

ORIGINAL RESEARCH

Open access

Bridging Hand-Crafted Features and Deep Learning Through an Explainable Shallow CNN for Medical Image Segmentation

Sara Al-Fahd^{1*}, Noura Al-Khalifa¹, Omar Al-Turki²

Abstract

In clinical environments, the transparency of artificial intelligence models is paramount, as understanding the rationale behind automated choices is vital. Although black-box AI frameworks exhibit remarkable efficacy, their deployment in healthcare remains impractical unless their underlying mechanisms can be clarified for medical professionals. Deep neural networks (DNNs) achieve state-of-the-art results, yet the justifications they provide are frequently incomprehensible to human observers. Conversely, traditional machine learning approaches and manually designed features extracted to capture distinct data characteristics offer greater transparency, though human constraints in feature engineering often limit their overall effectiveness. To bridge this gap, we present ExShall-CNN, a novel and transparent shallow convolutional neural network tailored for medical image analytics. This framework builds upon hand-crafted feature methodologies to preserve human interpretability, guaranteeing clear and verifiable decision-making. We introduce the explainable shallow convolutional neural network (ExShall-CNN), which merges the high performance of sophisticated deep architectures like U-Net with the inherent clarity of hand-crafted features for segmenting medical images. Grounded in contemporary machine learning progress, ExShall-CNN integrates widely utilized kernels while maintaining structural openness, allowing medical practitioners to visually inspect and comprehend its outputs. This balanced strategy delivers the precision expected of deep learning paradigms alongside the explainability required for clinical deployment.

Keywords Explainability, Image segmentation, Shallow convolutional neural network

*Correspondence:

Sara Al-Fahd

s.alfahd@ksu.edu.sa

¹ Department of Medical Image Segmentation and Explainable Models, College of Medicine, King Saud University, Riyadh, Saudi Arabia

² Department of Shallow CNNs and Feature Engineering, College of Engineering, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

Introduction

At its core, a machine learning model serves as a function that maps input variables to target outputs. This mapping must demonstrate reliability and generalization when applied to unobserved data instances. Consequently, it cannot function as a rigid lookup table; instead, output values must be algorithmically derived from the inputs. Machine learning systems are therefore typically formulated as mathematical functions [1].

The design of these systems involves two primary phases: selecting the appropriate model architecture and optimizing it through training. Neural networks represent a prominent class of frameworks that have proven highly effective at managing both stages [2].

Theoretically, shallow neural networks—such as a multilayer perceptron (MLP) consisting of two layers equipped with activation functions—possess the capability to approximate any arbitrary input-output relationship or mapping, provided they contain sufficient non-linearity and adjustable parameters [3].

Research indicates that deep learning architectures demand significantly fewer trainable parameters to achieve performance levels comparable to expansive shallow networks [4]. Minimizing the parameter count accelerates optimization and enhances trainability, allowing the system to locate superior local optima within a constrained computational window. Deep learning models leverage a diverse array of structural components and architectures, including convolutional neural networks (CNNs) [2], recurrent neural networks (RNNs) [5], and transformer neural networks (TNNs) [6]. Both CNNs and TNNs are heavily utilized in computer vision and image analysis; the choice between them often depends on the scale of the dataset and the model capacity. Generally, TNNs excel when applied to massive datasets, whereas CNNs exhibit superior efficacy on smaller data collections [7, 8].

The domain of medical image processing spans numerous clinical specialties, including radiology and pathology. However, despite the immense possibilities within this discipline, only a constrained subset of these applications benefits from the availability of massive, meticulously curated datasets [9]. For this reason, while transformer architectures demonstrate minor advantages on large-scale cohorts, deep CNNs remain preferred for medical image analysis due to their robust performance on smaller data repositories [7, 10].

Two prominent deep CNN architectures utilized for image segmentation tasks are fully convolutional neural networks (FCNNs) [11] and U-Nets [12]. Nonetheless, because FCNNs and U-Nets are deep architectures, their internal mathematical operations cannot be straightforwardly explained [13]. While several approaches outlined in [14–18] are available to introduce explainability, very few have gained widespread acceptance in clinical practice, primarily because the direct visual representation of their internal features remains uninterpretable to humans [19]. This opacity presents significant hurdles in medical environments where the capacity to verify and audit an algorithm's decision-making pathways is imperative. Transparent models mitigate this issue by exposing their internal reasoning, thereby empowering clinicians to cross-examine and validate AI-generated predictions using their expert medical knowledge [20].

In summary, the primary innovations presented in this study are detailed below:

- We introduce an explainable shallow convolutional neural network (ExShall-CNN) that achieves segmentation accuracy comparable to benchmark architectures like U-Net in medical image tasks.
- ExShall-CNN is designed based on sequential developments in machine learning for image segmentation, enabling it to encapsulate the most prominent kernel configurations.
- The internal features generated by ExShall-CNN are transparent, allowing medical professionals and clinicians to evaluate them through direct visual inspection.

In the subsequent section, we provide an in-depth analysis of manually designed, human-interpretable features. Concurrently, we construct a chronological theoretical framework from first principles to integrate the most fundamental hand-crafted and machine-learning-driven features into a transparent, shallow convolutional network structure. The ExShall-CNN framework and its source code are openly accessible at <https://github.com/MLBC-lab/ExShall-CNN> (Access date: 1 February 2025).

Theory and Calculation

Otsu [21] introduced a global thresholding methodology that minimizes intra-class variance, an operation mathematically equivalent to maximizing the variance between the background and foreground classes for image segmentation.

$$T_{otsu} = \underset{t}{\operatorname{argmin}} \left\{ P_t(PL - 1 - P(t)) \left(\frac{\sum_{i=0}^{t-1} iP(i)}{p(t)} - \frac{\sum_{i=t}^{L-1} iP(i)}{P(L-1) - P(t)} \right) \right\} \quad (1)$$

In this formulation, L represents the span of the pixel intensity scale (e.g., 256 discrete levels for standard grayscale imagery), while p and P denote the probability density and cumulative probability functions, respectively. The specific intensity value t that minimizes the expression within the brackets defines the global Otsu threshold, T_{otsu} .

This optimization problem is typically resolved through iterative computation. The execution is generally rapid even when processing high-resolution imagery, as it relies solely on the frequency distribution of color tones or grayscale shades to derive the threshold. While computationally efficient, this technique establishes a uniform threshold across the entire image landscape, completely disregarding the spatial organization and context of pixel intensities. Consequently, it can fail to account for vital regional variations and contextual details across different image zones, which may compromise segmentation fidelity and analytical precision. By ignoring the spatial distribution of pixel intensities, this approach often falls short of capturing the fine details required for complex computer vision operations. To resolve the discrepancy between regional and global intensity distributions, Sauvola [22] introduced an alternative localized thresholding strategy:

$$T_{\text{sauvola}} = \mu_n \left(1 + k \left(\frac{\sigma_n}{r} - 1 \right) \right) \quad (2)$$

where n signifies the diameter of the localized neighborhood, while μ_n and σ_n represent the local mean and standard deviation computed within that neighborhood, respectively. The terms k and r are application-specific empirical coefficients. A standard convention sets the value of r to equal half of the maximum possible pixel intensity. Consequently, achieving optimal performance relies on tuning the two parameters n (the neighborhood diameter) and k . When ground-truth segmentation masks are accessible, these two parameters can be adaptively optimized according to the specific characteristics of the dataset.

It can be deduced that the local mean (μ_n) and local standard deviation (σ_n) function as regional descriptors, where Eq. 2 serves as the dividing decision boundary. Thus, the performance benefits of the Sauvola technique over the Otsu approach underline that regional descriptors are highly effective at isolating foreground elements from the background in classification problems. As a result, a unified methodology can center on the extraction of localized descriptors followed by a classification step. The fundamental raw inputs available consist solely of individual pixel values and their spatial coordinates, meaning all secondary descriptors must be constructed from this baseline information. The mean and standard deviation represent primary examples of descriptors derived from raw data, which can be expanded into more sophisticated variations.

Raw inputs originate from two distinct points of reference: the individual pixel intensities themselves and the intensity values of neighboring pixels. Descriptors can be characterized as any mathematical transformation that maps these raw inputs into an alternative domain to facilitate easier classification, such as linear transformations. With this principle in mind, a variety of spatial mappings have been developed, collectively known as kernel methods [23]. A selection of frequently utilized kernels is summarized in Table 1 [24].

Table 1. Common kernel mappings; x_i is an RGB vector of the current or neighbors' pixels.

Kernel Name	Equation
Linear	$X_1^T X_2$
Cosine	$\frac{X_1 X_2^T}{\ x_1\ \ x_2\ }$

Polynomial	$(\gamma X_1^T X_2 + C_0)^d$
Sigmoid	$\tanh(\gamma X_1^T X_2 + C_0)$
Radial Basis Function (RBF)	$\exp\left(-\gamma \left\ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\ ^2\right)$
Laplacian	$\exp\left(-\gamma \left\ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\ _1\right)$
Chi-Squared	$\exp\left(-\gamma \sum_i \frac{\left(\frac{x_1[i] - x_2[i]}{x_1[i] + x_2[i]}\right)^2}{2}\right)$

Kernel methodologies fully encompass both the Otsu and Sauvola techniques, given that both approaches rely on two primary descriptors: the mean and the standard deviation. These descriptors correspond to linear or second-order combinations of adjacent pixel intensities. Consequently, if a more advanced framework integrates all the specified kernel transformations, it natively covers the functionalities of the Otsu and Sauvola methods as well. Every kernel comprises basic mathematical operations, such as addition, subtraction, multiplication, division, power, and exponentiation.

The Otsu, Sauvola, and kernel-based frameworks are notable for their high degree of interpretability, particularly through visual representation. Under these approaches, the resulting classifications correspond directly to visual content. Furthermore, they feature mathematical transparency, explicitly demonstrating the structural interactions between neighboring pixels throughout the classification procedure.

We will demonstrate that the Otsu, Sauvola, and kernel-driven approaches can be unified within a shallow convolutional neural network when equipped with appropriate activation functions. A CNN performs a weighted summation of raw data, while specialized activation functions enable the model to emulate kernel transformations. To integrate the core mathematical operations—namely addition/subtraction, multiplication/division, power, and exponentiation—we introduce the following activation functions:

$$e(x) = \exp(k^T x) \quad (3)$$

$$l(x) = \log(k^T x) \quad (4)$$

where x represents the input vector and k denotes the convolution kernel vector. Given that spatial convolution can be interpreted as a cross-correlation operation, the two vectors undergo element-wise multiplication followed by a summation. The exponential and logarithmic transformations are applied element-wise across mini-batches. Consequently, the length of both vectors corresponds directly to the kernel dimensions. To demonstrate that every fundamental mathematical operation can be executed using these activation functions, a basic two-dimensional input is utilized as an illustrative baseline. Incorporating additional dimensions serves merely as a direct extension of this simplified case. The formulations presented in **Table 2** are integrated into a shallow convolutional neural network architecture.

Table 2. Feasibility of required operations with proposed activation functions.

Weighted Operation	Induction	Implementation
Summation/Subtraction	$k_1 x_1 + k_2 x_2$	$k_1 x_1 + k_2 x_2$

Multiplication	$k_{1 \times 1} * k_{2 \times 2}$	$\exp(\log(k_{1 \times 1}) + \log(k_{2 \times 2}))$
Division	$k_{1 \times 1} / k_{2 \times 2}$	$\exp(\log(k_{1 \times 1}) - \log(k_{2 \times 2}))$
Power	$(k_{1 \times 1})^{k_{2 \times 2}}$	$\exp(\log(k_{1 \times 1}) + \log(\log(k_{2 \times 2})))$

As demonstrated in [Table 2](#), all necessary mathematical operations can be successfully executed using shallow convolutional neural networks combined with the proposed activation functions, provided sufficient residual connections are integrated. Residual mappings, which represent a major milestone in CNN design, were originally introduced in [\[25\]](#). Since their inception, they have been widely integrated into virtually all high-performing CNN architectures. It is straightforward to conceptualize that by appropriately pairing and configuring these operational units, every kernel detail listed in [Table 1](#) can be successfully recreated, along with numerous highly complex interactions that have yet to be formalized into standard kernel categories. While the exponentiation of a logarithm functions as an identity mapping, introducing distinct coefficients allows these components to construct highly non-linear functions. A critical constraint in these formulations (particularly regarding logarithmic transformations) is that negative inputs do not possess real-valued logarithms. To overcome this limitation, we engineered complex convolutional neural networks (CCNNs) rather than relying on standard, real-valued CNN architectures. From a theoretical standpoint, the imaginary components derived from the expressions in [Table 2](#) should equal zero when the input values consist exclusively of real numbers. Because our framework utilizes normalized, real-valued pixel intensities, the imaginary components can be safely discarded following the terminal activation function of each module. Although the imaginary elements will not resolve exactly to zero due to floating-point computational errors, their magnitude remains negligible. Operating within the complex number domain is essential because intermediate algebraic steps frequently generate substantial imaginary components.

Furthermore, it is well established that the receptive fields of CNN architectures expand primarily as a function of network depth [\[26, 27\]](#). Because our approach prioritizes a shallow CNN structure to preserve superior explainability, we cannot achieve expansive receptive fields via deep layering; instead, this limitation must be offset by increasing the spatial dimensions of the convolution kernels. In deep CNN configurations, kernel dimensions of 3, 5, and occasionally 7 are standard. Assuming a center-aligned kernel configuration, an expansive kernel dimension of 7 incorporates two adjacent pixels in every spatial direction into the calculation. Across numerous sequential layers in a deep CNN, this receptive field expands substantially, scaling proportionally with the kernel dimensions, stride, and dilation factors of the hidden layers. In a shallow CNN architecture, this depth-driven expansion of the receptive field is absent; thus, a diverse selection of kernel dimensions and dilation rates must bear the responsibility of capturing receptive fields comparable to those found in deep networks. Larger kernel sizes cause non-linear formulations to converge at a slower rate during backpropagation. Consequently, they typically demand extended optimization periods, even though their total parameter footprint is often substantially smaller than that of deep CNN configurations.

Materials and Methods

This research utilizes the Retina Blood Vessel segmentation [\[28\]](#) and the International Skin Imaging Collaboration (ISIC) [\[29-31\]](#) datasets, which represent two benchmark reference collections for evaluating medical image segmentation algorithms. The Retina Blood Vessel dataset comprises 100 color images alongside corresponding binary ground-truth masks. This collection is partitioned into two distinct subsets: 80 images paired with their respective masks are allocated for model training, while the remaining 20 images and masks are set aside for evaluation. This dataset provides a highly curated repository of retinal fundus photographs, with each sample containing highly accurate annotations of vascular structures. Precise segmentation of the retinal vasculature is vital in clinical ophthalmology, as it supports the early detection and management of ocular diseases, such as diabetic retinopathy and macular degeneration [\[32\]](#). The

combined ISIC 2016, 2017, and 2018 datasets encompass 7,723 images along with their associated binary target masks. The data from each of these three annual iterations includes separate partitions for model training, validation, and testing. From this total repository, we randomly extracted 100 images across the annual cohorts, assigning 60 samples from the training subsets, 20 from the validation subsets, and 20 from the testing subsets.

The architecture of our proposed shallow CNN framework is illustrated in **Figure 1**. As detailed in the schematic, the network incorporates three distinct structural module variations: Conv, Log-Conv-Exp (LCE), and Conv-Log-Conv-Exp (CLCE), with each module implemented across multiple kernel sizes. The standard Conv layer calculates a weighted summation of adjacent pixel values, emulating the behavior of a linear kernel (**Table 2**, row 1). The LCE block computes products and quotients of neighboring pixel values, effectively replicating cosine, polynomial, and certain sigmoid kernel operations (**Table 2**, rows 2 and 3). Finally, the CLCE block approximates the mathematical characteristics of RBF, Laplacian, and Chi-Squared kernels (**Table 2**, row 4). Because the weight matrices and bias vectors of the convolutional operations can take on either positive or negative values, the resulting logarithmic outputs yield complex numbers. Consequently, the network is executed entirely within the complex number domain, with the sole exception of the terminal aggregation layer, which collects and combines real numbers. In contrast to the other processing blocks, the Conv and Aggregate operations do not utilize non-linear activation functions. In this design, we integrate kernel sizes of 1, 3, 5, 9, 13, 17, 21, and 25. This configuration yields a total of $8 \times 3 = 24$ parallel processing modules.

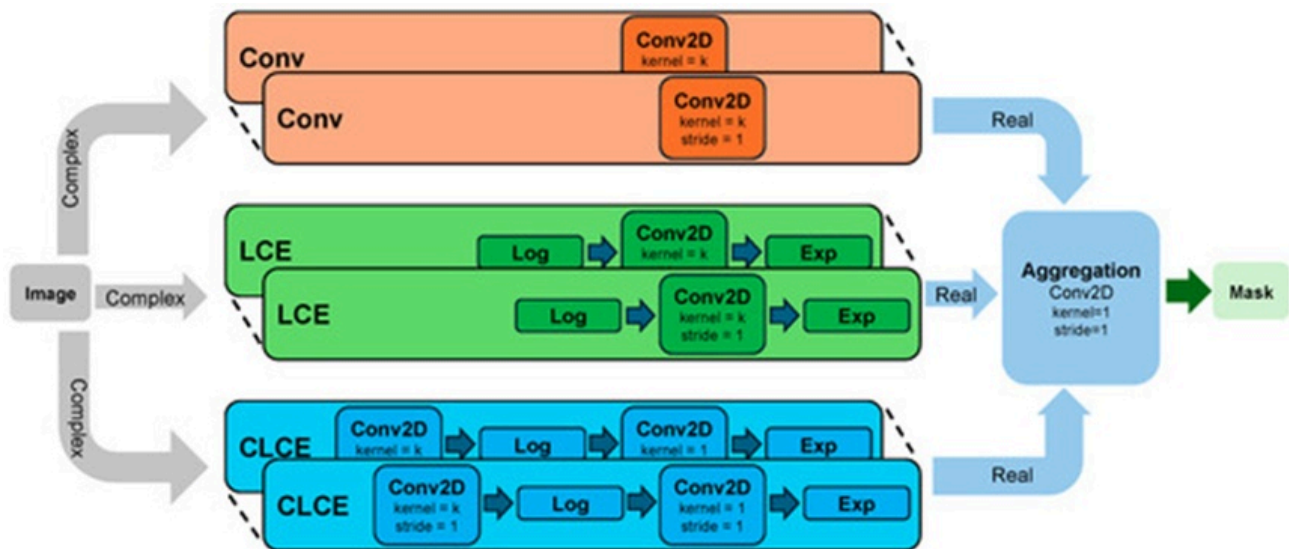


Figure 1. ExShall-CNN with 24 parallel complex modules, outputting real aggregation.

The software implementation is constructed within the Python 3.10 and PyTorch 2.0 [33] programming environments. A uniform batch size of 4 is maintained across all evaluated frameworks and datasets. The Adam optimization algorithm [34] was evaluated using a range of learning rates (1×10^{-2} , 1×10^{-3} , 1×10^{-4} , 1×10^{-5}) for all architectural configurations. The optimal performance configurations achieved by each network configuration on each dataset are selected and documented.

We benchmark our explainable shallow CNN against two prominent baselines in the field of image segmentation: the FCNN [11] and the U-Net [12]. The FCNN framework features a structural layout similar to our proposed model but incorporates significantly greater depth, whereas the U-Net architecture serves as a primary state-of-the-art reference model for medical image segmentation tasks. Both baseline models differ from our proposed shallow CNN in two critical architectural aspects: both are deep neural networks utilizing multiple hidden layers, and both rely exclusively on standard ReLU activation functions.

All three architectural frameworks are trained and validated using identical training data partitions and subsequently appraised on the independent testing sets. The objective function utilized during optimization is the negative logarithm of the Jaccard index, minimized via the Adam optimizer configured with a baseline learning rate of 1×10^{-4} .

Results and Discussion

This section outlines our empirical findings and contextualizes them against benchmark deep learning architectures, followed by a detailed breakdown of our system's explainability properties.

Evaluation and comparative analysis

To quantitatively gauge segmentation fidelity, we leverage the Jaccard similarity index alongside Sørensen–Dice parameters. Given the binary classification framework of this task (distinguishing exclusively between foreground targets and background elements), the computed Dice metrics align precisely with standard F1 performance scores. Quantitative profiles for the three evaluated networks are organized within [Table 3](#) and [Table 4](#). Our analytical focus centers primarily on positioning our findings against state-of-the-art deep architectures. Documentation hosted on the Kaggle platform [\[28\]](#) establishes that these deep learning frameworks drastically outpace classical machine learning pipelines. Nevertheless, such setups typically employ massive structural depth, rendering them entirely opaque. This investigation explicitly confronts this limitation by engineering a streamlined, interpretable deep learning framework.

During training phases, the proposed Shallow-CNN registers lower accuracy metrics than both FCNN and U-Net variants. Conversely, when deploying the models against novel validation sets, Shallow-CNN outperforms the FCNN pipeline while exhibiting parity with the U-Net architecture. We attribute the superior generalization capability of Shallow-CNN over FCNN to its radically constrained parameter footprint, an architectural property that inherently minimizes overfitting hazards. For context, a compressed U-Net variant utilizing 0.36×10^6 parameters (designated as U-Net-G) tested on an equivalent dataset achieved a DICE metric of 0.88, as documented in [\[35\]](#). While the absolute performance of the U-Net remains superior to ExShall-CNN, one must emphasize that standard U-Net topologies require a substantially larger parameter allocation—exceeding our proposed lightweight model by at least an order of magnitude. By capitalizing on hand-crafted feature paradigms and explicitly interpretable data transformations, Shallow-CNN successfully strikes an equilibrium between structural complexity and expressive feature representation, yielding robust behavior on unobserved data profiles. These indicators clarify that while Shallow-CNN does not surpass U-Net benchmarks, it provides far better generalization than an FCNN. Furthermore, whereas deep models like FCNN and U-Net create a barrier to interpretability, this shallow network provides explicit structural transparency, a characteristic parsed thoroughly in the subsequent subsection.

Table 3. Performance of the three models on Retina Blood Vessel.

Network Architecture	Number of Parameters	Retina Blood Vessel Segmentation			
		Test Set	Test Set	Training Set	Training Set
		DICE (%)	Jaccard (%)	DICE (%)	Jaccard (%)
Fully Convolutional Neural Network (Fully CNN)	54,304,086	72.4	56.8	77.6	63.3
U-Net Architecture	31,043,521	79.0	65.9	77.7	63.5
Lightweight CNN	39,698	73.6	58.3	72.0	56.2

Table 4. Performance of the three models on International Skin Imaging Collaboration (ISIC).

Network Architecture	Number of Parameters	ISIC 2016, 2017, 2018 Dataset			
		Test Set	Test Set	Training Set	Training Set
		DICE (%)	Jaccard (%)	DICE (%)	Jaccard (%)
Fully Convolutional Neural Network (Fully CNN)	54,304,086	77.9	63.8	90.6	82.8
U-Net Architecture	31,043,521	81.9	69.4	71.2	56.0
Lightweight CNN	39,698	81.3	68.5	62.5	45.5

A fundamental divergence separating shallow network designs from deep convolutional networks concerns the scale of their spatial receptive fields. A prevailing principle in computer vision dictates that stacking multiple hidden layers automatically yields wider receptive fields than those achievable by compact networks. To counter this limitation within ExShall-CNN, we artificially extended our receptive field profiles by integrating expanded spatial dimensions. Initially, the curse of dimensionality obstructed the framework's capability to accurately map latent spatial patterns, which compromised initial test runs. In practice, however, configuring kernel dimensions to values of 21 and 25 equipped ExShall-CNN with a sufficiently vast receptive field, directly counteracting the constraints of a shallow architecture and ultimately elevating overall segmentation performance.

The evaluation and cross-examination of these three specific setups involve certain constraints. In this study, our appraisals are restricted to a duo of datasets featuring relatively modest sample distributions. A broader empirical investigation drawing from expanded image repositories would yield more definitive insights into the operational discrepancies among these architectures. Additionally, the Otsu, Sauvola, and kernel-based operations are intrinsically manually designed feature extraction pipelines. While we acknowledge that representations learned dynamically by deep networks frequently deliver maximum utility, our primary motive was to analyze the visual properties of these hand-crafted extraction behaviors. To manifest this, we engineered ExShall-CNN to synthesize features matching the characteristics of hand-crafted operations while capturing the performance advantages inherent to deep learning feature optimization. Consequently, we targeted our benchmarking strategies toward its primary algorithmic alternatives—namely, deep learning networks—rather than historical, standalone hand-crafted extraction techniques.

Explainability

In direct contrast to typical deep learning setups, compact, shallow network structures possess organic explainability properties [36]. This interpretability clarifies the underlying rationale guiding a model's classifications, serving as a pillar for establishing clinical trust in automated diagnostic assistance [16]. Because deep learning architectures rely on highly intricate, multi-layered hierarchies, identifying the specific hidden layer that yields the most coherent visual explanation is immensely difficult. For instance, the authors in [16] thoroughly evaluate multiple distinct explainability frameworks. Visual transparency provides distinct benefits, including direct usability and immediate verification. While numerous methodologies can generate visual explanations, our architectural setup functions similarly to Class Activation Mapping (CAM) concepts [37]. However, whereas CAM yields a derivative, indirect trace of the collective influence of sequential

layers (which grow highly numerous in networks such as VGG [38]), our paradigm directly reflects the explicit diagnostic weight of each operational block due to the streamlined, shallow nature of the architecture.

We decipher the inner workings of the proposed Shallow-CNN by isolating the most critical processing modules outlined in **Figure 1** using data from the Retina Blood Vessel cohort. Each individual module behaves as an optimized mathematical transformation (kernel) tailored to the training data. We implement a model-driven ablation analysis to isolate the exact performance impact of each individual module across the global architecture [39]. As illustrated in the structural diagram of **Figure 1**, our network integrates 24 distinct modules. During each sequential trial, we systematically extract a single module from the active pipeline and recompute performance metrics, such as the total Dice score. If a specific module serves a critical diagnostic role, its removal triggers a sharp decline in overall accuracy. We utilize the following mathematical scoring function to quantify the absolute operational impact of each individual module:

$$Impact_i = \frac{DICE_{max} - DICE_i}{DICE_{max}} = 1 - \frac{DICE_i}{DICE_{max}} \quad (5)$$

where $DICE_{max}$ represents the peak accuracy achieved by the system when utilizing all modules simultaneously, $DICE_i$ indicates the accuracy score recorded after isolating and removing module i , and $Impact_i$ defines the finalized contribution weight of that specific module. The term $Impact_i$ yields a bounded metric between 0 and 1, where a value of 0 denotes zero operational contribution and 1 represents the absolute highest tier of importance. To ensure the integrity of **Eq. 5**, it is imperative that the network is never retrained following the extraction of a layer; instead, it must be evaluated strictly using its pre-optimized parameters. **Eq. 5** offers an authentic, logical breakdown of module significance exclusively while the network remains shallow. Consequently, this diagnostic formulation is valid for compact topologies but cannot be applied to deep systems. The target of **Eq. 5** is to fully deconstruct the final Aggregation module and expose its underlying mechanics. Throughout the optimization phase, the Aggregation layer automatically assigns higher scaling coefficients to the most expressive features derived from the internal processing modules. As a result, dropping these vital modules out of the loop heavily degrades the performance of ExShall-CNN, an effect precisely captured by **Eq. 5**. It is crucial to highlight that the Aggregation mechanism determines its output for each discrete pixel by executing a weighted summation of corresponding spatial pixels across all active modules at that identical coordinate, because the spatial kernel size of the Aggregation layer is restricted to 1. The resulting impact distribution for Shallow-CNN has been mapped and is illustrated in **Figure 2**. As shown in the graph, specific processing modules—most notably 1, 9, and 15—exhibit the highest overall impact, whereas alternative blocks like 2, 5, and 12 register negligible influence. This demonstrates that by preserving only the highly influential modules and pruning the low-impact blocks, we can maintain performance limits that closely match the complete Shallow-CNN baseline.

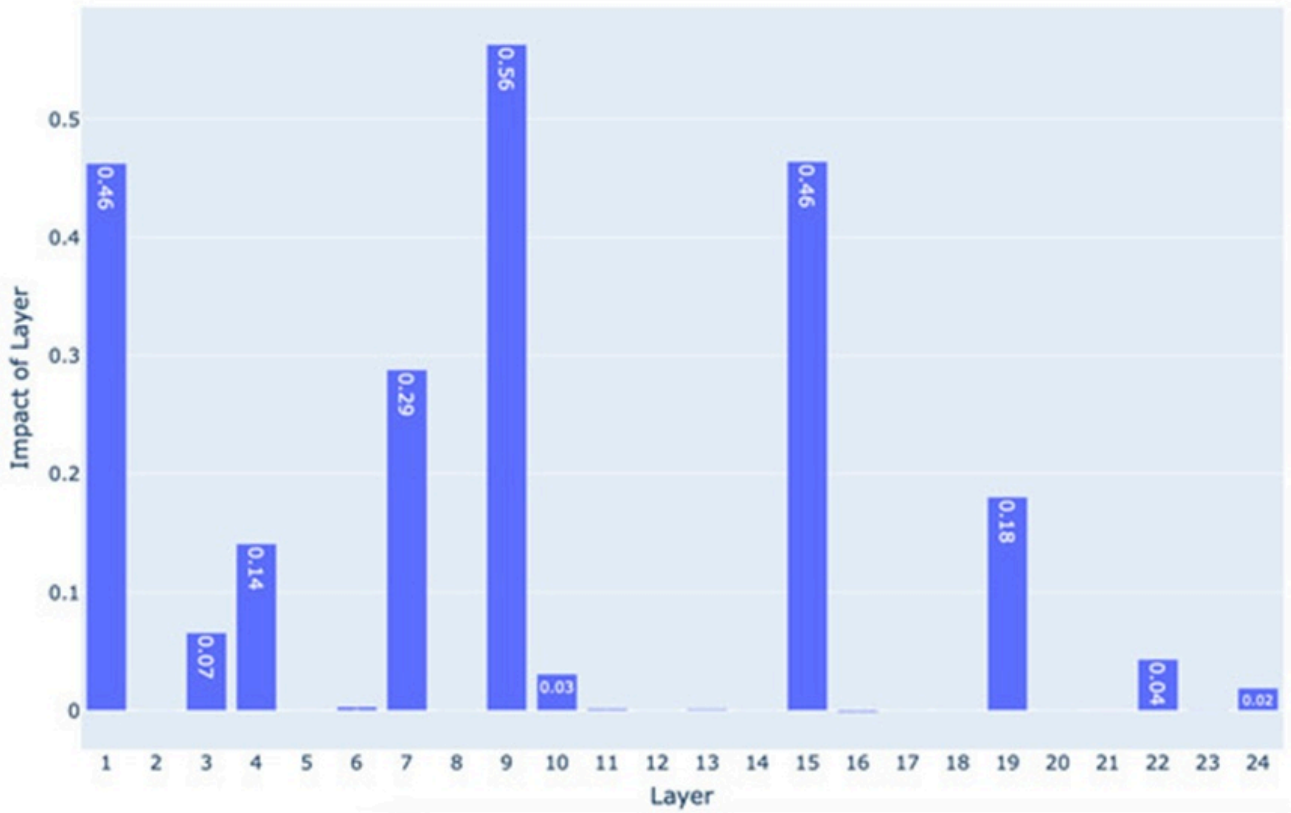


Figure 2. Comparison of layer impact on the performance of the Shallow-CNN. The higher values show more impact. Layers 9, 1, 15, 7, and 19 have the most impactful role in the decisions, respectively.

We flag specific modules that achieve an impact score scaling beyond 80\% of the interquartile range calculated across all 24 baseline impact scores, resulting in the selection of modules 1, 7, 9, 15, and 19 as detailed in Table 5. The exact image transformations produced by these isolated modules are visualized in Figure 3. As documented in Figure 3, while the mathematical mapping of kernel 1 remains strictly linear, the remaining operational kernels produce a highly non-linear response to variations in input color intensity. A comparative inspection of background elements and foreground structures between the raw input and the transformed variants reinforces this behavior.

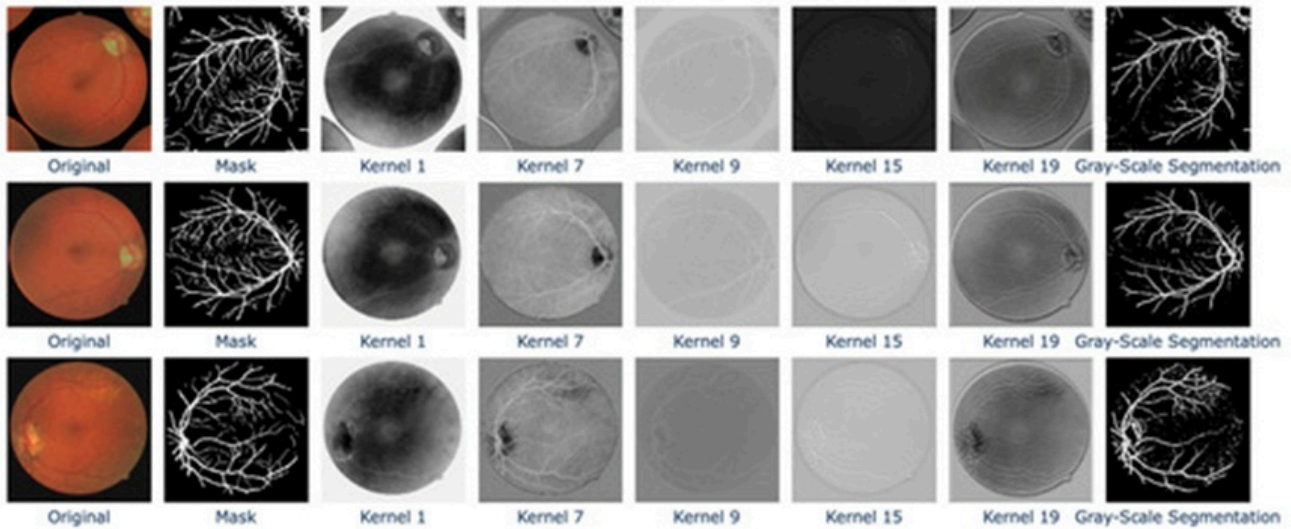


Figure 3. Transformations of the most important kernels on the original image. The mask serves as the ground truth, while the grayscale segmentation represents the model's output. Visual outputs from the most important layers—1, 7, 9, 15, and 19—are displayed. These outputs help to clarify the final decisions made by the ExShall-CNN model.

Table 5. Characteristics of the most impactful modules.

Kernel Index	Module Type	Kernel Size
1	Conv	1
7	Conv	21
9	LCE	1
15	LCE	17
19	CLCE	3

The interpretability of this framework is fundamentally realized through explicit visual data representation, as demonstrated by the case study presented in [Figure 3](#). Our central motivation for engineering these customized activation formulas ([Eqs. 3 and 4](#)) was to enhance visual transparency while simultaneously instantiating classic, well-established kernel models. Standard deep learning architectures typically consist of dozens of stacked layers, obscuring exactly which operational stage offers the most reliable visual justification.

Exposing these transformed internal images significantly strengthens the clinical reliability of otherwise opaque AI models within medical environments. This transparency allows medical professionals to scrutinize the foundational reasons behind automated choices, allowing them to determine if the internal mappings and ultimate clinical outputs are valid. For example, clinicians and radiologists frequently track specific statistical markers, such as tissue variance, to diagnose localized abnormalities within organs like the liver or lungs. Consequently, computational frameworks that base their diagnostic outputs on these identical transparent metrics naturally gain clinical trust. The visually transformed output explicitly maps out which spatial details of the raw source medical image exerted the greatest leverage over the model's ultimate clinical prediction.

Conclusion

Integrating artificial intelligence frameworks within clinical environments demands rigorous scrutiny, with a primary emphasis on model explainability. Although deep neural networks have established remarkable diagnostic capabilities that frequently match or exceed human performance, they fundamentally operate as black-box systems lacking necessary structural openness. Throughout the classical machine learning era, investigators prioritized human-interpretable feature engineering, which explicitly clarified the underlying logic of a model's predictions and removed algorithmic opacity. Nevertheless, the predictive accuracy of these manually designed descriptors proved insufficient because they failed to fully capitalize on optimization methodologies driven by raw data.

This research explored the historical progression of hand-crafted feature extraction tailored for spatial image segmentation. We demonstrated that an optimally engineered, explainable shallow convolutional neural network (ExShall-CNN) can match the performance thresholds of deep CNN architectures while simultaneously delivering vastly superior interpretability. ExShall-CNN is structured as a data-driven evolution of classical kernel methods, capable of orchestrating intricate non-linear mappings and encompassing the most prominent kernel archetypes. Regarding its clinical utility, we contend that our proposed framework can serve as an effective substitute for standard deep learning models in medical image computing, particularly in scenarios where both interpretability and structural simplicity are paramount.

In summary, this paper introduces the explainable shallow convolutional neural network (ExShall-CNN), establishing a viable equilibrium between predictive accuracy and transparency for medical image segmentation. While it does not surpass the U-Net benchmark, ExShall-CNN exhibits superior generalization compared to fully convolutional neural networks (FCNN)—particularly when confronting unobserved data instances—owing to its tightly constrained parameter footprint and minimized susceptibility to overfitting. By synthesizing manually designed descriptors with transparent architectural transformations, the model delivers the clarity and interpretability critical for clinical integration. Although this study is constrained by localized sample volumes and a fundamental reliance on hand-crafted feature paradigms, the empirical outcomes underscore the viability of shallow configurations for transparent, high-fidelity medical image segmentation.

Moving forward, we intend to advance the capabilities of ExShall-CNN to accommodate multi-resolution image analysis, thereby mitigating the dimensionality constraints tied to expansive kernel sizes. We also plan to rigorously investigate both the theoretical properties and the empirical boundaries of the proposed network. This pursuit will encompass standardized mathematical analysis alongside an investigation of real-world variables, including structural pattern complexity and database scale. Furthermore, we intend to appraise the performance of ExShall-CNN across a more diverse spectrum of biomedical imaging modalities, allowing for subsequent optimization and architectural refinement.

While the introductory findings established in this article provide a robust baseline, we acknowledge that comprehensive validation across a more heterogenous collection of datasets will strengthen our conclusions. Given the designated scope of the current inquiry, our focus was directed toward baseline experimentation, though we are actively scaling our validation pipeline. Our future research will focus on executing a more extensive evaluation involving larger, more varied imaging repositories.

The ExShall-CNN architecture and its complete source code are openly accessible at <https://github.com/MLBC-lab/ExShall-CNN> (Access Date: 1 February 2025).

Abbreviations

The following abbreviations are used in this manuscript:

DNN	Deep neural network
CNN	Convolutional neural network
RNN	Recurrent neural network
TNN	Transformer neural network
FCNN	Fully convolutional neural network
MLP	Multilayer perceptron

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 31 Aug 2021

Revised: 26 Sep 2021

Accepted: 10 Nov 2021

Published online: 15 March 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Murphy KP. Machine learning: a probabilistic perspective. Cambridge (MA): MIT Press; 2012.

LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-44.
<https://doi.org/10.1038/nature14539>.

Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Netw*. 1989;2(5):359-66.
[https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8).

Mhaskar H, Liao Q, Poggio T. When and why are deep networks better than shallow ones? In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*; 2017; San Francisco, CA, USA. p. 2343-9.
<https://doi.org/10.1609/aaai.v31i1.10913>.

Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, editors. *Parallel distributed processing: explorations in the microstructure of cognition*. Vol. 1. Cambridge (MA): MIT Press; 1986. p. 318-62.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; 2017; Long Beach, CA, USA. p. 5998-6008.

Murphy ZR, Venkatesh K, Sulam J, Yi PH. Visual transformers and convolutional neural networks for disease classification on radiographs: a comparison of performance, sample efficiency, and hidden stratification. *Radiol Artif Intell*. 2022;4(6):e220012.
<https://doi.org/10.1148/ryai.220012>.

Shamshirband S, Fathi M, Dehzangi A, Chronopoulos AT, Alinejad-Rokny H. A review on deep learning approaches in healthcare systems: taxonomies, challenges, and open issues. *J Biomed Inform*. 2021;113:103627.
<https://doi.org/10.1016/j.jbi.2020.103627>.

Guan H, Yap PT, Bozoki A, Liu M. Federated learning for medical image analysis: a survey. *Pattern Recognit*. 2024;151:110424.

Khan MSI, Rahman A, Debnath T, Karim MR, Nasir MK, Band SS, et al. Accurate brain tumor detection using deep convolutional neural network. *Comput Struct Biotechnol J*. 2022;20:4733-45.
<https://doi.org/10.1016/j.csbj.2022.08.039>.

Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2015; Boston, MA, USA. p. 3431-40.
<https://doi.org/10.1109/CVPR.2015.7298965>.

Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical image computing and computer-assisted intervention—MICCAI 2015*. Cham: Springer; 2015. p. 234-41.
https://doi.org/10.1007/978-3-319-24574-4_28.

Ibrahim R, Shafiq MO. Explainable convolutional neural networks: a taxonomy, review, and future directions. *ACM Comput Surv*. 2023;55.

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.
<https://doi.org/10.1016/j.inffus.2019.12.012>.

Band SS, Yarahmadi A, Hsu CC, Biyari M, Sookhak M, Ameri R, et al. Application of explainable artificial intelligence in medical health: a systematic review of interpretability methods. *Inform Med Unlocked*. 2023;40:101286.

van der Velden BHM, Kuijff HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med Image Anal*. 2022;79:102470.
<https://doi.org/10.1016/j.media.2022.102470>.

Zhang Q, Wu YN, Zhu SC. Interpretable convolutional neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2018; Salt Lake City, UT, USA. p. 8827-36.

Huang S, Xu Z, Tao D, Zhang Y. Part-stacked CNN for fine-grained visual categorization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016; Las Vegas, NV, USA. p. 1173-82.

Zhang Y, Tino P, Leonardis A, Tang K. A survey on neural network interpretability. *IEEE Trans Emerg Top Comput Intell*. 2021;5(5):726-42.

Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. *Comput Biol Med*. 2022;140:105111.
<https://doi.org/10.1016/j.combiomed.2021.105111>.

Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*. 1979;9(1):62-6.

Sauvola J, Pietikäinen M. Adaptive document image binarization. *Pattern Recognit*. 2000;33(2):225-36.
[https://doi.org/10.1016/S0031-3203\(99\)00055-2](https://doi.org/10.1016/S0031-3203(99)00055-2).

Schölkopf B, Smola AJ. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge (MA): MIT Press; 2001.

Zhang J, Marszałek M, Lazebnik S, Schmid C. Local features and kernels for classification of texture and object categories: a comprehensive study. *Int J Comput Vis*. 2007;73(2):213-38.
<https://doi.org/10.1007/s11263-006-9794-1>.

He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016; Las Vegas, NV, USA. p. 770-8.
<https://doi.org/10.1109/CVPR.2016.90>.

Araujo A, Norris W, Sim J. Computing receptive fields of convolutional neural networks. *Distill*. 2019;4:e21.
<https://doi.org/10.23915/distill.00021>.

Luo W, Li Y, Urtasun R, Zemel R. Understanding the effective receptive field in deep convolutional neural networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems; 2016; Barcelona, Spain. p. 4905-13.

Retina Blood Vessel [dataset]. Kaggle; 2024 [cited 2024 Jul 1].

Gutman D, Codella NCF, Celebi E, Helba B, Marchetti M, Mishra N, et al. Skin lesion analysis toward melanoma detection: a challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *arXiv [Preprint]*. 2016. arXiv:1605.01397.

Codella NCF, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). In: Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 2018; Washington, DC, USA. p. 168-72.

Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, Gutman D, et al. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv [Preprint]*. 2019. arXiv:1902.03368.

Cervantes J, Cervantes J, García-Lamont F, Yee-Rendon A, Cabrera JE, Jalili LD. A comprehensive survey on segmentation techniques for retinal vessel segmentation. *Neurocomputing*. 2023;556:126626.

Paszke A, Gross S, Massa F, Lerer A, Bradbury A, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. *Advances in Neural Information Processing Systems 32*. Red Hook (NY): Curran Associates; 2019. p. 8024-35.

Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv [Preprint]*. 2014. arXiv:1412.6980.

Wang KYW, Hugonot J, Fua P, Salzmann M. Recurrent U-Net for resource-constrained segmentation. In: 2019 IEEE International Conference on Computer Vision Workshops (ICCVW); 2019.
<https://doi.org/10.1109/ICCVW.2019.00517>.

Marques dos Santos JD, Marques dos Santos JP. Towards XAI: interpretable shallow neural network used to model HCP's fMRI motor paradigm data. In: *Bioinformatics and Biomedical Engineering*. Cham: Springer; 2022. p. 260-74.

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016; Las Vegas, NV, USA. p. 2921-9.
<https://doi.org/10.1109/CVPR.2016.319>.

Liu S, Deng W. Very deep convolutional neural network based image classification using small training sample size. In: Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR); 2015; Kuala Lumpur, Malaysia. p. 730-4.

Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res*. 2019;20:1-81.