A Telemedicine Analytic Framework for Fully and Semi-Automatic Alzheimer's Disease Screening Using Clock Drawing Test

Wei Bo, Suzanne S. Sullivan, Xiaoyu Zhang, Mingchen Gao, and Wenyao Xu, *Senior Member, IEEE*

***Abstract*—More than 6 million Americans are at risk for Alzheimer's Disease Related Dementias (ADRD), most of whom are 65 or older. The clock drawing test (CDT) is a quick, simple, and effective technique that has the potential advantage of self-management and screening for ADRD patients. Current CDT-based ADRD screening studies focus more on efficacy, involving many handcrafted features, ignoring data modalities, and lacking validation. This paper aims to propose a unified telemedicine framework for fully and semi-automatic effective early ADRD screening based on multimodal and agile data fusion, focusing on the interpretability and validation of the model by using gradient-weighted class activation mapping (Grad-CAM) and locally linear embedding (LLE). The datasets for this work include 1,662 samples of CDT images and related demographic and cognitive information. The fully automatic case involving only CDT images can achieve the highest AUC of 81% with a 75% recall rate in binary screening. The multimodal data fusion in the semi-automatic case can achieve up to 90% AUC with an 83% recall rate. The visualization of the Convolutional Neural Networks (CNNs) shows that it can automatically obtain critical information about the outline, scale, and clock hands from CDT images, and the analysis of structured features shows that the memory test is key to effective ADRD screening.**

***Index Terms*—Clock drawing test, gradient-weighted class activation mapping, locally linear embedding, multimodal data fusion.**

I. Introduction

AGE-RELATED cognitive decline and neurodegenerative disorders, such as Alzheimer's disease and related dementias (ADRD), pose significant challenges to the health and well-being of older adults. As individuals age, the risk of developing ADRD increases substantially, with the majority of cases occurring in people older than 65. There are more than 55 million people living with dementia worldwide, and nearly 10 million new cases appear each year [1]. It has a profound impact on individuals, their families, and society as a whole [1], [2].

The rising prevalence of ADRD among the aging population has prompted significant research efforts to better understand the underlying causes and explore early interventions for prevention [3]. The Clock Drawing Test (CDT) then becomes one of the most attractive measures of cognitive function. It has the advantages of being low cost, easy to understand, requiring no special medical hardware, and the capacity for remote operation, which is very friendly to people in rural areas and middle-income countries. In addition, it can minimize the intervention of professionals through automated machine learning methods, which has the advantages of real-time and resource-saving. Compared with other screening tools, such as the Mini-Mental State Examination (MMSE), it not only compensates for the time-consuming and complex shortcomings, but it is also relatively unaffected by language, cultural, and ethnic influences [4].

However, it is crucial to acknowledge the existing deficiencies in current CDT-based studies. The first notable concern is the predominant focus on effectiveness and accuracy [5], [6], [7], leading to the inclusion of handcrafted features and extensive manual intervention. These practices carry the risk of introducing excessive external bias and compromising the fairness of the screening process. Secondly, certain CDT studies exhibit a limited focus solely on the CDT tool itself, lacking the necessary flexibility to incorporate diverse inputs [8], [9]. These studies fail to consider the practical application scenarios where additional data, such as demographic information, play a significant role. Thirdly, studies that incorporate heterogeneous inputs tend to concentrate primarily on the relationship between the CDT tools and ADRD. Unfortunately, they often overlook the potential of multimodal and agile data fusion techniques [10], halting their analysis at a statistical level without further verification and validation steps [4], [11].

Our work proposes a unified telemedicine framework for ADRD screening. The framework encompasses fully automatic

Received 28 September 2023; revised 21 January 2024 and 13 May 2024; accepted 18 June 2024. Date of publication 25 June 2024; date of current version 6 December 2024. This work was supported by the NIH under Grant R03AG067159. *(Corresponding author: Wenyao Xu.)*

Wei Bo, Xiaoyu Zhang, Mingchen Gao, and Wenyao Xu are with the Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY 14260 USA, and also with the State University of New York, Buffalo, NY 14260 USA (e-mail: weibo@buffalo.edu; zhang376@buffalo.edu; mgao8@buffalo.edu; wenyaoxu@buffalo.edu).

Suzanne S. Sullivan is with the School of Nursing, University at Buffalo, Buffalo, NY 14214 USA, also with the State University of New York, Buffalo, NY 14214 USA, also with the College of Nursing, Upstate Medical University, Syracuse, NY 13210 USA, and also with the State University of New York, Syracuse, NY 13210 USA (e-mail: suzanney@buffalo.edu).

Digital Object Identifier 10.1109/JBHI.2024.3419059

scoring and screening based on CDT images, as well as semi-automatic cases involving heterogeneous inputs and multimodal data fusion. In the fully automated scenario, participants can operate autonomously without human intervention or handcrafted features. The semi-automatic case allows for customization by including additional information, such as demographic and cognitive data. This flexibility facilitates the integration of handcrafted features into the system. The framework is adaptable for quantitative scoring or qualitative binary diagnosis of ADRD, making it highly applicable. We prioritize the interpretability and validity verification of the model to enhance comprehensibility and credibility, addressing key concerns in mHealth application design and user satisfaction. The current scope of our study focuses on establishing benchmarks, especially utilizing CDT images for automatic scoring and screening for ADRD with attention to interpretability and validation. This paper, as a prospective study, will be of great help to the actual implementation of the application. The contributions of our work are three-fold:

- A unified telemedicine framework for the screening of ADRD, which is powered by machine learning algorithms, enabling automatic scoring and screening based on CDT images, thereby eliminating the need for manual intervention.
- The framework can also apply for heterogeneous inputs and enable focus on multimodal and agile data fusion in semi-automatic cases.
- Exploring the effectiveness and reliability of advanced visualization method and manifold learning.

The remainder of this paper is organized as follows. Section II gives a brief background of CDT and an overview of existing work on CDT-based ADRD studies. Section III introduces the basic principles of the proposed framework and describes the datasets and procedures used in this work, and Section IV provides comprehensive and concrete analysis and results. Section V offers a final discussion and insights. We conclude our work in Section VI.

II. Related Works

CDT has emerged as a prominent focus in research on early screening of ADRD, owing to its effectiveness. Various methods have been explored to validate its efficacy, including statistical analysis and rating scales [12], [13], as well as machine learning techniques for processing CDT images [8], [14]. However, these approaches often employ intricate scoring systems, rely on numerous handcrafted features, and involve extensive manual interventions, leading to potential bias during model training and posing challenges for flexible applications.

Some studies have combined CDT images with other information for further analysis. Seigerschmidt et al. [11] claimed that age, gender, and level of education need to be considered with CDT scores together for dementia screening. Additional studies have also demonstrated that motor ability decreases significantly with age, which affects CDT scores [5], [15]. In a highly educated population, clock drawing is influenced by educational level [4]. These studies provide insights for combining multiple sources of information, indicating it would be more conducive to ADRD screening. However, they usually focus too much on the validation of the results or prediction performance, ignoring the interpretation of these multi-modal data fusion processes and the procedural analysis of this information.

Relevant studies have shown that machine learning can be effectively and widely used in the auxiliary diagnosis and screening of ADRD [8], [10], [16]. However, few CDT-based studies have comprehensive considerations, such as not only considering the effectiveness of ADRD screening but also focusing on the combination of CDT images and other valuable information, as well as explaining and validating the model to help users better understand.

Our research aims to remedy the mentioned deficiencies with a unified framework that maximizes the applicability of the proposed framework by discovering key features during model analysis for effective ADRD screening without any manual intervention.

III. Proposed Framework and Methods

In this study, we proposed a unified and comprehensive telemedicine framework for early screening of ADRD. The framework encompasses two main scenarios: fully automatic ADRD screening using only CDT images and semi-automatic screening involving multimodal data fusion. The framework consists of six modules, each serving a specific purpose as depicted in Fig. 1.

The initial approach presented in this study offers a standardized procedure for both quantitative score measurement and qualitative binary screening. It emphasizes the effectiveness and application of CDT itself, striving to minimize human intervention and achieve rapid screening through artificial intelligence. Advanced visualization [17] and manifold learning algorithms [18] are employed to enhance interpretability and explore the model's validity and reliability. The second semi-automatic approach focuses on the multi-modal data fusion procedure [19], [20], encompassing all three fusion stages, to take advantage of additional information, such as the word recall test and demographic data. Furthermore, it explores the use of predicted quantitative scores as inputs for the final qualitative binary diagnosis, aiming to achieve a more automated prediction process.

The framework aligns rigorously with the established principles of mobile application development [21], facilitating users in executing CDT tests with pen-and-paper convenience. Subsequent to this initial step, users can seamlessly capture CDT images using their mobile devices and subsequently transmit the acquired data for processing. The results of prediction or analysis are efficiently relayed back to the mobile device. Notably, the application operates predominantly in a local capacity, requiring the network connection solely during data transmission. This operational design significantly mitigates concerns associated with model size, speed, file management, and thread performance [22]. Furthermore, the framework operates without real-time constraints, with data transfer efficiency not being

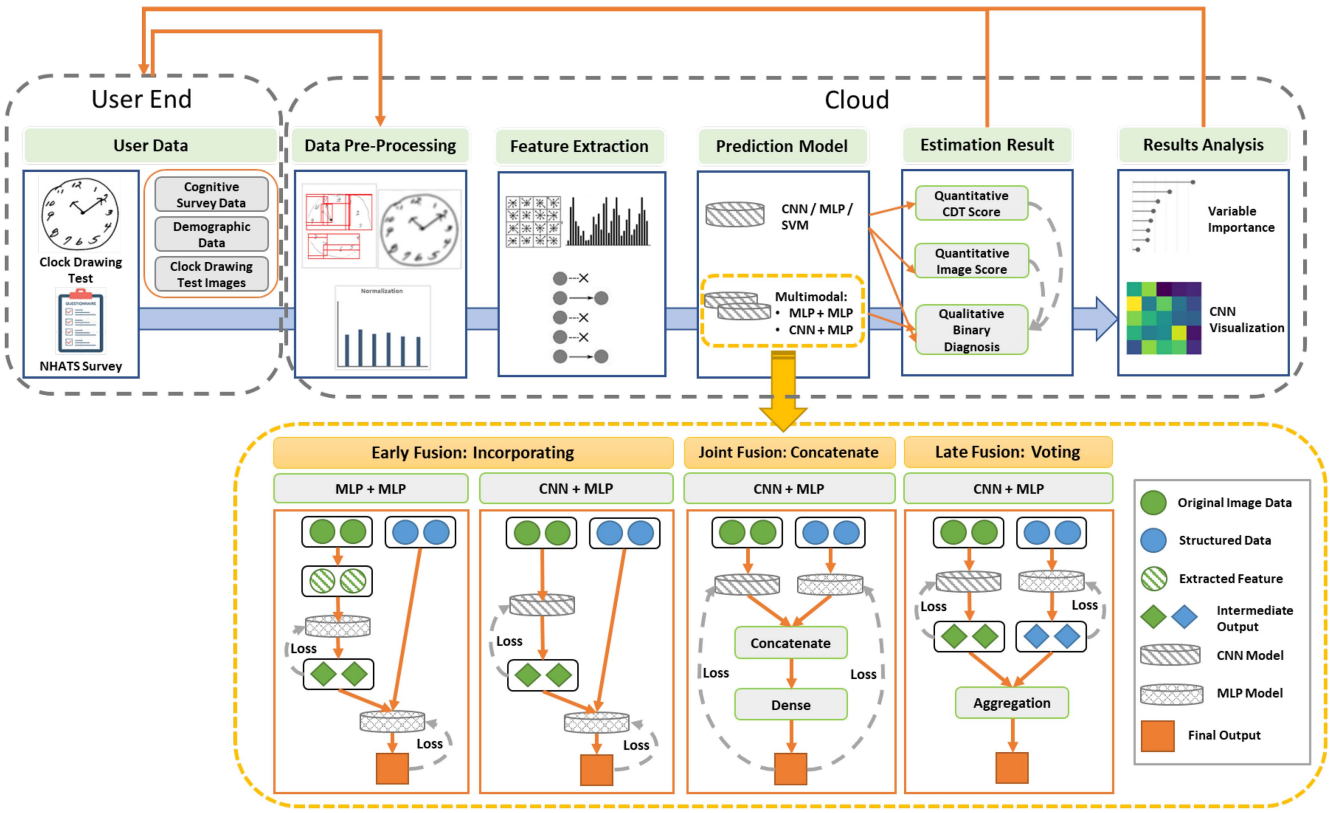

Fig. 1. A unified telemedicine framework for early screening of ADRD. This framework consists of six modules: user data collection, data pre-processing, feature extraction, prediction model, estimation result, and result analysis. The prediction model encompasses two main scenarios: fully automatic ADRD screening using only CDT images with CNN, MLP, and SVM models, and semi-automatic screening involving multimodal data fusion with three fusion stages of early, joint, and late fusion.

the primary focus. The fully automatic mode yields an average raw data size of 373.81 KB per event in telemedicine screening, whereas the semi-automatic mode records an average size of 373.90 KB per event, encompassing both image and structured data. Presently, the prevailing Wi-Fi standard in most households, 802.11ac, supports data rates ranging from several hundred Mbps to Gbps [23]. Comparatively, 4G LTE in cellular networks typically delivers speeds between several tens to hundreds of Mbps, while 5G networks can potentially attain multi-Gbps speeds contingent upon implementation and frequency band utilization [24]. For instance, at a typical data rate of 100 Mbps, data transmission takes approximately 0.0299 seconds per event in both fully automatic and semi-automatic modes. The framework's versatility caters comprehensively to diverse user categories, offering valuable insights to both general and clinical users in the telemedicine domain.

A. Dataset

The data used in this work includes CDT images and structured data, which consists of demographic and cognitive information. All these datasets come from the National Health and Aging Trends Study (NHATS) [25], which has been collecting information on a national sample of Medicare beneficiaries aged 65 and older annually since 2011. NHATS was chosen for this study due to its nationally representative data with great richness in data content and high diversity, inclusivity and data quality. Its enrollment plan ensures the findings are generalizable to the entire elderly population covered by Medicare, and its sampling method is stratified to ensure adequate representation across various demographic segments, including age, sex, race, and geographic location [26]. The study uses rigorous data collection methods, including in-person interviews and performance tests, to ensure the accuracy and reliability of the data. NHATS is sponsored by the National Institute on Aging (grant number NIA U01AG32947) and is conducted by Johns Hopkins University. It oversamples African Americans and older adults by design [27].

The feature space of this study, shown in Table I, is constructed based on NHATS survey data and corresponding CDT images. The predictive features include demographic and cognitive information from the survey.

Demographic information includes age, weight, height and gender of the subject and cognitive information contains scores of memory test and CDT. The variable 'intvrage' indicates the interval of the subject's age level. This is a categorical variable whose value is from 1 to 6 to indicate the subject's age level varies from 65 to 90+, and every 5 years indicates a class. For example, value 1 means this subject's age is in the interval of 65 to 69. We didn't include sensitive data in this study, such as the exact age of the subjects, due to the consideration of

TABLE I
FEATURE SPACE

Feature Type	Feature Name	Description	Values
Demographic Information	weight	Subject's weight	Numerical data
	height	Subject's height	Numerical data
	gender	Subject's gender	2 classes: Female, Male
	race	Subject's race	3 classes: White, African American, Other
	intvrage	Subject's age	6 classes: From 65 to 90+, every 5 years per class
Cognitive Information	wrdimmrc	Score of Immediate Word Recall Test	From 0 to 10
	wrddlyrc	Score of Delayed Word Recall Test	From 0 to 10 (Actually from 0 to 8)
Cognitive Information & Target Variable for Quantitative Prediction	clkdraw	Score of Clock Drawing Test	From score 0 to score 5: Not recognizable, Severely distorted, Moderately distorted, Mildly distorted, Accurate, and Reasonably accurate
	clkimgcl	Image Clarity	From score 1 to score 4: Very clear, Somewhat clear, Somewhat unclear, Very unclear
Target Variable for Dementia Screening	dementia	Subjects has dementia or not	2 classes: 0 for no, 1 for yes

TABLE II
STATISTICS OF (PART OF) DEMOGRAPHIC INFORMATION

Feature Name	Values	Non-ADRD	ADRD
intvrage	65-69	102	7
	70-74	341	23
	75-79	347	58
	80-84	294	71
	85-89	181	66
	90+	120	52
gender	Male	619	108
	Female	766	169
race	White	1011	174
	African American	256	70
	Other	118	33

TABLE III
STATISTICS OF CDT SCORES

Feature Name	Values	Non-ADRD	ADRD
clkdraw	0	7 (0.51%)	11 (3.97%)
	1	35 (2.53%)	34 (12.27%)
	2	148 (10.69%)	53 (19.13%)
	3	248 (17.91%)	85 (30.69%)
	4	519 (37.47%)	66 (23.83%)
	5	428 (30.90%)	28 (10.11%)
	Mean	3.8202	2.8845
clkimgcl	1	1323 (95.52%)	245 (88.45%)
	2	43 (3.10%)	20 (7.22%)
	3	12 (0.87%)	9 (3.25%)
	4	7 (0.51%)	3 (1.08%)
	Mean	1.0635	1.1697

mitigating the risk of information leakage and de-identification of the subjects.

Cognitive information includes memory tests and CDT. The memory tests are immediate and delayed 10 item word recall tests. A list of 10 nouns is read to subjects and they were asked to recall as many words as possible, in any order [26]. The immediate word recall test focuses on short-term or working memory, the subjects were asked to recall the words immediately after the presentation of the 10 words [28]. However, the delayed word recall test evaluates long-term memory or the ability to store and retrieve information over longer periods [28]. The subjects were asked to recall the words after the clock drawing test or other cognitive tests [26]. There are two measurements corresponding to CDT scores, one is the score of CDT (clkdraw), the other one is image clarity (clkimgcl). The score of CDT ranges from 0 to 5, 0 means the image cannot be recognized as a clock, 5 means the image is an accurate depiction of a clock. The score of image clarity ranges from 1 to 4, 1 means the image is very clear, 4 means the image is very unclear. These scores are assessed by trained lay coders and clinical coders. The specific scoring guideline, coder training, and selection process can be found in the NHATS user manual [26].

The qualitative target variable for binary ADRD screening comes from the NHATS health condition section, participants will be asked whether they have been informed by a doctor that they have ADRD since the last interview [26]. If they answered yes, it means that they have been clearly diagnosed with dementia or AD in the past year. The participants who answered yes would be regarded as positive samples in this work, and those who answered no would be negative samples. However, it is worth pointing out that the participants who answered no do not absolutely mean that they are not at risk for dementia or ADRD. It may be that they have not had this examination in the past year, or they may have possible dementia, but it was not detected.

NHATS has conducted 12 rounds of surveys, with 10 rounds completed at the time of this study. From the 6–10 rounds of data, a set of 1662 samples were randomly selected as ground truth, comprising 277 positive samples and 1385 negative samples. As Tables II and III showed, among all the samples, there were 727 men and 935 women, most of them were white. Because the subjects of NHATS are 65 age or older, there are 89.17% positive samples over 75 years old, and 83.97% of negative samples are between 70–89 years old. In addition, approximately 97% of negative samples get CDT scores greater than or equal to 2, and

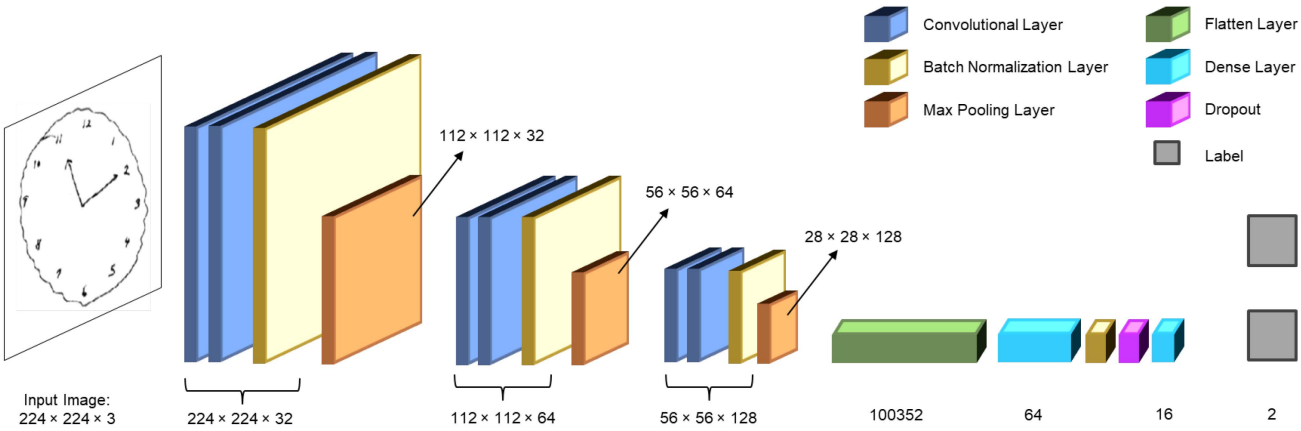

Fig. 2. A detailed architecture of customized CNN for qualitative binary ADRD screening. It processes 224 × 224 pixel CDT images and features three convolutional layers with increasing filters, batch normalization, max pooling, dense layers reducing from 64 to 16, and a dropout layer to prevent overfitting. This architecture ensures efficient feature extraction and robust image classification.

86.28% of them are greater than 2. For positive samples, 83.76% of them got CDT scores greater than or equal to 2. The mean score of positive samples is nearly 1 point larger than that of negative samples. For all samples, about 94.34% of them have an image clarity score of 1, which means that the image is clearly recognizable.

B. Experimental Setup

For the fully automatic scoring and screening scenario, three single models (CNN, Multilayer Perceptron (MLP), and support vector machine(SVM)) were designed to facilitate both qualitative and quantitative screening using only CDT images. In contrast, the semi-automatic scenario focuses on qualitative binary screening and incorporates heterogeneous inputs. Further details on the model design can be found in Section III-D.

Fig. 2 illustrates the detailed architecture of our custom CNN model, used for qualitative ADRD binary screening, forming the foundation for the semi-automatic case and subsequent processes of interpretability and validity verification.

This model is specifically tailored to process 224 × 224 pixel color images, featuring an architecture that incorporates multiple layers designed to efficiently capture and analyze features for classification. It includes three convolutional layers, each equipped with an increasing number of 3 × 3 filters (32, 64, and 128), essential for extracting a hierarchical spectrum of features from simple edges to complex patterns [29]. Following each convolutional layer, batch normalization is applied to enhance training stability and efficiency [30]. Subsequently, each convolutional layer is followed by a max pooling layer that reduces the spatial dimensions of the feature maps by half, sequentially decreasing from 224 × 224 to 112 × 112, 56 × 56, and finally 28 × 28. This reduction not only decreases the number of parameters and computational complexity but also increases the robustness of the feature representations against variations in feature positioning within the input images [29]. After reducing and extracting features, the model transforms the 3D feature maps into a 1D vector. This vector then feeds into a sequence of dense layers, which gradually decrease in size from 64 to 16, and finally to a size that corresponds to the number of classes (e.g., 2 for binary classification). These layers are crucial for integrating the learned features to formulate a basis for the final classification decision [29]. To prevent overfitting, a dropout layer is included between the dense layers, which randomly sets a fraction of the input units to zero during training [31]. This approach encourages the model to learn more robust and generalizable features.

In this study, SVM and MLP models were selected for their proven effectiveness in binary classification, image processing, and multimodal data fusion. SVM excels in high-dimensional spaces by efficiently creating optimal hyperplanes, which is crucial for achieving clear class separation, especially in precision-sensitive clinical settings [32], [33], [34]. Meanwhile, MLPs are key in deep learning, adept at extracting complex patterns through their hierarchical processing layers and learning nonlinear feature interactions, making them ideal for intricate image data analysis [35], [36], [37]. The effectiveness of both SVM and MLP models was rigorously assessed using several key performance metrics in our study. The results confirmed their suitability for the task, with both models demonstrating satisfactory outcomes, thereby validating their selection for this research.

The decision to employ classification methods over regression for evaluating CDT scores, which range quantitatively, is based on several key considerations that enhance clinical relevance and practical utility:

- *Clinical Relevance:* The CDT scores, though numerically continuous, often represent distinct categories of cognitive impairment levels — ranging from 'not recognizable' to 'reasonably accurate'. Each score corresponds to a clinically distinct stage of cognitive function, which aligns more closely with categorical outcomes rather than a continuous spectrum. Thus, classification allows us to focus on the differential diagnosis relevant to clinical settings.

- *Model Performance and Interpretability:* In our study, classification models have demonstrated robust performance in distinguishing between these set categories effectively. Using classification accuracy and other discrete metrics like precision and recall, we can provide clear and interpretable results for clinical practitioners, which is crucial for rapid decision-making in medical diagnostics [38].
- *Alignment with Previous Studies:* Many studies in the field of cognitive assessment using CDT scores have successfully employed classification methods [6], [8], [9], [10], providing a benchmark and validation for our methodological choices. This approach allows for consistency and comparability across studies, enhancing the reliability of findings across different research contexts.
- *Simplicity and Accessibility:* Classification models are inherently simpler to implement and interpret compared to regression models [39], [40]. In the context of telemedicine applications, where our tool might be used by practitioners with varied levels of technical expertise, simplicity ensures broader accessibility and usability [38].
- *Risk Stratification:* Our classification method also allows for explicit risk stratification, which is essential in clinical settings. By categorizing patients into discrete risk groups based on CDT scores, clinicians can prioritize interventions more effectively.

C. Preprocessing and Feature Extraction

Digital CDT images undergo grayscale conversion and Gaussian filtering to capture object gradients, enabling enhanced feature extraction and effective noise reduction [41], [42]. Normalization is applied in both CDT images and structured data, so as to eliminate the influence of other transformation functions on the image, that is, convert it into a unique standard form to resist affine transformation [43], [44], and remove the unit or magnitude limitation of the structured data to convert it into a dimensionless pure value [45]. When using the gradient descent method to train a neural network in this work, normalization can speed up the solution speed of gradient descent, thereby speeding up the convergence of the network [46]. The images used in the actual experiment in this work are scanned, they have non-fixed size and multiple blank invalid regions around the drawing clock. To alleviate the varied quality and scale issue and enhance data validity, a selective search algorithm is employed for image cropping in CDT images. The resulting cropped images are resized to a dimension of 224 pixels $\times$ 224 pixels to facilitate subsequent processing.

Given the requirements for image feature representation devoid of handcrafting or human intervention, the variability in photo quality uploaded by users, and the limited color and texture information inherent in CDT images, this paper employs the Scale-Invariant Feature Transform (SIFT) [47] and Histogram of Oriented Gradient (HOG) [48] methods to extract local features from CDT images for representation purposes.

The SIFT descriptors can be obtained by building a gradient orientation histogram for a small region around each key point, and the key point can be obtained by computing the maxima or minima in the stack of Difference of Gaussians (DoG) images. The DoG image can be represented as

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma), \tag{1}$$

where $L(x, y, \sigma)$ means the convolution of the original image $I(x, y)$ with a Gaussian filter $G(x, y, \sigma)$.

The core of HOG is to divide the entire image into multiple small connected cells and calculate the gradient or edge direction histogram of each cell. The combination of these histograms can be used to form a feature descriptor.

D. Agile Multimodal Data Fusion Clusters

Multimodal data fusion encompasses the integration of two or more data modalities to extract enhanced features by leveraging the informational synergy among diverse data modalities. This process is typically categorized into three stages: early fusion, joint fusion, and late fusion, as delineated in existing literature [19], [20]. In this context, we propose the introduction of four agile multimodal data fusion clusters utilizing CDT images and structured data across all three fusion stages. This novel approach aims to address the current research gap and contribute to a more comprehensive understanding of multimodal data fusion in the specific context of ADRD screening based on CDT.

Early fusion is a process wherein features from multiple modal data are integrated into a singular feature vector, serving as the input for subsequent machine learning model training. Within this study, two distinct methodologies are employed for processing CDT images. The first approach involves utilizing two classical image feature representation algorithms, SIFT and HOG, to initially extract features from the images. These features are then used as input for training an MLP model. The second approach employs CNN to directly process CDT images, bypassing a specific step for image feature extraction. Both methods yield an intermediate output from CDT images, and the loss associated with this output is propagated back to the training model based on the images. The predicted quantitative CDT and image scores serve as the final features for CDT images, facilitating the fusion with structured features. Subsequently, another MLP model takes this fused feature vector as input for qualitative ADRD binary screening. Importantly, the loss of the final output solely propagates back to the subsequent MLP model, excluding the involvement of the preceding image-based training models. This strategic approach ensures an effective fusion of features and enhances the overall efficiency of the ADRD screening process.

Joint fusion involves the transformation of data from diverse modalities into feature representations, which are then interconnected with the final output. Within the proposed framework, CDT images and structured data serve as inputs for CNN and MLP, respectively. The feature representations obtained from the intermediate layers of these neural networks are fused through a concatenation layer before the final dense layer. This resulting joint feature representation is then linked to the ultimate qualitative binary output for ADRD screening. In joint fusion, a pivotal distinction from the early fusion lies in the critical point

that the loss incurred during the training of the final output is propagated back to both the CNN and MLP models, facilitating a collaborative, iterative process. This iterative joint propagation ensures a more coherent integration of information from CDT images and structured data, contributing to the refinement of the joint feature representation and, subsequently, the accuracy of ADRD screening.

Late fusion is characterized by a more straightforward conceptualization in comparison to early and joint fusion, as it involves the aggregation of data from distinct modalities at the decision level. Specifically, this framework entails the training of two classifiers–one for CDT images and another for structured data–with qualitative binary variables as the target. The output of each model is aggregated to form the final output. In the context of ADRD screening, there is a deliberate emphasis on identifying positive samples. Consequently, in this scenario, the final outcome is considered positive as long as at least one classifier produces a positive output. This strategic approach ensures a simplified yet effective late fusion model, facilitating the integration of information from diverse modalities for improved accuracy in ADRD screening.

E. Validity Analysis

Our research aims to use a unified framework to maximize its applicability by discovering key features through validation, showing the analysis process of structured features, and dissecting the CNN model for interpretability and transparency.

We used gradient-weighted class activation mapping (Grad-CAM) [17] in each convolutional layer of CNN to explore the process of CDT image feature extraction. Locally linear embedding (LLE) [18] is applied to the heatmap of each convolutional layer and the original images to obtain the difference between positive and negative samples at each key step of CNN and the change of the entire learning process, thus showing the evolution trend of CDT images.

LLE stands as a pivotal technique in manifold learning, specifically designed to preserve the local linear characteristics inherent in samples during the process of dimensionality reduction. Renowned for its efficacy in unraveling intricate structures within high-dimensional datasets, LLE represents a non-linear dimensionality reduction approach that finds widespread application across diverse domains. Its unique capability to discern complex patterns makes it particularly well-suited for tasks wherein the preservation of local relationships within the data is deemed paramount. LLE is often used to visualize high-dimensional data in two or three dimensions. By preserving the local structure, LLE can provide intuitive visualizations that help in understanding the intrinsic geometry of the data. In pattern recognition, LLE helps in identifying and classifying different styles by effectively mapping the high-dimensional data of image characters [49]. Therefore, LLE is employed here to reveal and understand the complex, high-dimensional structures in CDT images. Specifically, for each n-dimensional sample x_i in sample set $D = \{x_1, x_2, \ldots, x_m\}$, it can be represented by a linear combination of its k nearest neighbors, then the loss function is

$$J(w) = \sum_{i=1}^{m} \left\| x_i - \sum_{j=1}^{k} w_{ij} x_j \right\|^2. \tag{2}$$

The w_{ij} means the weight coefficient, and it can be obtained by minimizing (2) with the constraint

$$\sum_{j=1}^{k} w_{ij} = 1. \tag{3}$$

Finally, each n-dimensional sample x_i can be mapped into a d-dimensional $(d < n)$ sample y_i by minimizing the loss function

$$J(y) = \sum_{i=1}^{m} \left\| y_i - \sum_{j=1}^{k} w_{ij} y_j \right\|^2. \tag{4}$$

In addition, the information gain ratio (IGR) is used to explain the structured feature importance. Considering it is only applied in categorical variables, then recursive feature elimination (RFE) has been applied in all the structured features.

The reason for using IGR, a typical feature selection metric, is to lessen the bias of information gain, which tends to favor the features with more categories. IGR can be represented as

$$IGR(X, a) = \frac{H(X) - H(X|a)}{-\sum_i^n p_i log(p_i)}. \tag{5}$$

where X is a random variable, $H(X)$ is the entropy of X, $H(X|a)$ is the entropy of X given the value of attribute a, p_i is the proportion of class i in the dataset, and n is the total number of classes.

RFE is another feature selection technique, and the logistic regression model is used here as the algorithmic model to operate the RFE technique. The model will initially be trained on the full feature set, and RFE will recursively remove the least important features and fit the given algorithmic model on the pruned feature set until a specified number of features or a desired level of performance is achieved.

F. Performance Metrics

Four different metrics have been used to evaluate the model performance to reduce measurement bias for the imbalanced data as much as possible. For the qualitative binary screening, the prediction can be divided into four situations:

- *TP:* True Positive, predict the positive class as the positive class;
- *FP:* False Positive, predict the negative class as a positive class;
- *TN:* True Negative, predict the negative class as a negative class;
- *FN:* False Negative, predict the positive class as a negative class.

Then

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{6}$$

TABLE IV
FULLY AUTOMATIC ADRD SCREENING RESULTS

	AUC	Recall	Precision	F1-Score
	Quantitative CDT Score			
CNN	72%	41%	42%	0.41
SIFT + MLP	78%	38%	37%	0.37
SIFT + SVM	63%	38%	41%	0.39
HOG + MLP	87%	55%	53%	0.54
HOG + SVM	67%	45%	48%	0.46
	Quantitative Image Score			
CNN	97%	94%	91%	0.92
SIFT + MLP	98%	94%	91%	0.92
SIFT + SVM	84%	75%	92%	0.83
HOG + MLP	98%	92%	90%	0.91
HOG + SVM	82%	74%	92%	0.82
	Qualitative ADRD Binary Screening			
CNN	79%	74%	78%	0.76
SIFT + MLP	81%	73%	78%	0.75
SIFT + SVM	61%	68%	78%	0.73
HOG + MLP	81%	75%	74%	0.74
HOG + SVM	62%	65%	78%	0.71

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{7}$$

$$F1 - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{8}$$

The weighted average method is applied for multi-classification scenarios.

Also, the Area Under the Curve (AUC) is used to measure the entire two-dimensional area underneath the receiver operating characteristic (ROC) curve from (0, 0) to (1, 1). It tells how much the model is capable of distinguishing between classes, and the higher, the better. Compared with accuracy, it is not sensitive to whether data is balanced.

For the above four evaluation metrics, precision highlights the accuracy of positive predictions, recall gives an insight into how effectively the model identifies actual positives, and the F1-score provides a balanced measure of both. The AUC-ROC complements the precision, recall, and F1-score by providing a threshold-independent measure of model performance. While precision, recall, and F1-score give us threshold-dependent measures at specific operating points, the AUC-ROC provides a global view of the model's performance across all thresholds. These metrics provide a more nuanced view of the model's performance, particularly in terms of its ability to correctly identify positive samples.

IV. RESULTS AND ANALYSIS

A. Fully Automatic ADRD Screening

As Table IV shows, CNN and image feature extraction techniques both could have considerable performance for qualitative ADRD binary screening, and MLP (Neural Networks) can give more accurate results than SVM (traditional method) when modeling high-dimensional, heterogeneous, clinical data. From the quantitative perspective, the models (CNN, MLP and SVM) usually got better performance on the image scores than CDT scores because 94.34% samples have a score of 1 on image clarity. In addition, feature extraction techniques could capture effective features of CDT images for both quantitative and qualitative screening of ADRD.

TABLE V
SEMI-AUTOMATIC ADRD SCREENING RESULTS

	AUC	Recall	Precision	F1-Score
	Qualitative ADRD Binary Screening			
CNN + MLP: Concatenate	90%	83%	83%	0.83
CNN + MLP: Voting	77%	71%	85%	0.77
CNN + MLP: Incorporating	83%	77%	86%	0.81
MLP (SIFT) + MLP: Incorporating	83%	76%	86%	0.81
MLP (HOG) + MLP: Incorporating	82%	75%	86%	0.80

TABLE VI
FEATURE IMPORTANCE BASED ON RFE

Features	gender	race	intvrage, weight, height, wrdimmrc, wrddlyrc, clkdraw,clkimgcl
Rankings	3	2	1

B. Semi-Automatic ADRD Screening

Table V shows the results in semi-automatic screening case. The joint fusion method takes advantage of both CNN and MLP to achieve an AUC of 90% on ADRD prediction, and the other three metrics are improved to 83%, which means this fusion model is more sensitive for screening ADRD. The voting method, which is in the late fusion stage, cares more about positive samples than overall performance, which introduces more errors because it will include more false positives to cause overfitting, so it performs worse than using only images. It is worth noting that the incorporating method of early fusion takes the predicted CDT and image scores rather than the true values for dementia screening. And three image processing techniques obtained similar performance, which is slightly better than using only images. The results indicate that multimodal data fusion models could achieve better prediction performance compared to using only CDT images, and the predicted CDT and image scores could, instead of true scores, be used for effective qualitative binary ADRD screening.

C. Feature Importance Analysis

IGR was performed on categorical structured features, and according to the results shown in Fig. 3, cognitive variables are more important than demographic variables; the most important two are the score of the immediate word recall test and the delayed word recall test, respectively.

RFE then is applied to all the structured features, and the results are shown in Table VI where a lower ranking value means higher importance. It can be concluded that all the cognitive features tend to be more important than gender and race, and the interval age weight, and height (which have been normalized during pre-processing) also seem to be important by using RFE.

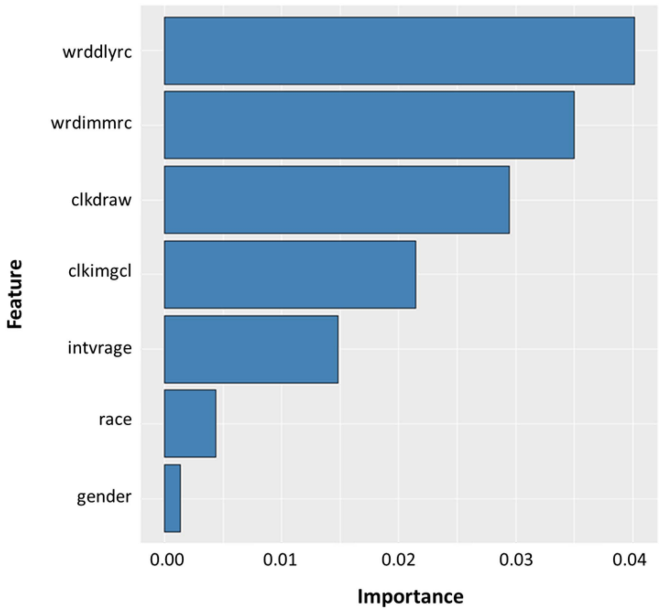

Fig. 3. Feature Importance based on IGR. IGR was performed on categorical structured features, showing cognitive variables are more important than demographic variables.

Although it's not capable of getting detailed rankings among the cognitive features, the RFE results are somewhat consistent with the results of IGR.

The integration of these two methods yields the finding that cognitive features, particularly those associated with memory tests, hold significant importance in ADRD screening. On the other hand, the significance of demographic features is relatively lower, with age intervals demonstrating a noteworthy correlation. This observation aligns with the established understanding that the risk of ADRD typically escalates with advancing age.

D. Visualization of CNN

Given the CNN model has good performance in the fully automatic instance for both quantitative and qualitative screening, it is worth investigating which pattern of the CDT images plays a decisive role. Fig. 4 shows original images of several random selected samples and their outputs from the first convolutional layer to the last one, and the valuable information can be known from the gradual changes of the heatmaps:

1) The first convolutional layer is a collection of various edge detectors. At this stage, the activation function retains almost all the information in the original image.
2) As layers go deeper, the information learned becomes more abstract and harder to understand intuitively. The deeper the layer, the less information there is about the visual content of the image.
3) The highlighted parts focus on the clock hands in heatmaps, including the center point of hands, the hands themselves and the direction and position of the hands. The minute hand seems to be more important than the hour hand. Some numbers and contours of the clock were also learned, although they were various among the images, mostly concentrated in the second and fourth quadrants.

Fig. 4 may reveal that positive samples tend to exhibit stronger commonalities compared to negative samples, potentially challenging the model's sensitivity in identifying subtler cases. However, this concern is effectively mitigated by our comprehensive approach to model design and implementation. Our advanced data preprocessing and feature extraction techniques are designed to enhance the model's ability to discern and learn from these variations. Our customized model employs robust regularization strategies, including dropout layer and batch normalization, to prevent overfitting and enhance generalization to diverse negative samples. Additionally, we experiment with various models in both fully and semi-automatic cases, and data fusion techniques in the semi-automatic case to improve predictive accuracy for clear ADRD cases, helping the model to recognize patterns common in positive samples essential for reliable screening.

In order to better explain the effectiveness of the CNN model and gain deeper insights, the original images and the outputs of the three convolutional layers are reduced to points in a standard two-dimensional coordinate system by modified LLE to observe the changing trend of the entire dataset. Fig. 5 shows the modified LLE outputs of the whole dataset from the original images to the last convolutional layer, each point represents an image, and the red point indicates that the sample is positive, the green point indicates that it is negative. Based on the LLE visualization, the original images are disorganized after LLE, and the positive and negative data are mixed together, which cannot be classified. The first convolutional layer classified a small part of positive samples and most of the negative samples, but there is still a large number of samples on the left side of the vertex that are mixed together. The second convolutional layer identifies more positive samples, but a considerable number of samples still cannot be classified. The last convolutional layer classified most of the positive samples, and only a small part of the samples on the left near the vertices are mixed with negative samples. This progressive change trend shows the working process of the CNN model, and to a certain extent unveils the mystery of the black box caused by the complexity of the abstract inference of CNN, which is important for improving the credibility of the model and providing transparency of the prediction results. It also further illustrates the effectiveness of this CNN model in predicting ADRD.

To gain deeper insights into the relationship between the prediction results and the extracted features by CNN, so as to further improve the interpretability of the model, Fig. 6 demonstrates the trend of the image features learned by the CNN model in the last convolutional layer by showing some samples, we also illustrate the class label of each sample. When it is assumed that a clock consists of three elements: outline, scale, and pointer, it would be very helpful for us to understand this trend more clearly.

On the left side of the figure are a large number of positive samples, from sample #1 to #5, it is known that the images of these samples are usually chaotic and have too much invalid information, and it seems that only some outline information

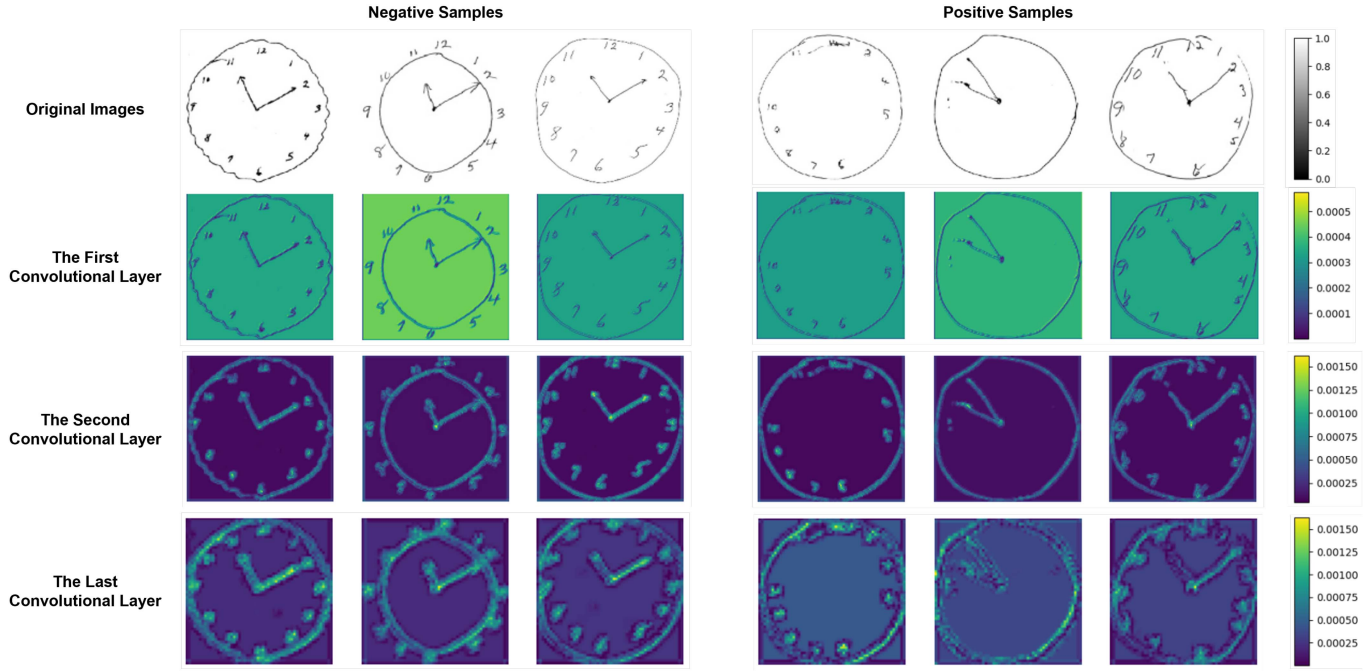

Fig. 4. Grad-CAM Visualization of CNN. It shows original images of several random selected samples and their outputs from the first convolutional layer to the last one, with the progression of layers, the acquired information becomes increasingly abstract. Heatmaps highlight significant features related to clock hands, encompassing the center point of the hands, the hands themselves, and the direction and position of the hands.

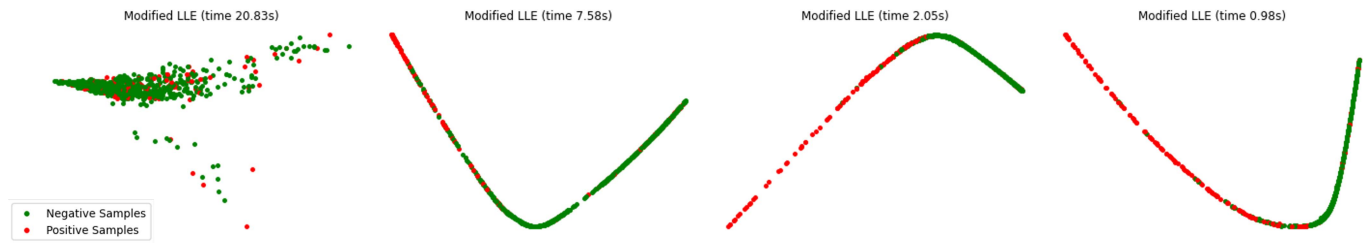

Fig. 5. LLE Visualization of CNN. The modified LLE outputs visually represent the entire dataset's transformation from original images to the last convolutional layer. Positive samples are denoted by red points, while negative samples are indicated by green points. The LLE visualization underscores the disorganization in the original images and illuminates the progressive changes achieved through each layer of the CNN model.

can be identified, but detailed features may not be obtained. On the left side of the vertex, there is a mixture of positive and negative samples. In this part, the outline of the clock is relatively clear, some images have scale information (sample #6 to #10), and some have pointer information (sample #8, #10 and #11), but the positive samples (sample #6, #7, and #11) usually contain only two of the three messages. More importantly, in this area, the images of positive and negative samples are very confusing, for example, for samples #9, #10, #11, of which #9 and #10 are negative, but their images are relatively chaotic, the indicated time also is wrong. On the contrary, the positive sample #11 has a clear outline and pointer, and the time indication is correct. It could be interpreted that the negative sample in this area will have a great risk of developing ADRD. In the area on the right side of the figure where the negative samples are clustered, the images are very clear, the standard circular outline, the position and orientation of the pointer, and the distribution of the clock scale, and even the numbers on the scale all can be recognized. That is to say, CNN has learned all three messages of outline, scale, and pointer in this part, as well as a lot of detailed information.

However, while the importance of outline, scale, and pointer features for ADRD screening is evident, further subdividing the dataset labels, such as missing only numbers or pointers, may complicate training without necessarily improving performance in this study. Such granularity could fragment the dataset, reduce statistical power, and shift focus from general cognitive impairment to specific drawing errors, misaligning with our goal of early ADRD detection. The absence of specific ground truth labels for these conditions could also undermine model training and validation.

V. Discussion, Insights and Future Plan

A. Discussion and Insights

Effectiveness: Based on the experimental results of this work, CDT has been shown to be an effective tool in fully and semi-automatic early ADRD screening. Both CNN and traditional

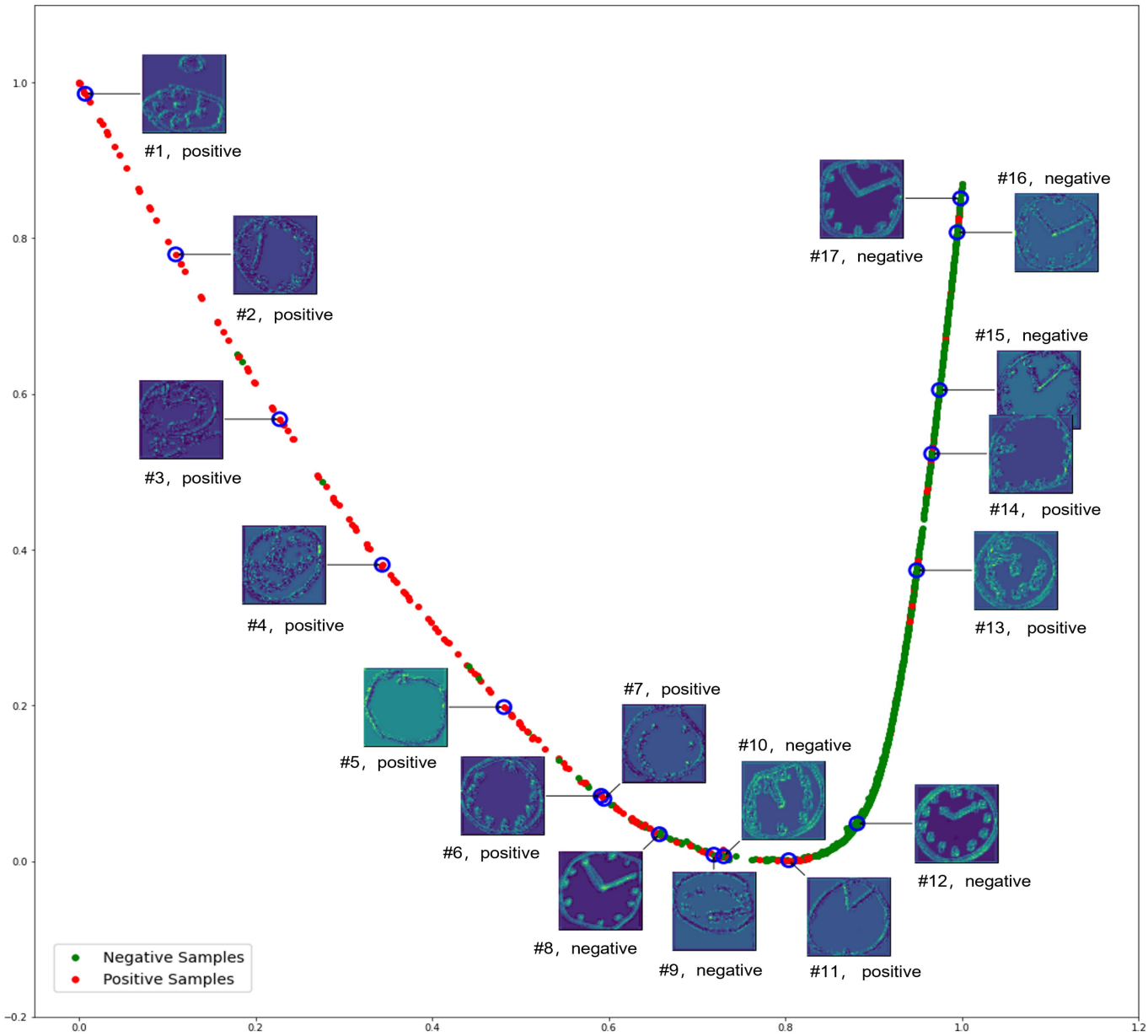

Fig. 6. Trend of the image features learned by CNN based on LLE. It elucidates the trend of image features learned by the CNN model in the last convolutional layer, enhancing interpretability. Notably, on the left side, positive samples exhibit chaotic images with identifiable outlines but lack detailed features. Towards the vertex, a mixture of positive and negative samples reveals varying clarity in clock elements and often lack one of the three components (outline, scale, or pointer). Conversely, the right side, where negative samples cluster, showcases clear images with recognizable outlines, scales, pointers, and detailed information.

image feature extraction techniques can capture key features of CDT images for accurate screening in both quantitative and qualitative terms. The neural network algorithm has been proved to be more effective in early ADRD screening than the classical SVM algorithm. The fusion of multimodal data, especially the cognitive features related to memory, can further improve performance.

In this study, we opted for a customized CNN model instead of larger or more sophisticated models due to considerations of dataset size, computational constraints, and the need for model interpretability. Although our dataset is substantial, it may not support the training of very deep networks without a significant risk of overfitting [50]. Additionally, deeper networks require extensive computational resources [51], which can be a challenge in telemedicine environments where such resources are often limited. Therefore, a lighter, simpler model is preferred to ensure effective deployment without high-end hardware, making the solution more accessible for typical clinical settings. The simpler architecture of our customized CNN also offers advantages in terms of modifiability, tuning,and interpretability, which are crucial in medical applications. The ability to easily understand and adjust the model based on preliminary results or specific clinical feedback ensures that our approach is not only effective but also practical and adaptable for clinical use.

Applicability: The proposed unified framework provides a detailed process for fully and semi-automatic ADRD screening process and also a comprehensive instance for multi-modal data fusion. It considers the fusion strategies of each stage, only when the loss of the final output is propagated back to the both models during training, it can more effectively help the model to extract features, thereby improving the performance.

In addition, the consideration of multi-modal data and telehealth-based framework design has great potential for ADRD screening by using other sensor data in the future. For example, high-dimensional heterogeneity data about gait and voice was obtained through the accelerometer and microphone of the mobile phone. What's more, users can receive simple instructions locally, and perform repeatable tests without strict environmental requirements. All calculations and data storage will be performed in the cloud. The feedback can be in real-time, and there is no requirement for their mobile device.

Transparency and Interpretability: Unlike most existing studies, which pre-set the errors or score items of CDT before training the model, this framework aims to reduce the bias introduced by human intervention as little as possible. The feature importance analysis and visualization of the model make it transparent and interpretable for understanding CDT-based early ADRD screening models and the process of extracting features, so as to obtain natural, rather than hand-crafted, key features. Also, four different metrics were used to eliminate as much as possible the bias introduced by unbalanced data and a single metric.

Dataset Sufficiency Validation Efforts: Our methodologies mitigate the limitations of relatively small and unbalanced dataset through several techniques that enhance the training process and model robustness. We employ models and architectures optimized for smaller datasets, incorporating regularization techniques such as dropout and batch normalization to prevent overfitting and ensure generalization to new data [30], [31]. The successful performance of these models is demonstrated through robust external validation and metrics such as AUC, precision-recall, and F1-scores, affirming the adequacy of the dataset size and the strategies' effectiveness in addressing data imbalances [52], [53].

Additionally, advanced visualization and interpretation techniques like Grad-CAM and LLE confirm that the dataset provides sufficient statistical power to identify meaningful differences and performances. In domain-specific applications like ADRD screening using the CDT, the domain knowledge and expert annotations, such as the CDT images and the cognitive assessment data in this study, inherent in the dataset's creation and the detailed annotations add depth [54], compensating for the dataset size and imbalances.

To further mitigate the dataset imbalances, we down-sampled the predominantly negative raw dataset to achieve a 1:5 ratio of positive to negative samples. We didn't adjust the dataset to a 1:1 ratio for several reasons:

First, the existing distribution of samples, although uneven, better reflects the real-world prevalence of ADRD within the general population to some extent [55]. By maintaining this distribution, we enhance the external validity and applicability of our models to actual clinical environments. Adjusting the ratio to 1:1 could introduce an artificial bias that might skew the model's performance metrics away from realistic conditions.

Second, we have implemented class weights during model training to address the imbalance issue effectively [56], [57]. And we assessed our models using metrics sensitive to data imbalances [52], [53]. By training our models on this unbalanced data, we aim to achieve a balanced sensitivity and specificity in detection. This is critical in clinical settings to avoid the significant consequences associated with both over-diagnosis and under-diagnosis. Our approach ensures that the models are sensitive to the less frequent positive cases while remaining accurate in identifying the more common negative cases.

However, there are two main limitations of this study:

- *Unbalanced and insufficient data:* The dataset's size of 1662 samples, while manageable, is smaller than ideal for neural network training. Despite our above efforts to mitigate class imbalance, we also adjust class weights during training, the existing 1:5 ratio of positive to negative samples poses ongoing concerns. This disparity could be even more problematic in real-world applications, where the ratio of positive to negative subjects could widen to 1:50, potentially impacting the model's generalizability and effectiveness in broader clinical settings.
- *Data quality:* The quality disparities between CDT images for binary classification pose a significant challenge. Negative samples often exhibit distorted lines, complicating accurate identification. This discrepancy arises from the NHATS survey data source, where subjects, despite reporting no dementia, may be at potential risk for ADRD or in early stages of mild cognitive impairment unbeknownst to them. Additionally, the presence of other diseases like Parkinson's Disease (PD) and stroke may influence CDT performance. Consequently, the CNN model, relying solely on raw clock drawing images, may encounter limitations in discerning distinctive features for effective screening.

B. Future Plan

Based on the above discussion and insights, our focus shifts towards pragmatic considerations, outlining actionable strategies and concrete plans for future developments for this study.

In future studies, a significant avenue for exploration lies in the development and utilization of customized models tailored specifically to the unique challenges and nuances of the ADRD-related research domain. Customized models offer the potential for greater accuracy and efficiency by incorporating ADRD domain-specific knowledge and data characteristics directly into their architecture. This approach can lead to models that are more attuned to the subtle patterns and intricacies inherent in data of subjects who have a high risk of ADRD, potentially uncovering insights that more generic models might miss.

Another promising area for future research is the application of advanced techniques for model fusion, which can combine insights from multiple models or data sources and is crucial

in scenarios where no single model provides a complete picture. Advanced fusion techniques, such as self-attention mechanisms [58] or neural network ensemble methods [59], can provide more nuanced and effective ways of integrating diverse sources of information.

As this benchmark study has concluded that cognitive features, especially those related to the memory test, play a more important role in ADRD screening than demographic features. Future research endeavors in the realm of ADRD screening have a significant opportunity to focus on the detailed analysis of cognitive information. Cognitive symptoms are among the earliest signs of ADRD, and a more nuanced understanding of these changes can lead to earlier and more accurate screening [60]. Then, a compelling area warranting further exploration involves the longitudinal analysis of cognitive data [61], [62]. NHATS actually provides us with this potential opportunity. It follows the same individuals over time, allowing researchers to observe changes and trends in health and functioning as people age. This design is crucial for studying the dynamics of aging, including the progression of chronic diseases, such as ADRD. Tracking cognitive function over time in individuals can provide valuable insights into the progression and potential early signs of cognitive decline. This approach can help in distinguishing between normal aging-related changes and those indicative of ADRD. Advanced statistical methods and machine learning models can be employed to analyze this longitudinal data, identifying trends and changes that are predictive of ADRD.

We acknowledge that class imbalance is an ongoing challenge in this field, and we aim to explore further methodologies in our future work to continue improving the robustness and reproducibility of our findings. One of the potential improvement methodologies will be exploring and refining advanced sampling methods. We intend to explore variations of the Synthetic Minority Oversampling Technique (SMOTE) [63], which generates synthetic samples in the feature space. This can be further enhanced by integrating SMOTE with the undersampling of the majority class to create a more balanced dataset without significant information loss. Another possible way is exploring advanced boosting techniques like AdaBoost [64], [65] and Gradient Boosting [65] with modifications for imbalance. These methods focus more on the misclassified examples and can be adapted to give more weight to the minority class.

VI. Conclusion

A mobile framework focusing on multimodal data fusion and model interpretation for fully and semi-automatic ADRD screening has been established in this study. It can take high-dimensional multi-modal data as inputs and reveal the key features of CDT images learned by the CNN model and its evolution to eliminate the manual intervention during model training.

The framework proves that traditional image feature techniques (SIFT and HOG) and neural networks are both effective for CDT-based ADRD screening. MLP performs much better than SVM when processing the extracted high-dimensional image features. The multi-modal data fusion technique could increase the prediction performance, and the additional cognitive variables, especially the memory test features, are more important than demographic variables. Also, the predicted CDT and image scores have great potential to substitute true values for binary qualitative ADRD screening.

The visualization and interpretation of the CNN model show that outline, scale, and clock hands are three key points for ADRD screening. The center point of hands, the hands themselves, and the direction and position of the hands, especially the minute hand, seem to play an essential role in screening.

References

[1] W. H. Organization, "Dementia," 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] N. C. f. C. D. P. Division of Population Health and H. Promotion, "Alzheimer's disease and related dementias," 2018. [Online]. Available: https://www.cdc.gov/aging/publications/features/alzheimers-disease-dementia.html

[3] N. N. I. Aging, "Alzheimer's disease fact sheet," 2021. [Online]. Available: https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet

[4] A. Paganini-Hill, L. J. Clark, V. W. Henderson, and S. J. Birge, "Clock drawing: Analysis in a retirement community," *J. Amer. Geriatrics Soc.*, vol. 49, no. 7, pp. 941–947, 2001.

[5] D. H. Yoo and J. S. Lee, "Clinical usefulness of the clock drawing test applying rasch analysis in predicting of cognitive impairment," *J. Phys. Ther. Sci.*, vol. 28, no. 7, pp. 2140–2143, 2016.

[6] Y. C. Youn et al., "Use of the clock drawing test and the Rey–Osterrieth complex figure test-copy with convolutional neural networks to predict cognitive impairment," *Alzheimer's Res. Ther.*, vol. 13, no. 1, pp. 1–7, 2021.

[7] L. Babins, M.-E. Slater, V. Whitehead, and H. Chertkow, "Can an 18-point clock-drawing scoring system predict dementia in elderly individuals with mild cognitive impairment?," *J. Clin. Exp. Neuropsychol.*, vol. 30, no. 2, pp. 173–186, 2008.

[8] S. Chen, D. Stromer, H. A. Alabdalrahim, S. Schwab, M. Weih, and A. Maier, "Automatic dementia screening and scoring by applying deep learning on clock-drawing tests," *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020.

[9] C. Qian and M. Liao, "An intelligent screening mobile application for Alzheimer's disease using clock drawing test," in *Proc. 4th Int. Conf. Signal Process. Mach. Learn.*, 2021, pp. 112–116.

[10] S. Amini et al., "An artificial intelligence-assisted method for dementia detection using images from the clock drawing test," *J. Alzheimer's Dis.*, vol. 83, no. 2, pp. 581–589, 2021.

[11] E. Seigerschmidt, E. Mösch, M. Siemen, H. Förstl, and H. Bickel, "The clock drawing test and questionable dementia: Reliability and validity," *Int. J. Geriatr. Psychiatry*, vol. 17, no. 11, pp. 1048–1054, 2002.

[12] A. Y. Lee, J. S. Kim, B. H. Choi, and E. H. Sohn, "Characteristics of clock drawing test (CDT) errors by the dementia type: Quantitative and qualitative analyses," *Arch. Gerontol. Geriatrics*, vol. 48, no. 1, pp. 58–60, 2009.

[13] S. Cosentino, A. Jefferson, D. L. Chute, E. Kaplan, and D. J. Libon, "Clock drawing errors in dementia: Neuropsychological and neuroanatomical considerations," *Cogn. Behav. Neurol.*, vol. 17, no. 2, pp. 74–84, 2004.

[14] I. Park and U. Lee, "Automatic, qualitative scoring of the clock drawing test (CDT) based on U-Net, CNN and mobile sensor data," *Sensors*, vol. 21, no. 15, 2021, Art. no. 5239.

[15] N. A. Talwar et al., "The neural correlates of the clock-drawing test in healthy aging," *Front. Hum. Neurosci.*, vol. 13, 2019, Art. no. 25.

[16] A. Merkin, R. Krishnamurthi, and O. N. Medvedev, "Machine learning, artificial intelligence and the prediction of dementia," *Curr. Opin. Psychiatry*, vol. 35, no. 2, pp. 123–129, 2022.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[18] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[19] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–9, 2020.

[20] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[21] B. Fling, *Mobile Design and Development: Practical Concepts and Techniques for Creating Mobile Sites and Web Apps*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2009.

[22] D. S. Kolluru and P. B. Reddy, "Analysis of parameters used for measuring performance of mobile applications," *Int. J. Anal. Appl.*, vol. 19, no. 4, pp. 587–603, 2021.

[23] M. Abbas, "Wi-fi generations definition, timeline and benefits," Oct. 2021. [Online]. Available: https://5ghub.us/wi-fi-generations-definitiontimeline-and-benefits/

[24] "Understanding the difference between 3G 4G and 5G networks," Jul. 2023. [Online]. Available: https://utilitiesone.com/understanding-the-difference-between-3g-4g-and-5g-networks

[25] N. NHATS, "National health and aging trends study (NHATS)–NHATS," 2022. [Online]. Available: https://nhats.org/researcher/nhats

[26] J. A. S. Vicki, A. Freedman, and M. E. Skehan, "National health and aging trends study user guide: Rounds 1–12 final release," 2022. [Online]. Available: https://nhats.org/researcher/nhats/methods-documentation?id=user_guide

[27] V. A. Freedman and J. D. Kasper, "Cohort profile: The national health and aging trends study (NHATS)," *Int. J. Epidemiol.*, vol. 48, no. 4, pp. 1044–1045, 2019.

[28] M. L. F. Chaves and A. L. Camozzato, "How many items from a word list can Alzheimer's disease patients and normal controls recall? Do they recall in a similar way?," *Dement. Neuropsychologia*, vol. 1, pp. 52–58, 2007.

[29] J. Wu, "Introduction to convolutional neural networks," *Nat. Key Lab Novel Softw. Technol. Nanjing Univ. China*, vol. 5, no. 23, 2017, Art. no. 495.

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] B. Ghaddar and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines," *Eur. J. Oper. Res.*, vol. 265, no. 3, pp. 993–1004, 2018.

[33] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: A study on high-dimensional spaces," *Knowl. Inf. Syst.*, vol. 12, pp. 95–116, 2007.

[34] D. Anguita, A. Ghio, N. Greco, L. Oneto, and S. Ridella, "Model selection for support vector machines: Advantages and disadvantages of the machine learning theory," in *Proc. Int. Joint Conf. Neural Netw.*, 2010, pp. 1–8.

[35] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Trans. Circuits Syst.*, vol. 8, no. 7, pp. 579–588, 2009.

[36] A. Vehtari and J. Lampinen, "Bayesian MLP neural networks for image analysis," *Pattern Recognit. Lett.*, vol. 21, no. 13/14, pp. 1183–1191, 2000.

[37] P. Naraei, A. Abhari, and A. Sadeghian, "Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data," in *Proc. Future Technol. Conf.*, 2016, pp. 848–852.

[38] N. Mateussi et al., "Clinical applications of machine learning," *Ann. Surg. Open*, vol. 5, no. 2, 2024, Art. no. e423.

[39] GeeksforGeeks, "Advantages and disadvantages of different classification models — GeeksforGeeks," Sep. 2020. [Online]. Available: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-classification-models/

[40] E. Nasarian, R. Alizadehsani, U. R. Acharya, and K.-L. Tsui, "Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework," *Inf. Fusion*, vol. 108, 2024, Art. no. 102412.

[41] H. M. Bui, M. Lech, E. Cheng, K. Neville, and I. S. Burnett, "Using grayscale images for object recognition with convolutional-recursive neural network," in *Proc. IEEE 6th Int. Conf. Commun. Electron.*, 2016, pp. 321–325.

[42] S. Das, J. Saikia, S. Das, and N. Goni, "Comparative study of different noise filtering techniques in digital images," *Int. J. Eng. Res. Gen. Sci.*, vol. 3, no. 5, pp. 180–191, 2015.

[43] K.-M. Koo and E.-Y. Cha, "Image recognition performance enhancements using image normalization," *Hum.-Centric Comput. Inf. Sci.*, vol. 7, no. 1, pp. 1–11, 2017.

[44] S.-C. Pei and C.-N. Lin, "Image normalization for pattern recognition," *Image Vis. Comput.*, vol. 13, no. 10, pp. 711–723, 1995.

[45] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, vol. 72. Berlin, Germany: Springer, 2015.

[46] J. Sola and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," *IEEE Trans. Nucl. Sci.*, vol. 44, no. 3, pp. 1464–1468, Jun. 1997.

[47] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[48] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.

[49] R. Hettiarachchi and J. F. Peters, "Multi-manifold LLE learning in pattern recognition," *Pattern Recognit.*, vol. 48, no. 9, pp. 2947–2960, 2015.

[50] B. Sabiri, B. El Asri, and M. Rhanoui, "Mechanism of overfitting avoidance techniques for training deep neural networks," in *Proc. Int. Conf. Enterprise Inf. Syst.*, vol. 1, 2022, pp. 418–427.

[51] H. Gao, Z. Wang, and S. Ji, "Large-scale learnable graph convolutional networks," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1416–1424.

[52] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf. Sci.*, vol. 513, pp. 429–441, 2020.

[53] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *Int. J. Data Mining Knowl. Manage. Process*, vol. 5, no. 2, 2015, Art. no. 1.

[54] B. Ljubic et al., "Influence of medical domain knowledge on deep learning for Alzheimer's disease prediction," *Comput. Methods Programs Biomed.*, vol. 197, 2020, Art. no. 105765.

[55] A. Association, "Alzheimer's disease facts and figures." 2024. [Online]. Available: https://www.alz.org/alzheimers-dementia/facts-figures

[56] T. Core, "Classification on imbalanced data," Aug. 20, 2024, [Online]. Available: https://www.tensorflow.org/tutorials/structureddata/imbalanced_data

[57] G. J., "Using class weight to compensate for imbalanced data – medium," Jul. 2022. [Online]. Available: https://medium.com/bubbapora_76246/using-class-weight-to-compensate-for-imbalanced-data-6eff370185d3

[58] H. Zhu, Z. Wang, Y. Shi, Y. Hua, G. Xu, and L. Deng, "Multimodal fusion method based on self-attention mechanism," *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1–8, 2020.

[59] J. Liu, S. Shang, K. Zheng, and J.-R. Wen, "Multi-view ensemble learning for dementia diagnosis from neuroimaging: An artificial neural network approach," *Neurocomputing*, vol. 195, pp. 112–116, 2016.

[60] J. C. Morris et al., "Mild cognitive impairment represents early-stage Alzheimer disease," *Arch. Neurol.*, vol. 58, no. 3, pp. 397–405, 2001.

[61] S. S. Sullivan, W. Bo, C.-S. Li, W. Xu, and Y.-P. Chang, "Predicting hospice transitions in dementia caregiving dyads: An exploratory machine learning approach," *Innov. Aging*, vol. 6, no. 6, 2022, Art. no. igac051.

[62] F. Li, H. Takechi, A. Kokuryu, and R. Takahashi, "Longitudinal changes in performance on cognitive screening tests in patients with mild cognitive impairment and Alzheimer disease," *Dement. Geriatr. Cogn. Disord. Extra*, vol. 7, no. 3, pp. 366–373, 2018.

[63] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018.

[64] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statist. Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[65] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: An experimental review," *J. Big Data*, vol. 7, pp. 1–47, 2020.