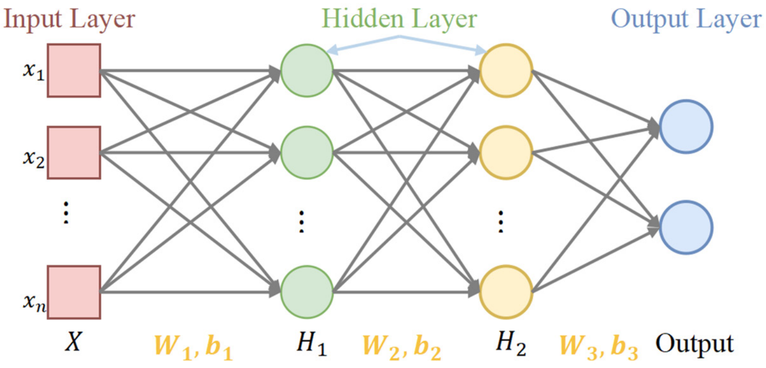

Figure 1. The network structure of the Kernel Extreme Learning Machine. The diagram illustrates the KELM architecture, which consists of an input layer, a hidden layer where input data is implicitly mapped to a high-dimensional feature space via a kernel function (e.g., radial basis function kernel), and an output layer. The output weights are calculated analytically, not iteratively trained. KELM: Kernel Extreme Learning Machine.

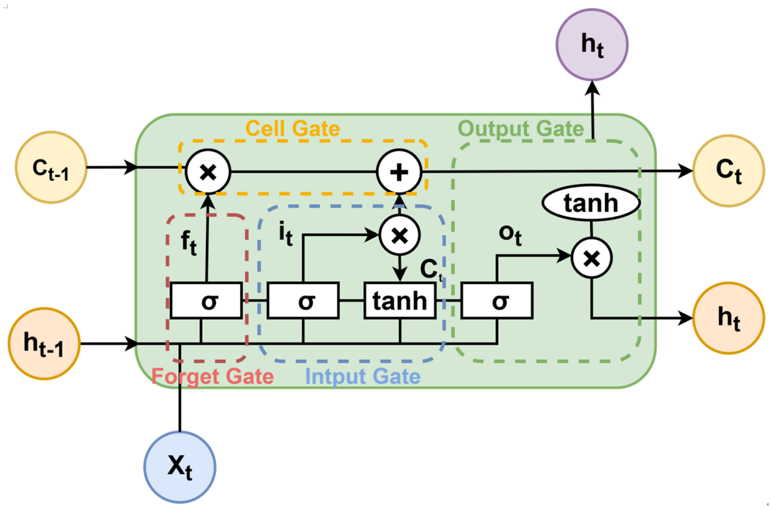

Figure 2. The network structure of the LSTM unit. This schematic shows the core components of an LSTM cell, including the cell state, which acts as a memory conveyor, and the three primary gates: the forget gate, the input gate, and the output gate. These gates regulate the flow of information, enabling the network to learn long-term dependencies. KELM: Kernel Extreme Learning Machine, LSTM: Long Short-Term Memory.

regularization term in this solution helps control model complexity and improve generalization. Using the derived weights and the same kernel function, KELM predicts outcomes for new samples efficiently.

KELM offers several advantages: extremely fast training speed, flexible handling of nonlinear problems, and stable generalization performance. These strengths have led to its wide application in classification, regression, and feature learning tasks across diverse research fields.

LSTM: a powerful architecture for long-term sequence learning

LSTM is a specialized form of the recurrent neural network (RNN) designed to address the vanishing and exploding gradient problems that commonly occur when processing long sequential data, thereby enabling the effective learning of long-term dependencies [29, 30]. The network structure of LSTM is shown in **Figure 2**, and its core innovation lies in the gating mechanism and cell state structure.

The cell state C_t, functioning as an information highway through the entire time series, allows for the relatively lossless propagation of information:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \qquad (4.1)$$

An LSTM unit contains three gates, the forget gate, input gate, and output gate, each implemented using a sigmoid activation function and element-wise multiplication.

The forget gate determines which information from the previous cell state should be discarded:

$$f_t = \sigma\,(W_f \cdot [h_{t-1}, x_t] + b_f) \qquad (4.2)$$

Values of f_t close to 0 indicate "forget", while values near 1 indicate "retain".

The input gate controls which new information is added to the cell state:

$$i_t = \sigma\,(W_i \cdot [h_{t-1}, x_t] + b_i) \qquad (4.3)$$

$$\tilde{C}_t = \tanh\,(W_c \cdot [h_{t-1}, x_t] + b_c) \qquad (4.4)$$

Here, i_t regulates the update ratio, while $\tilde{C}_t$ represents the candidate cell state value generated by the tanh layer. The cell state is then updated by combining the effects of forgetting

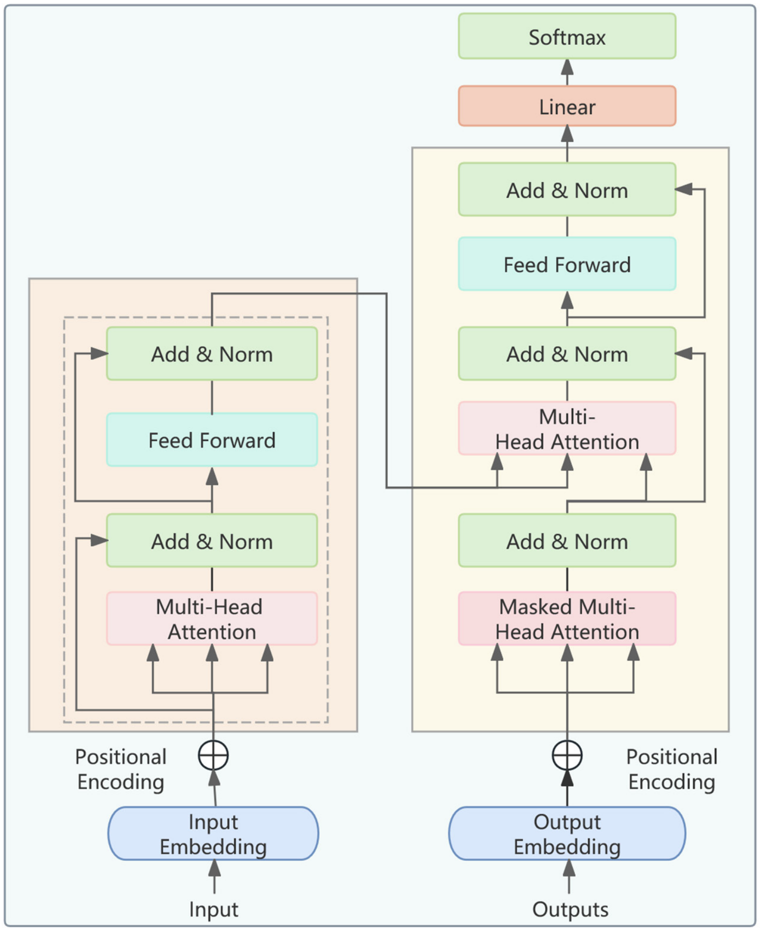

Figure 3. The network structure of the Transformer model. The figure depicts the overall architecture of the Transformer, highlighting its encoder-decoder structure. Key components such as multi-head self-attention mechanisms, position-wise feed-forward networks, positional encodings, and residual connections with layer normalization are shown.

old information and incorporating new information, as described in Equation (4.1).

The output gate determines which parts of the cell state contribute to the hidden state:

$$o_t = \sigma\,(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{4.5}$$

$$h_t = o_t * \tanh(C_t) \tag{4.6}$$

Thus, the new hidden state h_t is a gated, nonlinear transformation of the updated cell state. This precise gating mechanism enables LSTM networks to autonomously learn when to remember, forget, and output information, effectively capturing both short-term and long-term temporal dependencies in sequential data.

Transformer: a revolutionary architecture for sequence data processing

The transformer is a revolutionary neural network architecture that fundamentally transformed sequence modeling and natural language processing [34]. Its structure, shown in **Figure 3**, eliminates traditional recurrence (RNN) and convolution (CNN), relying instead on the self-attention mechanism to model dependencies among all elements in a sequence. The model consists of stacked encoder and decoder layers.

The encoder processes an input sequence (e.g., a sentence) by embedding tokens and incorporating positional encodings to preserve order information:

$$E = X + P \tag{4.7}$$

where *X* is the token embedding matrix and *P* represents the positional encodings. This input is passed through a multi-head self-attention mechanism that computes:

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.8}$$

Here, Q, K, and V are learned linear projections of E, and d_k is the dimension of the key vectors. Multi-head attention concatenates multiple attention heads and projects the result:

$$\text{MultiHead}\,(E) = \text{Contat}(\text{head}_1, \ldots \text{head}_h)W^o \tag{4.9}$$

$$\text{head}_i = \text{Attention}\,(EW_i^Q, EW_i^k, EW_i^v) \tag{4.10}$$

The attention output is then passed through a position-wise feedforward network (FFN), with residual connections and layer normalization applied after each sublayer to stabilize training:

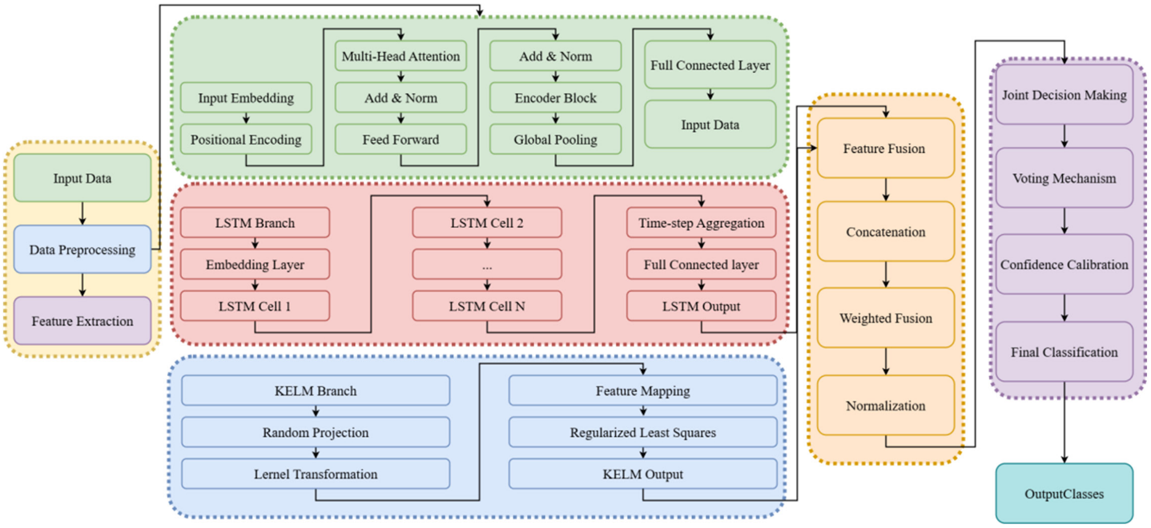

Figure 4. The network structure of the proposed KELM-LSTM-Transformer joint model. This flowchart illustrates how input data is processed in parallel by the LSTM and Transformer modules to extract temporal and global contextual features, respectively. These features are then concatenated and fed into the KELM classifier for the final prediction. KELM: Kernel Extreme Learning Machine, LSTM: Long Short-Term Memory.

$$H_1 = \mathrm{LayerNorm}(E+\mathrm{MultiHead}(E)) \quad (4.11)$$

$$H_2 = \mathrm{LayerNorm}(H_1+\mathrm{FFN}(H_1)) \quad (4.12)$$

where $FFN(x) = max(0,xW_1+b_1)W_2+b_2$ is the position-wise feedforward transformation. This design allows each token to integrate global contextual information and undergo nonlinear transformation, with residual paths ensuring stable gradient flow during training. Positional encodings compensate for the lack of inherent order sensitivity in self-attention mechanisms.

The decoder has a structure similar to the encoder but introduces two critical modifications: (1) Masked self-attention to prevent the model from attending to future positions during autoregressive generation, implemented as:

$$\mathrm{MaskedAttention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V \quad (4.13)$$

where M is a mask with large negative values that suppress attention to future tokens. (2) Cross-attention, allowing the decoder to attend to the encoder outputs:

$$\mathrm{CrossAttention}(Q_d,K_e,V_e) = \mathrm{Attention}(Q_d,K_e,V_e) \quad (4.14)$$

The decoder produces hidden representations h_t and predicts the next token probabilities using a linear projection followed by softmax:

$$P(y_t|y_{<t},x) = \mathrm{softmax}(w_o,h_t) \quad (4.15)$$

The model is trained end-to-end by minimizing the cross-entropy loss, equivalent to maximizing the likelihood of the correct target sequence:

$$\Gamma = -\sum_t \log P(y_t^* \mid y_{<t}^*, x) \quad (4.16)$$

The Transformer's parallel computation and global attention mechanism enable it to capture long-range dependencies efficiently, making it the cornerstone of modern large language models [31-33].

KELM-LSTM-Transformer: a hybrid approach for multimodal feature fusion and classification

The KELM-LSTM-Transformer joint model integrates the temporal modeling capability of LSTM, the global dependency learning power of the Transformer, and the efficient nonlinear classification of KELM to form a multimodal feature fusion and classification framework. The network architecture is illustrated in **Figure 4**.

In the first stage, input data undergo standardized preprocessing and are simultaneously fed into two parallel feature extraction modules: LSTM and Transformer.

The LSTM module, leveraging its unique gating mechanism, sequentially processes the input data, retains long-term dependencies in the cell state, and extracts the hidden state from the final time step as temporal features.

In parallel, the Transformer module employs a multi-head self-attention mechanism to compute global correlation weights between sequence elements. Positional encoding preserves sequence order, while stacked encoder layers, comprising self-attention and feedforward sublayers, capture contextual relationships. Average pooling is then applied to generate high-level contextual feature representations.

These two modules yield complementary features: the LSTM captures local dynamic evolution, while the Transformer encodes global structural relationships across the data.

In the feature fusion stage, the temporal feature vectors from the LSTM and contextual vectors from the Transformer are concatenated to form a comprehensive multimodal feature representation. This fused feature vector is then passed to the KELM classifier for final decision-making.

Within KELM, features are first projected into a high-dimensional space through random mapping. The radial basis function (RBF) kernel is then used to compute nonlinear similarities between samples and construct the kernel matrix. The output weights are obtained by solving a regularized least squares problem, producing a closed-form solution.

This hybrid design combines the deep feature extraction capabilities of neural networks with the computational efficiency of extreme learning machines. The complementary information captured by LSTM and Transformer enhances the model's representational power for complex temporal-contextual patterns, while the KELM component provides robust nonlinear decision boundaries. As a result, the proposed system achieves high classification accuracy while substantially reducing the computational overhead of backpropagation training compared to conventional deep learning models.

Statistical analysis

All statistical analyses were conducted using MATLAB R2024a. Model performance was comprehensively evaluated using standard metrics, including accuracy, precision, recall (sensitivity), F1 score, and area under the receiver operating characteristic curve (AUC).

(1) Accuracy represented the proportion of correctly classified samples. (2) Precision was defined as the ratio of true positives to the sum of true and false positives. (3) Recall (sensitivity) was the ratio of true positives to the sum of true positives and false negatives. (4) The F1 score was calculated as the harmonic mean of precision and recall. (5) The AUC was used to evaluate overall discriminative performance.

All performance metrics for the internal test set were averaged over three independent experimental runs to ensure result stability and reproducibility. For the external validation cohort, 95% confidence intervals (CIs) of the performance metrics were estimated using bootstrapped resampling (1,000 iterations). Descriptive statistics were used to summarize the baseline characteristics of the study cohorts.

Results

In the classification experiments of the KELM-LSTM-Transformer hybrid model, the following parameter configurations were used.

The hardware setup comprised an NVIDIA RTX 3090 GPU (24 GB VRAM), an AMD Ryzen 95950X processor, and 64 GB DDR4 memory. The software environment was MATLAB R2024a.

For model parameters, the LSTM layer contained 128 hidden units with a fixed sequence length of 50. The Transformer encoder consisted of two stacked layers, each equipped with a 4-head self-attention mechanism and a 256-dimensional feedforward network. The KELM classifier employed a RBF kernel $\gamma = 0.01$ and a regularization coefficient $\lambda = 1e\text{-}5$.

During training, the Adam optimizer (initial learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$) was used with a batch size of 64. Early stopping was triggered when validation loss failed to decrease for 10 consecutive epochs, with a maximum of 200 iterations. The cross-entropy loss function was combined with L2 weight decay for regularization. Weighted feature fusion was applied between the two feature extraction modules, assigning an LSTM weight of 0.4 and a Transformer weight of 0.6. The

Table 2. The specific experimental parameter settings

Specific parameter items	Parameter values
LSTM module	
Number of hidden units in LSTM	One hundred and twenty-eight
Input sequence length	Fifty
Transformer encoder module	
Encoder layers (N)	Two
Number of Attention Heads (h)	Four
Feedforward network dimension (d_ff)	Two hundred and fifty-six
Model dimension (d_model)	One hundred and twenty-eight
Dropout rate	Zero point one
KELM classifier	
kernel function	Radial basis function
Nuclear parameter (γ)	Zero point zero one
Regularization coefficient (λ/C)	Zero point zero zero zero zero one
Combined	Weighted feature fusion
Optimizer	Adam
Initial learning rate (lr)	Zero point zero zero one
Adam parameters (β_1, β_2)	(0.9, 0.999)
Batch Size	Sixty-four
Maximum number of training epochs (Epochs)	Two hundred
Early Stop Mechanism (Patience)	Ten
Loss function	CrossEntropyLoss
Weight decay (L2 regularization coefficient)	Zero point zero zero zero one

KELM: Kernel Extreme Learning Machine, LSTM: Long Short-Term Memory.

detailed parameter settings are presented in **Table 2**.

Comparative model performance

In this study, nine advanced ML models were evaluated for structured (Excel-type) data classification. The core modelwas the KELM-LSTM Transformer joint architecture, integrating the nonlinear classification capability of KELM, the temporal feature extraction power of LSTM, and the global dependency modeling ability of the Transformer.

The comparative models included three transformer-based architectures optimized for tabular data: tabtransformer, tailored for categorical-numerical feature integration; Self-Attention with Interaction Transformer (SAINT), employing dual row-column attention; and Feature Tokenizer Transformer (FT-Transformer), which enables high-dimensional feature interaction via tokenization [35].

Five additional advanced models were included for benchmarking: XGBoost-Line, combining gradient boosting trees and linear models; Neural Oblivious Decision Trees (NODE), integrating differentiable decision trees; DeepFM, which unifies deep neural networks and factorization machines; TabNet, an interpretable model based on sequential attention; and AutoInt, an automatic feature interaction network.

All models were trained and tested under identical data partitions, and each experiment was repeated three times for consistency. The performance results for the KELM-LSTM-Transformer model are illustrated in **Figure 5**, and overall metrics are summarized in **Table 3**.

Model performance was comprehensively evaluated using five indicators: Accuracy, reflecting overall prediction correctness; Precision, measuring the reliability of positive classifications; Recall (Sensitivity), assessing the completeness of positive sample detection; F1 score, balancing precision and recall; and Area Under the ROC Curve (AUC), quantifying discriminative ability.

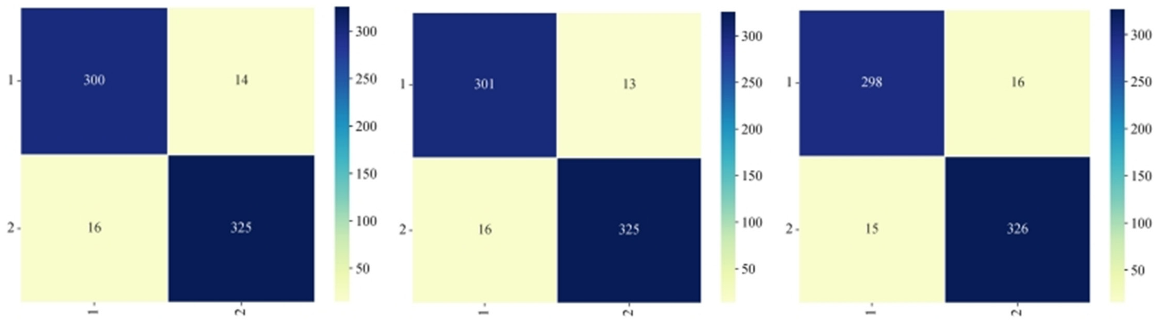

Figure 5. Performance of the KELM-LSTM-Transformer model across three repeated experiments on the test set. The bar chart displays the mean and standard deviation for key performance metrics (Accuracy, Precision, Recall, F1 Score, and AUC), demonstrating the model's consistent and stable high performance. KELM: Kernel Extreme Learning Machine, LSTM: Long Short-Term Memory.

Table 3. The experimental results

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC
Tab-Transformer	89.71	88.23	89.45	88.83	0.936
SAINT	88.92	87.65	88.17	87.91	0.928
TabNet	82.38	81.24	80.97	81.1	0.872
XGBoost-Linear	85.24	83.97	84.32	84.14	0.902
DeepFM	83.95	82.76	82.43	82.59	0.887
NODE	84.67	83.51	83.08	83.29	0.896
AutoInt	81.46	80.33	79.85	80.08	0.863
FT-Transformer	87.53	86.42	86.89	86.65	0.919
KELM-LSTM-Transformer	95.42	94.87	95.63	95.25	0.981

KELM: Kernel Extreme Learning Machine, LSTM: Long Short-Term Memory, SAINT: Self-Attention with Interaction Transformer.

The KELM-LSTM-Transformer model achieved remarkable improvements in classification performance. It reached an accuracy of 95.42%, surpassing all comparison models (the next highest being 89.71%), and recorded an AUC of 0.981, confirming the robustness of its hybrid design that integrates LSTM-based temporal feature extraction, Transformer-driven global correlation modeling, and KELM's efficient nonlinear classification.

Notably, the model attained a recall of 95.63% and an F1 score of 95.25%, exceeding competing models by over 6 percentage points, demonstrating its superior ability to identify positive samples and minimize false negatives. Although the three enhanced Transformer variants (TabTransformer, SAINT, FT-Transformer) improved feature interaction, they lacked explicit temporal modeling, resulting in weaker performance on datasets containing inter-row dependencies, with accuracy lower by 5.71-7.89 percentage points. Traditional models such as XGBoost-Line (85.24%) and TabNet (82.38%) further highlighted the innovation of the proposed hybrid architecture. The LSTM component effectively captured intra-row dynamic evolution, the transformer component modeled cross-column global interactions, and the KELM classifier addressed high-dimensional nonlinear boundaries through kernel mapping. Together, these elements formed a spatiotemporal joint learning paradigm that outperformed all single-structure and partially integrated models [36, 37].

Ablation study of model components

The ablation study results are summarized in **Table 4**. The complete KELM-LSTM-Transformer model achieved the highest performance across all metrics - accuracy (95.42%), precision (94.87%), recall (95.63%), and F1 score (95.25%) - demonstrating the synergistic benefit of integrating all three modules.

Table 4. The results of the ablation experiment

Model	Accuracy	Precision	Recall	F1	AUC
KELM-LSTM-Transformer	95.42	94.87	95.63	95.25	0.981
LSTM-Transformer	88.26	87.05	87.91	87.47	0.909
KELM-Transformer	87.39	86.18	86.85	86.51	0.901
KELM-LSTM	85.72	84.43	84.97	84.69	0.887
Transformer	83.95	82.67	83.24	82.95	0.872
LSTM	82.18	80.92	81.53	81.22	0.856
KELM	80.31	78.98	79.62	79.29	0.839

KELM: Kernel Extreme Learning Machine, LSTM: Long Short-Term Memory.

Table 5. Performance of the KELM-LSTM-Transformer model in predicting 3-year ad conversion from mild cognitive impairment

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC
KELM-LSTM-Transformer (Ours)	92.78	92.15	93.55	92.84	0.965
FT-Transformer	86.15	85.9	86.44	86.17	0.913
SAINT	84.32	83.71	85.02	84.36	0.898
XGBoost Line	82.05	81.66	82.51	82.08	0.881
TabNet	79.88	79.24	80.45	79.84	0.856

KELM: Kernel Extreme Learning Machine, LSTM: Long Short-Term Memory, SAINT: Self-Attention with Interaction Transformer.

When any component was removed, performance declined substantially. The LSTM-Transformer model (without KELM) achieved 88.26% accuracy and 87.47% F1, the KELM-Transformer model (without LSTM) reached 87.39% accuracy and 86.51% F1, and the KELM-LSTM model (without Transformer) showed the largest performance drop, with 85.72% accuracy and 84.69% F1.

Single-component models performed weakest - Transformer (83.95%), LSTM (82.18%), and KELM (80.31%) - indicating that multimodal integration was essential for optimal performance.

The AUC values for all ablation variants ranged from 0.839 to 0.909, consistently lower than that of the complete model (0.981), reinforcing the importance of feature fusion in enhancing robustness and discriminative ability.

Overall, the ablation results emphasize that the KELM-LSTM-Transformer architecture effectively captures complex temporal-contextual relationships through complementary feature learning, offering significantly superior performance to any partial or single-model configuration.

Early prediction of ad conversion

To validate the model's capability for early prediction, one of the key objectives of this study, we conducted a longitudinal analysis in a subcohort of patients with MCI. From the initial dataset, 450 participants diagnosed with MCI at baseline and followed clinically for at least 3 years were identified. These patients were divided into two groups according to their diagnosis at the 3-year follow-up: MCI converters (MCI-C), who progressed to AD (n = 210); and MCI stable (MCI-S), who remained at the MCI stage (n = 240).

The predictive task was to determine whether the model could accurately forecast these 3-year outcomes using only baseline (Year 0) data. The proposed KELM-LSTM-Transformer model was benchmarked against other high-performing models for this task, with results summarized in **Table 5**.

Our model demonstrated outstanding predictive performance, achieving 92.78% accuracy, 92.15% precision, 93.55% recall, 92.84% F1 score, and an AUC of 0.965. Notably, the model outperformed all comparison algorithms, exceeding the next-best model (FT-Transformer)

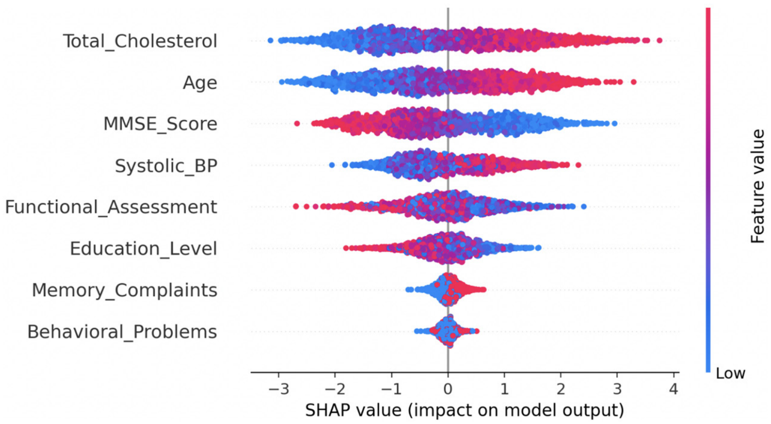

Figure 6. SHAP summary plot illustrating feature importance and impact on the model's prediction. Each point on the plot represents a Shapley value for a feature and a patient. Features are ranked by importance on the y-axis. The x-axis represents the SHAP value, where positive values push the prediction towards AD. The color indicates the feature's value, with red representing higher values and blue representing lower values, revealing how specific feature values influence the prediction outcome.

by more than 6 percentage points in accuracy. The high recall rate (93.55%) is particularly significant clinically, reflecting the model's strong ability to identify individuals at high risk of progression to AD-crucial for timely intervention. These findings provide robust evidence for the model's effectiveness in early AD prediction, demonstrating its potential to forecast disease onset approximately three years in advance using only baseline clinical data.

Model interpretability and feature importance

To address the "black-box" nature of deep learning models and enhance clinical transparency, we employed the SHAPLEY ADDITIVE EXPLANATIONS (SHAP) framework, a state-of-the-art explainable AI (XAI) approach. SHAP quantifies the contribution of each clinical feature to individual predictions, thereby elucidating the model's internal decision-making process [38].

The SHAP summary plot for our model (**Figure 6**) ranks features by their mean absolute SHAPLEY value, representing overall importance. Each point corresponds to a single patient-feature pair, where color denotes feature magnitude (red = high, blue = low), and the x-axis position indicates the feature's impact on model output. Positive SHAP values drive the prediction toward AD, whereas negative values move it away.

The results clearly show that the Mini-Mental State Examination (MMSE) score is the most influential feature. Lower MMSE scores (blue points) are predominantly associated with high positive SHAP values, meaning that cognitive decline strongly pushes predictions toward AD. Conversely, higher MMSE scores (red points) have negative SHAP values, reducing predicted risk - an outcome perfectly consistent with clinical understanding.

Age ranked as the second most important predictor, with higher ages (red points) corresponding to positive SHAP values, reaffirming age as a well-established AD risk factor. Additionally, Functional Assessment scores and the presence of memory complaints also showed substantial influence. Lower functional scores (blue) and reported memory complaints (feature value = 1, red) both increased AD risk predictions, in line with early symptomatic progression.

In summary, the SHAP analysis reveals that the model's predictions are grounded in clinically interpretable features rather than arbitrary correlations. This interpretability enhances clinical trust and applicability, allowing physicians to understand the rationale behind AI-driven predictions - an essential step toward integrating XAI into real-world clinical workflows.

External validation and generalizability

To rigorously assess model generalizability, we conducted external validation using an independent dataset not involved in model training or internal validation [39].

This external cohort, derived from an open-access clinical database, included 1,012 patients with complete demographic, lifestyle, clinical, and cognitive data. The distribution between AD and non-AD cases was balanced (AD: 48.3%, non-AD: 51.7%), ensuring comparability with the internal dataset. Variables used for validation mirrored those in the train-

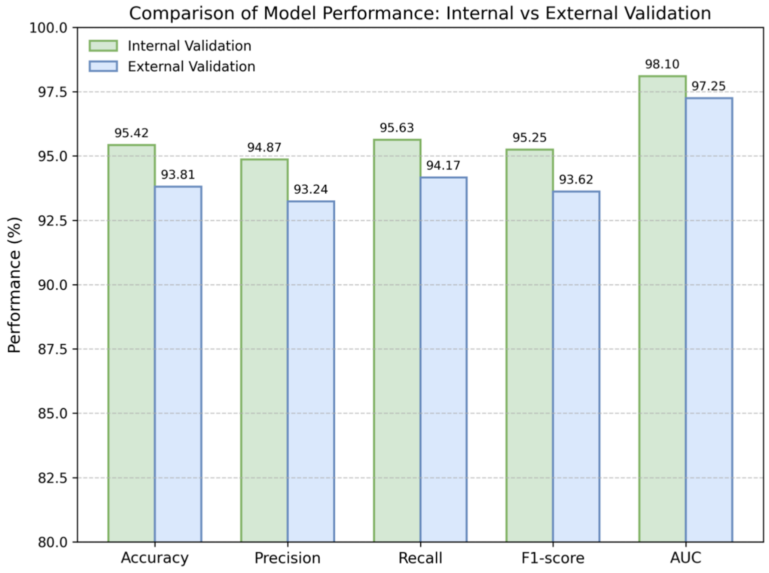

Figure 7. Comparison of model performance between internal and external validation. This chart compares the key performance metrics of the model on the internal validation set (n = 2,149) and the independent external validation set (n = 1,012). The consistent high performance across both datasets highlights the model's strong generalizability and robustness.

ing cohort, including age, gender, education, BMI, comorbidities (hypertension, diabetes, cardiovascular disease), lifestyle factors (smoking, alcohol consumption, physical activity), and clinical measures (blood pressure, cholesterol levels). Cognitive assessments included MMSE, activities of daily living, memory complaints, and behavioral changes. All categorical variables were binarized, and continuous variables were standardized (zero mean, unit variance) using statistics from the external dataset.

The trained KELM-LSTM-Transformer model was directly applied to this dataset without retraining or fine-tuning, ensuring a strict out-of-sample evaluation. Performance metrics-accuracy, precision, recall (sensitivity), F1 score, and AUC-were computed, with bootstrapped resampling (1,000 iterations per run) used to estimate 95% confidence intervals (CIs).

On the external dataset, the model achieved: accuracy: 93.81% (95% CI: 93.1-94.4%); precision: 93.24% (95% CI: 92.5-93.9%); recall: 94.17% (95% CI: 93.6-94.8%); F1 score: 93.62% (95% CI: 93.0-94.3%); AUC: 97.25% (95% CI: 96.7-97.8%).

Although slightly lower than the internal validation performance (accuracy 95.42%, precision 94.87%, recall 95.63%, F1 95.25%, AUC 98.10%), these results remain substantially superior to all baseline methods, confirming the model's robustness and stability across heterogeneous populations (**Figure 7**).

The consistent high recall underscores the model's clinical utility, reflecting excellent sensitivity for detecting early AD cases. These findings confirm that the model is not overfitted to a specific cohort and can generalize effectively to new patient populations. Overall, the external validation demonstrates strong potential for real-world deployment, while future studies should further evaluate its performance in multi-center settings incorporating neuroimaging and biomarker data [40].

Discussion

This study presents a novel deep fusion framework-KELM-LSTM-Transformer-for the diagnosis and prediction of AD. By combining three distinct architectures, the model achieves enhanced accuracy, robustness, and interpretability. The integration of the KELM, LSTM, and Transformer modules enables comprehensive modeling of structured clinical data, simultaneously capturing nonlinear relationships, temporal evolution, and global feature interactions.

The model achieved 95.42% accuracy, 95.63% recall, and an AUC of 0.981 in internal validation, demonstrating superior capability in addressing the complexity of AD prediction, especially in cases characterized by subtle symptoms and high-dimensional data.

A key strength of this work lies in the longitudinal validation of its predictive capacity. Using baseline data alone, the model accurately forecasted 3-year conversion from MCI to AD, underscoring its potential for early disease detection. This functionality moves beyond conventional diagnosis to serve as a prognostic tool capable of identifying high-risk individuals before irreversible neurodegeneration occurs. Such early prediction represents an important advance toward preventive intervention and personalized disease management, the central aims of modern AD research.

From a methodological perspective, this framework exemplifies multimodal feature fusion. The LSTM module excels in sequential pattern recognition, the transformer captures long-range contextual dependencies, and the KELM component enables rapid nonlinear classification with minimal training overhead. This synergy effectively bridges local dynamic dependencies with global structural relationships, producing a scalable solution for medical time-series analysis.

External validation further confirmed the model's robustness. On an independent dataset of 1,012 patients, it achieved 93.81% accuracy, 93.24% precision, 94.17% recall, 93.62% F1, and 97.25% AUC. Although slightly lower than internal results, these findings verify that the model generalizes well to heterogeneous populations without overfitting. The consistently high recall is particularly valuable in clinical applications, minimizing the risk of missed early-stage AD diagnoses (**Figure 6**).

Despite promising results, several limitations must be noted. First, the study utilized only structured tabular data-demographic, clinical, and cognitive assessments-excluding neuroimaging (e.g., MRI, PET) and fluid biomarkers (e.g., CSF Aβ42/tau). While such biomarkers enhance diagnostic precision, this work intentionally focused on routinely collected, low-cost, and non-invasive data to evaluate accessibility and real-world applicability. Consequently, the model is not intended to replace imaging or CSF analyses but to function as a first-line screening and risk-stratification tool in primary or resource-limited settings. Patients identified as high-risk could then be referred for confirmatory biomarker testing. Second, although external validation was conducted, additional multicenter studies are needed to confirm performance across diverse demographic and clinical populations.

Interpretability remains another challenge. Although the Transformer provides partial transparency through attention weights, it lacks full clinical explainability. Future work should integrate XAI techniques, such as SHAP or local interpretable model-agnostic explanations, to visualize feature importance and strengthen clinician trust. Additionally, further research should address class imbalance, subgroup bias (e.g., ethnicity or gender), and incomplete longitudinal records to ensure model fairness and reliability. Finally, while designed for AD, the proposed architecture is modular and generalizable. Its flexible structure can be readily extended to other chronic or progressive diseases with temporal and multimodal characteristics-such as Parkinson's disease, heart failure, or cancer prognosis-thereby broadening its potential applications in predictive healthcare.

In conclusion, this work proposes a hybrid deep learning framework-KELM-LSTM-Transformer-for early prediction and diagnosis of Alzheimer's disease using structured clinical data. By integrating the sequential modeling power of LSTM, the global attention capability of the Transformer, and the nonlinear classification efficiency of KELM, the model achieved state-of-the-art performance, outperforming both traditional machine-learning and recent Transformer-based models.

Ablation experiments confirmed the indispensable contribution of each component, with the transformer showing the greatest influence on global feature integration. External validation further demonstrated strong generalizability and clinical potential.

Future research should aim to enhance generalization through multi-center validations, expand the feature set to include imaging, biomarkers, and genetic data, and strengthen transparency using XAI frameworks. Addressing issues of fairness and subgroup bias will also be critical for clinical adoption. The model's modular design provides a promising foundation for broader implementation across time-sensitive and multimodal disease prediction tasks, advancing the field toward truly precision-driven medicine.

Acknowledgements

The authors would like to thank all colleagues and institutions that contributed to this work.

Disclosure of conflict of interest

None.

Address correspondence to: Ling Zhu, Department of Basic Medicine, Qujing University of Medicine & Health Sciences, No. 226 Sanjiang Avenue, Economic and Technological Development Zone, Qujing 655011, Yunnan, China. E-mail: aazz@qjyxgdzkxx12.wecom.work

References

[1] Alzheimer's Association. 2022 Alzheimer's disease facts and figures. Alzheimers Dement 2022; 18: 700-789.

[2] Nichols E, Steinmetz JD and Vollset SE; GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019. Lancet Public Health 2022; 7: e105-e125.

[3] Hardy J and Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science 2002; 297: 353-356.

[4] Wang J, Gu BJ, Masters CL and Wang YJ. A systemic view of Alzheimer disease-insights from amyloid-β metabolism beyond the brain. Nat Rev Neurol 2017; 13: 612-623.

[5] Jack CR Jr, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, Holtzman DM, Jagust W, Jessen F, Karlawish J, Liu E, Molinuevo JL, Montine T, Phelps C, Rankin KP, Rowe CC, Scheltens P, Siemers E, Snyder HM and Sperling R; Contributors. NIA-AA research framework: toward a biological definition of Alzheimer's disease. Alzheimers Dement 2018; 14: 535-562.

[6] Cummings J, Lee G, Zhong K, Fonseca J and Taghva K. Alzheimer's disease drug development pipeline: 2021. Alzheimers Dement (N Y) 2021; 7: e12179.

[7] Hampel H, O'Bryant SE, Molinuevo JL, Zetterberg H, Masters CL, Lista S, Kiddle SJ, Batrla R and Blennow K. Blood-based biomarkers for Alzheimer disease: mapping the road to the clinic. Nat Rev Neurol 2018; 14: 639-652.

[8] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B and Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal 2017; 42: 60-88.

[9] LeCun Y, Bengio Y and Hinton G. Deep learning. Nature 2015; 521: 436-444.

[10] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S and Dean J. A guide to deep learning in healthcare. Nat Med 2019; 25: 24-29.

[11] Lin W, Tong T, Gao Q, Guo D, Du X, Yang Y, Guo G, Xiao M, Du M and Qu X; Alzheimer's Disease Neuroimaging Initiative. Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. Front Neurosci 2018; 12: 777.

[12] Zhang L, Wang M, Liu M and Zhang D. Deep learning for diagnosis of Alzheimer's disease: a survey. Front Aging Neurosci 2020; 13: 620037.

[13] Basaia S, Agosta F, Wagner L, Canu E, Magnani G, Santangelo R and Filippi M; Alzheimer's Disease Neuroimaging Initiative. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. Neuroimage Clin 2019; 21: 101645.

[14] Lu D, Popuri K, Ding GW, Balachandar R and Beg MF; Alzheimer's Disease Neuroimaging Initiative. Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. Sci Rep 2018; 8: 5697.

[15] Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Dormont D, Durrleman S, Colliot O and Burgos N; Alzheimer's Disease Neuroimaging Initiative; Australian Imaging Biomarkers and Lifestyle flagship study of ageing. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med Image Anal 2020; 63: 101694.

[16] Chen T and Guestrin C. XGBoost: a scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min 2016; 785-794.

[17] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N and Lee SI. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020; 2: 56-67.

[18] Sarica A, Cerasa A and Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Front Aging Neurosci 2017; 9: 329.

[19] Ortiz A, Munilla J, Gorriz JM and Ramirez J. Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease. Int J Neural Syst 2016; 26: 1650025.

[20] Suk HI, Lee SW and Shen D; Alzheimer's Disease Neuroimaging Initiative. Deep ensemble learning of sparse regression models for brain disease diagnosis. Med Image Anal 2017; 37: 101-113.

[21] Islam J and Zhang Y. Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. Brain Inform 2018; 5: 2.

[22] Breiman L. Random forests. Mach Learn 2001; 45: 5-32.

[23] Kipf TN and Welling M. Semi-supervised classification with graph convolutional networks. ArXiv 2016; 1609.02907.

[24] Parisot S, Ktena SI, Ferrante E, Lee M, Guerrero R, Glocker B and Rueckert D. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. Med Image Anal 2018; 48: 117-130.

[25] Hochreiter S and Schmidhuber J. Long short-term memory. Neural Comput 1997; 9: 1735-1780.

[26] Zhang J, Liu M and Shen D. Integrating temporal information into medical image classification: recurrent neural networks for Alzheimer's disease diagnosis. Med Image Anal; 53: 111-122.

[27] Ma T and Zhang A. Integrating multi-omics data for disease prediction: a review of machine learning methods. Comput Struct Biotechnol J; 17: 1047-1058.

[28] Huang GB, Zhou H, Ding X and Zhang R. Extreme learning machine for regression and multiclass classification. IEEE Trans Syst Man Cybern B Cybern 2012; 42: 513-529.

[29] Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015; 61: 85-117.

[30] Yu K, Zhang D, Pan J, Li Y, Liu Y, Shen J and Han J. LSTM-based EEG classification in motor imagery tasks. IEEE Trans Neural Syst Rehabil Eng 2018; 26: 2086-2095.

[31] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L and Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst 2017; 30: 5998-6008.

[32] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. An image is worth 16x16 words: transformers for image recognition at scale. ArXiv 2010; 11929.

[33] Khan S, Naseer M, Hayat M, Zamir SW, Khan FS and Shah M. Transformers in vision: a survey. ACM Comput Surv 2022; 54: 1-41.

[34] Li Y, Zhang L and Liu M. A review of deep learning methods for brain disease diagnosis using neuroimaging modalities. IEEE Trans Neural Netw Learn Syst; 32: 1237-1257.

[35] Huang X, Khetan A, Cvitkovic M and Karnin Z. TabTransformer: tabular data modeling using contextual embeddings. ArXiv; 2012.06678.

[36] Arik SÖ and Pfister T. TabNet: attentive interpretable tabular learning. Proc AAAI Conf Artif Intell 2021; 35: 6679-6687.

[37] Popov S, Morozov S and Babenko A. Neural oblivious decision ensembles for deep learning on tabular data. ArXiv 2019; 1909.06312.

[38] Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, Fieguth P, Cao X, Khosravi A, Acharya UR, Makarenkov V and Nahavandi S. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf Fusion 2021; 76: 243-297.

[39] Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, Jack CR Jr, Jagust WJ, Shaw LM, Toga AW, Trojanowski JQ and Weiner MW. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology 2010; 74: 201-209.

[40] Whelan R; Alzheimer's Disease Neuroimaging Initiative. Effective connectivity predicts clinical outcome in mild cognitive impairment. Neuroimage 2021; 221: 117216.