RESEARCH-ARTICLE

An Interpretable AI Model for Alzheimer's Disease Detection using Retinal Images

SAEED JAMSHIDIHA, University of Tehran, Tehran, Tehran, Iran

ALIREZA REZAEE, University of Tehran, Tehran, Tehran, Iran

FARSHID HAJATI, University of New England Australia, Armidale, NSW, Australia

MOJTABA GOLZAN, University of Technology Sydney, Sydney, NSW, Australia

RAYMOND CHOO CHIONG, University of New England Australia, Armidale, NSW, Australia

Open Access Support provided by:

University of New England Australia

University of Technology Sydney

University of Tehran

Published: 16 May 2025

ICMHI 2025: 2025 9th International Conference on Medical and Health Informatics
May 16 - 18, 2025
Kyoto, Japan

https://doi.org/10.1145/3761712.3761745
ISBN: 9798400715143

An Interpretable AI Model for Alzheimer's Disease Detection using Retinal Images

Saeed Jamshidiha
Nanotechnology, Biotechnology, Information Technology and Cognitive Science Laboratory
University of Tehran
Tehran, Iran
s_jamshidiha@aut.ac.ir

Alireza Rezaee
Department of Mechatronics, School of Intelligent Systems, College of Interdisciplinary Science and Technology
University of Tehran
Tehran, Iran
arrezaee@ut.ac.ir

Farshid Hajati*
School of Science and Technology, Faculty of Science, Agriculture, Business and Law
University of New England
Armidale, Australia
fhajati@une.edu.au

Mojtaba Golzan
Graduate School of Health
University of Technology Sydney
Ultimo, Australia
Mojtaba.Golzan@uts.edu.au

Raymond Chiong
School of Science and Technology, Faculty of Science, Agriculture, Business and Law
University of New England
Armidale, Australia
School of Medicine and Public Health
The University of Newcastle
Callaghan, Australia
rchiong@une.edu.au

Abstract

Alzheimer's disease (AD) is a progressive cognitive and neurological disorder that affects millions of people worldwide. Retinal imaging has emerged as a cheap, non-invasive biomarker for AD, using which artificial intelligence (AI) models can be trained to streamline AD screening and diagnosis. Given the sensitivity of AD diagnosis, the lack of interpretability of conventional AI models, such as convolutional neural networks (CNNs), is an important barrier to their practical application. In this context, the Kolmogorov-Arnold network (KAN) architecture provides an intrinsically interpretable alternative to traditional deep learning models. Using KANs, we propose an interpretable AI model for AD detection that achieves an accuracy of 87%, outperforming state-of-the-art models by margins of more than 6%.

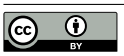
CCS Concepts

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision tasks**; • **Visual inspection**;

Keywords

Alzheimer's Disease, Kolmogorov-Arnold Networks, Interpretable Artificial Intelligence, Optical Coherence Tomography

*Corresponding Author:



ICMHI 2025, Kyoto, Japan

ACM ISBN 979-8-4007-1514-3/2025/05
https://doi.org/10.1145/3761712.3761745

ACM Reference Format:
Saeed Jamshidiha, Alireza Rezaee, Farshid Hajati, Mojtaba Golzan, and Raymond Chiong. 2025. An Interpretable AI Model for Alzheimer's Disease Detection using Retinal Images. In *2025 9th International Conference on Medical and Health Informatics (ICMHI 2025), May 16–18, 2025, Kyoto, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3761712.3761745

1 Introduction

Alzheimer's disease (AD) is a progressive neurological disorder characterised by cognitive decline and memory loss. Early detection of AD can significantly impact the management of the disease, making accurate screening and diagnostic tools essential for patient care [1]. Artificial intelligence (AI) and computer vision can streamline AD detection and improve its speed and accuracy [1]. Retinal imaging has recently emerged as a potential cheap and non-invasive biomarker for AD detection [2].

A modular machine learning (ML)-based method for AD detection using retinal images is presented in [3]. This method is comprised of three modules. The first module, which is a neural network ensemble, detects and excludes low-quality images. The second module, a U-Net model, performs blood vessel segmentation. The third module, a support vector machine (SVM), classifies retinal images with an accuracy of 82.44%. EfficientNet-B2 is utilised by the authors of [4] to detect AD using retinal imaging, achieving an accuracy of 83.6%. Residual neural networks (ResNets) are leveraged in [5] for AD detection using multi-modal retinal image data, with a precision of 74.4% and a recall of 72.7%. The authors of [6] compare the performance of multiple ML methods, reporting accuracies ranging from 67% to 74%. The authors of [7] use EfficientNetB-3 to capture the relationship between retinal fundus images and cognitive scores, explaining 22.4% of the variance of cognitive scores

using both retinal images and their respective metadata, and 9.3% of the variance using metadata alone. The study undertaken in [8] leverages VGG-16 and explainable artificial intelligence (XAI) methods to assess the continuum of AD, reporting an accuracy of 71.4%. The authors of [9] utilise MobileNet for AD classification using retinal fundus images and report an area under the curve of the receiver operating characteristic curve of 0.927.

While deep learning-based machine vision models, including convolutional neural networks (CNNs) have achieved high accuracy in various tasks, including AD detection, they are generally considered black box models due to their complex, non-linear transformations of input data, which is particularly concerning in medical applications where reliability and trustworthiness are paramount. While XAI methods can offer insight into the decisions made by these models, they are still intrinsically not interpretable.

The Kolmogorov-Arnold network (KAN) architecture offers an interpretable alternative to conventional deep learning models. The KAN model is a novel architecture motivated by the Kolmogorov-Arnold representation theorem, stating that any continuous multivariate function can be written as the sum of continuous univariate functions. Unlike other neural network models, the KAN model uses learnable activation functions [10]. B-Splines and wavelets are among the commonly used activation functions in the KAN architecture [10, 11]. The efficacy of KANs for image classification has been demonstrated in [12] by comparing the KAN model with MLPs and ResNets on CIFAR10 [13], CIFAR100 [13], and MNIST [14] datasets.

The interpretability of the KAN model makes it an attractive choice in the context of medical diagnosis, where transparency and reliability are crucial.

This study investigates the efficacy of KANs in AD detection using retinal images. Our findings contribute to the development of more accurate, reliable, and streamlined screening and diagnostic tools for AD detection.2 Materials and Methods

1.1 Data

The publicly available optical coherence tomography (OCT) dataset [15] is utilised in this research to train and test the proposed model and the benchmark algorithms. Two-dimensional slices are extracted from three-dimensional OCT images and down sampled to 50x50 pixels, resulting in 1120 two-dimensional 50x50 grayscale images.

1.2 Kolmogorov-Arnold Networks

Assuming that f is a continuous multi-variate function over a bounded domain, the Kolmogorov-Arnold theorem states that f can be decomposed into a sum of univariate functions [10]:

$$f(x) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right), \tag{1}$$

The Kolmogorov-Arnold representation theorem (1) is equivalent in formulation to a 2-layer neural network. This architecture can be generalised to an arbitrary number of layers (L) [10]:

$$f(x) = \sum_{i_{L-1}=1}^{n_{L-1}} \phi_{L,i_L,i_{L-1}} \left(\sum_{i_{L-2}=1}^{n_{L-2}} \phi_{L-1,i_{L-1},i_{L-2}} \left(\cdots \left(\sum_{i_0=1}^{n_0} \phi_{0,i_1,i_0}(x_{i_0}) \right) \cdots \right) \right), \tag{2}$$

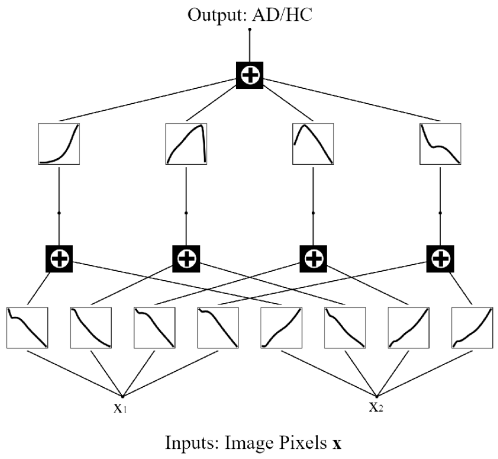


Figure 1: A 2 layer KAN

where n_i is the number of nodes in the i-th layer of the KAN, $\phi_{l,i,j}$ is the activation function connecting the i-th node in layer l to the j-th node in layer l+1, and the number of nodes of the final layer (n_L) is assumed to be one.

In the KAN architecture introduced in [10], the ϕThe The $\phi_{l,i,j}$ functions are defined as follows:

$$\phi_{l,i,j}(x) = w_{b,i,j,l} b(x) + w_{s,i,j,l} s(x), \tag{3}$$

where $w_{b,i,j,l}$ and $w_{s,i,j,l}$ are learnable weights of the fixed and learnable activation functions, respectively, s(x) denotes the B-spline function, and b(x) is a fixed activation function, such as the swish function:

$$b(x) = \frac{x}{1 + e^{-x}}. \tag{4}$$

While, in principle, the learnable weights can be absorbed into b(x) and s(x), they offer better control over the magnitude of the activation functions [10].

The optimal hyperparameters of the KAN model, obtained using (non-nested) 3-fold cross-validation and Bayesian optimisation are presented in Table 3, and the optimal hyperparameters of the benchmark models are presented in Table 4.

Fig. 1 depicts a KAN with 2 layers, a 2-dimensional input (on the lower side), and a 1-dimensional output (on the upper side). As shown in the figure, each input feature passes an activation function, and the results are all summed up, producing the features in the second layer of the network. The second layer features each pass another activation function and are summed up to produce the network output. Furthermore, the activation functions of the KAN model are learned, and the activation function of each neuron differs from the activation functions of other neurons. Each function is univariate, and only receives one input variable.

1.3 Computational Complexity

Denoting the number of KAN layers with L, the number of spline functions with N, the number of spline knots with G, and spline order with k, the computational complexity of the KAN model is

Table 1: The search space for the KAN model

Hyperparameter	Possible Values	Prior Distribution
Number of Layers	{1, 2, ..., 5}	Uniform
Number of Spline Functions per Layer	{10, 11, ..., 100}	Uniform
Number of Spline Knots	{0, 1, ..., 100}	Uniform
Fixed Activation Functions	{RELU, GELU, Sigmoid}	Uniform

Table 2: The search space for the CNN-based benchmarks

Hyperparameter	Possible Values	Prior Distribution
Number of MLP Layers	{1, 2, 3, 4}	Uniform
Number of Neurons per MLP Layer	{100, 101, ..., 4096}	Exponential

Table 3: KAN Hyperparameter Values

Hyperparameter	Value
Layers	3
Units per Layer	72
Spline Order	3
Grid Size	11
Fixed Activation Function	GELU

Table 4: The hyperparameters of the benchmark models

Model	MLP Layers	Neurons per Layer
EfficientNet-B3	3	829
MobileNet	3	870
ResNet-50	3	458
VGG-16	4	3599

$O(N^2L(G+k))$ [10]. The computational complexity of the CNN-based benchmarks is $O(N^2LD^2)$, where L denotes the number of CNN layers, N denotes the number of feature maps in each layer, and D denotes the dimensionality of each feature map [16].

2 Results

To assess the performance of the proposed method, its classification accuracy, precision, recall, and area under the curve (AUC) of its receiver operating characteristic curve are compared with RETFound, a transformer-based foundation model for retinopathy detection [17], as well as EfficientNet-B3, MobileNet, ResNet-50 and VGG-16. The RETFound model is initialised from the weights published by its authors [17], while the other benchmark methods are initialised from weights resulting from pretraining on the ImageNet dataset [18]. The final classifier layers of the benchmark models are discarded, and an MLP classifier head is added on top of the benchmark models. The number of MLP layers and the number of units per MLP layer constitute the hyperparameters of the benchmark models. The KAN model is trained from scratch using the OCT data.

Table 5: Comparison of classification metrics

Model	Accuracy	Precision	Recall	AUC
Proposed Model	**87%**	**88%**	**86%**	**88%**
RETFound	79%	81%	73%	79%
EfficientNet-B3	54%	53%	62%	60%
MobileNet	50%	26%	85%	56%
ResNet-50	73%	81%	76%	84%
VGG-16	81%	86%	80%	87%

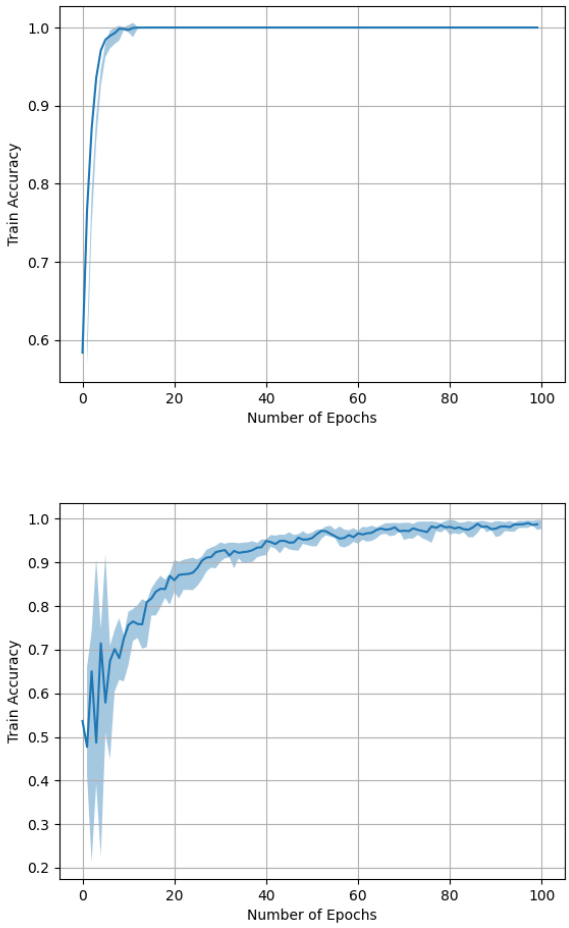


Figure 2: (a) Accuracy vs. Number of Training Epochs for the KAN model (b). Accuracy vs. Number of Training Epochs for the VGG-16 model

2.1 Hyperparameters

Instead of performing a grid search or a random search to obtain the optimal set of hyperparameters for the KAN model and the benchmark models, we have utilised Bayesian optimisation in this research. Instead of sampling the hyperparameter space at fixed intervals like grid search or at random points like random search, Bayesian optimisation samples the hyperparameter space at points where there is the highest probability of improvement over the best-observed performance (accuracy, precision, etc.). The hyperparameter search space and prior distributions over the hyperparameters are presented in Table 1 and Table 2 for KAN and the CNN-based benchmarks, respectively.

2.2 Performance Evaluation

In order to select the optimal hyperparameters for the KAN model as well as the benchmark models while performing model selection to determine the best model, we have utilised nested k-fold cross-validation along with Bayesian optimisation in this research. The inner folds are used to obtain the optimal set of hyperparameters, while the outer folds are used to assess the classification performance of the models. The results are presented in Table 5.

As evident in Table 5, the proposed KAN-based method outperforms the benchmarks in all classification metrics, showcasing KAN's potential in disease diagnosis using medical images.

2.3 Convergence Speed

In order to compare the convergence speed of KAN to the benchmarks, the train accuracies of 10 different randomly initiated KANs with the optimal hyperparameters obtained in Section 3.1 are averaged and presented in Fig. 2a. The shaded area highlights one standard deviation on each side of the average train accuracy. Fig. 2b presents the same curve for VGG-16, which is the best benchmark algorithm in Section 3.2.

Keep in mind that KAN is being trained from scratch while VGG-16 is initialised from weights obtained on the ImageNet dataset and is being fine-tuned on the OCT data.

Figure 3: samples of retinal OCT images

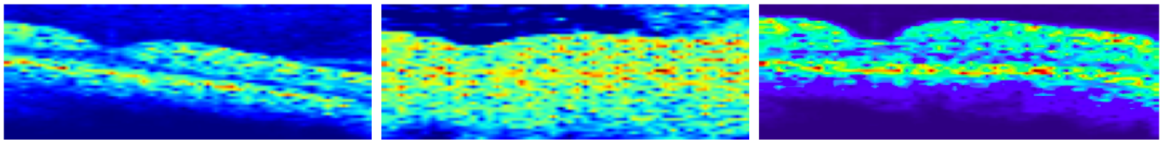


Figure 4: Samples of retinal OCT images multiplied by the weights of the KAN model

2.4 Interpretability

As stated earlier, the KAN model is interpretable and can offer insight into the input-output relationship being studied. Since we are applying the KAN model to a high-dimensional problem space (images), it is difficult to assess the input-output relationship learned by the KAN model using the methodology presented in [10].

In order to provide insight into the classifications provided by the KAN model, the pixel values of sample images from the OCT dataset, shown in Fig 3, are multiplied by the weights from the splines in the first layer of the KAN model ($w_{s,i,j,1}$) to highlight the areas in the images that are most relevant to the decision making of the KAN model. The resulting heatmaps are shown in Fig 3

The heatmaps show the highest weights being assigned to the regions in the image corresponding to the choroid layer of the retina. They also highlight the general area of the retina. These areas are in line with clinical studies, which suggest a link between retinal thickness and AD [2], and between the thickness of the choroid layer, responsible for retinal blood flow and AD [2, 19].

3 Discussion

In this research, we show the potential of the KAN model in medical machine vision tasks, specifically in AD detection using retinal images. It is shown that the KAN model can achieve superior classification performance in all of the assessed classification metrics with fewer parameters and smaller network sizes in comparison with CNN-based benchmarks.

The superior convergence speed of the KAN model can be attributed to its lower complexity compared to the VGG-16 model. While the computational complexity of a KAN model is similar to a VGG-16 model with the same number of layers and the same number of units in each layer, the KAN model can learn the classification problem using a much smaller network in comparison to the VGG-16 model, resulting in a significantly lower computational complexity. Here, the optimal set of KAN hyperparameters results in a model with approximately 200 thousand parameters, while the optimal set of VGG-16 hyperparameters results in a model with 42 million parameters.

Another advantage of the KAN model over other deep learning models, which are inherently black-box models, is its interpretability, offering insight into the classification process of the network. This enables experts, e.g. ophthalmologists, to review the predictions made by the KAN model. Leveraging this inherent interpretability, we show that the regions most relevant to the decision-making process of the model align well with clinical findings.

4 Conclusion

In this paper, a KAN-based interpretable AI model was proposed for AD detection using retinal imaging. The proposed model was evaluated using OCT data, outperforming state-of-the-art benchmarks by margins of more than 6%, with a much smaller network size in comparison to the CNN-based benchmark algorithms. It was also shown that the regions most important to the KAN model's classification align well with clinical findings. Overall, the results showcase the potential of KAN in developing disease detection algorithms using medical imaging. Future work can focus on exploring other KAN basis functions, such as wavelets and Fourier transforms.

Acknowledgments

This work was supported by the University of New England, Australia, under grant number A24/2676.

Author Statement

This paper was written entirely by the authors without the use of any artificial intelligence (AI) tools.

References

[1] Ebrahimighahnavieh MA, Luo S, Chiong R. Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. Computer methods and programs in biomedicine. 2020;187:105242.

[2] Cheung CY, Mok V, Foster PJ, Trucco E, Chen C, Wong TY. Retinal imaging in Alzheimer's disease. Journal of Neurology, Neurosurgery & Psychiatry. 2021;92(9):983-94.

[3] Tian J, Smith G, Guo H, Liu B, Pan Z, Wang Z, *et al.* Modular machine learning for Alzheimer's disease classification from retinal vasculature. Scientific Reports. 2021;11(1):238

[4] Cheung CY, Ran AR, Wang S, Chan VT, Sham K, Hilal S, *et al.* A deep learning model for detection of Alzheimer's disease based on retinal photographs: a retrospective, multicentre case-control study. The Lancet Digital Health. 2022;4(11):e806-15.

[5] Wisely CE, Wang D, Henao R, Grewal DS, Thompson AC, Robbins CB, *et al.* Convolutional neural network to identify symptomatic Alzheimer's disease using multimodal retinal imaging. British Journal of Ophthalmology. 2022;106(3):388-95.
[6] Wang X, Jiao B, Liu H, Wang Y, Hao X, Zhu Y, *et al.* Machine learning based on Optical Coherence Tomography images as a diagnostic tool for Alzheimer's disease. CNS Neuroscience & Therapeutics. 2022;28(12):2206-17.
[7] Corbin D, Lesage F. Assessment of the predictive potential of cognitive scores from retinal images and retinal fundus metadata via deep learning using the CLSA database. Scientific Reports. 2022;12(1):5767.
[8] Yousefzadeh N, Tran C, Ramirez-Zamora A, Chen J, Fang R, Thai MT. Neuron-level explainable AI for Alzheimer's Disease assessment from fundus images. Scientific Reports. 2024;14(1):7710.
[9] Kim HK, Lim YJ, Sunwoo MH, *et al.* Efficient Deep Retinal Fundus Image-Based Network for Alzheimer's Disease Diagnosis Using Mobile Device Applications. IEEE Access. 2024
[10] Liu Z, Wang Y, Vaidya S, Ruehle F, Halverson J, Soljaciˇ c M, *et al.* Kan: Kolmogorov-arnold networks.´ arXiv preprint arXiv:240419756. 2024.
[11] Bozorgasl Z, Chen H. Wav-kan: Wavelet kolmogorov-arnold networks. arXiv preprint arXiv:240512832. 2024.
[12] Cheon M. Demonstrating the efficacy of kolmogorov-arnold networks in vision tasks. arXiv preprint arXiv:240614916. 2024.
[13] Krizhevsky A, Hinton G, *et al.* Learning multiple layers of features from tiny images. 2009.
[14] Deng L. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. IEEE Signal Processing Magazine. 2012;29(6):141-2.
[15] Bissig D, Zhou C, Le V, Bernard J. A practical approach to functional optical coherence tomography shows abnormal retinal responses in Alzheimer's disease [Dataset]. Dryad.; 2020. https://doi.org/ 10.5061/dryad.msbcc2ftc.
[16] Ke X, Cao W, Lv F. Relationship between complexity and precision of convolutional neural networks. In: 2017 2nd International Symposium on Advances in Electrical, Electronics and Computer Engineering (ISAEECE 2017). Atlantis Press; 2017. p. 325-9.
[17] Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, *et al.* A foundation model for generalizable disease detection from retinal images. Nature. 2023;622(7981):156-63.
[18] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248-55.
[19] Lee CS, Apte RS. Retinal biomarkers of Alzheimer disease. American journal of ophthalmology. 2020;218:337-41