

Article

Explainable Deep Fuzzy Cognitive Map Diagnosis of Coronary Artery Disease: Integrating Myocardial Perfusion Imaging, Clinical Data, and Natural Language Insights

Anna Feleki ¹, Ioannis D. Apostolopoulos ¹, Serafeim Moustakidis ^{1,2,*}, Elpiniki I. Papageorgiou ¹, Nikolaos Papathanasiou ³, Dimitrios Apostolopoulos ³ and Nikolaos Papandrianos ¹

¹ Department of Energy Systems, University of Thessaly, Gaiopolis Campus, 41500 Larisa, Greece; annafele1@uth.gr (A.F.); apostolopoulos.john.dim@gmail.com (I.D.A.); elpinikipapageorgiou@uth.gr (E.I.P.); npapandrianos@uth.gr (N.P.)

² AIDEAS OÜ, Narva mnt 5, 10117 Tallinn, Estonia

³ Department of Nuclear Medicine, University Hospital of Patras, 26504 Rio, Greece; nikopapath@upatras.gr (N.P.); dimap@med.upatras.gr (D.A.)

* Correspondence: s.moustakidis@aideas.eu

Abstract: Myocardial Perfusion Imaging (MPI) has played a central role in the non-invasive identification of patients with Coronary Artery Disease (CAD). Clinical factors, such as recurrent diseases, predisposing factors, and diagnostic tests, also play a vital role. However, none of these factors offer a straightforward and reliable indication, making the diagnosis of CAD a non-trivial task for nuclear medicine experts. While Machine Learning (ML) and Deep Learning (DL) techniques have shown promise in this domain, their “black-box” nature remains a significant barrier to clinical adoption, a challenge that the existing literature has not yet fully addressed. This study introduces the Deep Fuzzy Cognitive Map (DeepFCM), a novel, transparent, and explainable model designed to diagnose CAD using imaging and clinical data. DeepFCM employs an inner Convolutional Neural Network (CNN) to classify MPI polar map images. The CNN’s prediction is combined with clinical data by the FCM-based classifier to reach an outcome regarding the presence of CAD. For the initialization of interconnections among DeepFCM concepts, expert knowledge is provided. Particle Swarm Optimization (PSO) is utilized to adjust the weight values to the correlated dataset and expert knowledge. The model’s key advantage lies in its explainability, provided through three main functionalities. First, DeepFCM integrates a Gradient Class Activation Mapping (Grad-CAM) algorithm to highlight significant regions on the polar maps. Second, DeepFCM discloses its internal weights and their impact on the diagnostic outcome. Third, the model employs the Generative Pre-trained Transformer (GPT) version 3.5 model to generate meaningful explanations for medical staff. Our dataset comprises 594 patients, who underwent invasive coronary angiography (ICA) at the department of Nuclear Medicine of the University Hospital of Patras in Greece. As far as the classification results are concerned, DeepFCM achieved an accuracy of 83.07%, a sensitivity of 86.21%, and a specificity of 79.99%. The explainability-enhancing methods were assessed by the medical experts on the authors’ team and are presented within. The proposed framework can have immediate application in daily routines and can also serve educational purposes.



Citation: Feleki, A.; Apostolopoulos, I.D.; Moustakidis, S.; Papageorgiou, E.I.; Papathanasiou, N.; Apostolopoulos, D.; Papandrianos, N. Explainable Deep Fuzzy Cognitive Map Diagnosis of Coronary Artery Disease: Integrating Myocardial Perfusion Imaging, Clinical Data, and Natural Language Insights. *Appl. Sci.* **2023**, *13*, 11953. <https://doi.org/10.3390/app132111953>

Academic Editor: Julio Garcia Flores

Received: 24 September 2023

Revised: 28 October 2023

Accepted: 30 October 2023

Published: 1 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: fuzzy cognitive maps; particle swarm optimization; convolutional neural networks; classification; feature selection; Grad-CAM; coronary artery disease; natural language processing

1. Introduction

1.1. Backdrop

In recent years, ML and DL techniques have achieved unprecedented success in diverse applications, ranging from image recognition to natural language processing (NLP). These powerful algorithms have the potential to revolutionize medical practice, driving

innovation in diagnostics, treatment, and patient care [1]. In particular, the intersection of ML and medicine holds great promise for improving the accuracy and efficiency of diagnosis in complex diseases, such as CAD [2], which is the main problem this paper deals with.

Despite the remarkable strides made by ML and DL in medical applications, their success comes with an inherent drawback—the “black-box” nature of these models [3]. Black-box models are characterized by their complex, high-dimensional architectures, where the relationship between inputs and outputs becomes increasingly inscrutable [4]. While these models demonstrate outstanding predictive capabilities, their lack of explainability poses significant challenges in critical domains, especially healthcare [5]. The opacity of black-box models not only hampers the acceptance of artificial intelligence (AI) solutions in medical practice but also raises concerns related to safety, ethics, and regulatory compliance [3]. Explainability, in the context of ML and DL, refers to the ability to comprehend and interpret the decision-making process of a model [6]. Traditionally, simpler models, such as linear regression and decision trees, offered transparent and understandable insights into their predictions [6]. However, the advent of more sophisticated algorithms, such as deep neural networks, has shifted the focus towards performance optimization at the expense of interpretability [7].

The field of medical diagnosis, and especially CAD diagnosis, heavily relies on the expertise of healthcare professionals, who meticulously analyze medical data, patient history, and other relevant factors to arrive at accurate diagnoses [8]. As AI algorithms are integrated into the diagnostic process, ensuring the transparency of their predictions becomes paramount. Physicians and clinicians demand an understanding of why an AI model arrives at a specific diagnosis to instill confidence in its use and facilitate better-informed clinical decisions [9]. Explainability bridges the gap between AI models’ predictive capabilities and human comprehension, fostering trust and facilitating AI adoption in healthcare [5,7]. Moreover, from an ethical perspective, explainability enhances transparency and accountability. The “right to explanation” is an emerging ethical principle, particularly in the context of AI deployment [9]. This principle asserts that individuals have the right to understand the rationale behind AI-driven decisions concerning their health and well-being. In medical practice, this principle not only aligns with regulatory requirements but also safeguards patients’ autonomy and informed consent [3].

Amid the growing concern over the black-box nature of AI models, researchers have explored alternative techniques that offer boosted interpretability. Fuzzy Cognitive Maps [10,11], an established method in the realm of soft computing, have emerged as a promising paradigm to address the challenge of explainability in ML and DL applications. FCMs are graph-based models that excel in capturing complex causal relationships in a dynamic system [11]. Comprising a collection of interconnected concepts represented as nodes, FCMs leverage fuzzy logic to model imprecise relationships between these concepts [10,11]. The strengths and directions of connections, represented by weighted edges, enable the representation of expert knowledge and domain expertise, making FCMs particularly suitable for medical applications [12–14]. Recent publications, such as those by Sovatzidi et al. [15–17], have further advanced the field by enabling the processing of images within the FCM framework. They achieved this using transfer learning and K-means clustering, thereby opening new avenues for more transparent and explainable models in medical domains, including CAD.

1.2. Related Studies

The referenced literature encompasses a diverse array of medical cases specifically related to CAD. These cases have been successfully addressed through the application of DL and FCM methodologies. These studies serve as foundational works, illustrating both the capabilities and limitations of existing techniques in CAD diagnostics and treatment planning.

1.2.1. CAD Diagnosis Using CNNs

Papandrianos et al. [7] aimed to develop an explainable DL methodology for the automatic 3-class CAD classification problem (infarction, ischemia, normal) with Single Photon Emission Computed Tomography (SPECT) images. The dataset included 625 patients in stress and rest representation. Data augmentation was utilized to expand the dataset and achieve generalization. The model presented efficient results with 93.3% accuracy and 94.58% AUC, where k-fold cross-validation was used to ensure the model's reliability. Grad-CAM was also utilized as an explainable tool for the classification outcomes of the proposed model.

Papandrianos et al. [18] developed RGB-CNN for two-class CAD classification with SPECT MPI images. A total of 224 patients were included in this study, and they were in stress and rest representation. Transfer learning was employed for comparison reasons against RGB-CNN with benchmark CNN models, like VGG-16, MobileNet, and InceptionV3. RGB-CNN outperformed these, with a $93.47 \pm 2.81\%$ accuracy. Papandrianos et al. [19] explored three CNN architectures for 3-class CAD classification: RGB-CNN, VGG-16, and DenseNet-121. The dataset contained 647 CAD instances in SPECT-MPI format, and data augmentation was performed to increase the number of available instances. RGB-CNN, VGG-16, and DenseNet-121 achieved an accuracy of 91.86%, 88.54%, and 86.11%, respectively. Ten-fold cross-validation was also utilized.

Apostolopoulos et al. [20] explored CNN to classify polar maps into normal and abnormal. The study consisted of 216 patient cases in stress and rest demonstrations, where the polar maps were in both attenuation-corrected (AC) and non-corrected (NAC) formats. VGG-16 was implemented with transfer learning, and 10-fold cross-validation was applied to estimate VGG-16's performance. Data augmentation was utilized as well. For comparison reasons, semi-quantitative methodologies were used and experts' analyses were performed, and VGG-16 performed best, with accuracy, sensitivity, and specificity values of 74.53%, 75%, and 3.43%, respectively. Semi-quantitative techniques attained 66.20% accuracy. Spier et al. [21] proposed Graph CNN for the automatic classification of myocardial polar maps into normal and abnormal. A total of 946 instances were included in the dataset in stress and rest representations. For further evaluation of the model's performance, a 4-fold cross-validation was developed. Regarding the interpretability of the results, heatmaps were generated that illustrate the region of abnormality in each instance. The proposed model attained 89.9% on rest polar maps and 91.1% on stress polar maps.

Otaki et al. [22] constructed a DL model for the accurate diagnosis of CAD in polar maps. For this research, 1160 patients without known CAD were included in upright and supine positions and only in stress demonstrations. Gender and BMI were also inserted into the dataset as clinical characteristics of patients. For the provision of explainability, Grad-CAM was applied to demonstrate the accountable regions for pathology. Concerning the result, the DL model achieved 82% sensitivity in men and 71% in women. For comparison reasons, the standard SSS (Summed Stress Score) was evaluated, yielding the following values: 75% and 71% for the upright Total Perfusion Deficit (U-TPD), 77% and 70% for the supine (S-TPD), and 73% and 65% in men and women, respectively. Based on the presented results, the study demonstrated significant variations in the diagnostic performance of DL compared to D-SPECT in predicting CAD in men and women. These differences may be attributed to the fact that men and women have different cardiac sizes, and certain factors differ between the genders. For the evaluation of results, leave-one-center-out external validation was applied.

Otaki et al. [23] constructed an explainable DL network from scratch to diagnose obstructive CAD in SPECT MPI images. A total of 3578 patients were included with suspected CAD only in stress representation. Age, gender, and cardiac volumes were fused in the final fully connected layer to enhance the patient's representation of CAD. The DL model performed well with 0.83% AUC against the readers' diagnosis, which produced 0.71% AUC. Ten-fold repeated testing was applied to ensure the reliability of DL's results.

Grad-CAM was applied for the generation of attention maps, highlighting the regions that correspond to the output.

Chen et al. [24] developed a three-dimensional CNN to classify MPI scans into normal and abnormal. A total of 979 instances were included in this study in Cadmium Zinc Telluride (CZT) format and a grayscale model. Grad-CAM was applied to detect the parts contributing to the corresponding prediction. The 3D model attained 87.64% accuracy, where a five-fold cross-validation technique was applied.

1.2.2. CAD Diagnosis Using FCMs

Khodadadi et al. [12] developed an FCM for the diagnosis of the risk of ischemic stroke. A non-linear Hebbian learning method was applied for FCM training to enhance its efficiency. The dataset included a total of 100 cases. The model achieved an overall accuracy of $93.6 \pm 4.5\%$ when using 10-fold cross-validation. The performance of the model was significantly improved with the incorporation of expert knowledge and fuzzy logic. For comparison reasons, Support Vector Machine (SVM) and K-nearest neighbors were developed, attaining accuracies of 86% and 80.2%, respectively.

Apostolopoulos et al. [13] employed a State Space Advanced FCM (AFCM) model and incorporated a rule-based mechanism to enhance the knowledge of the system and its interpretability for the automatic, non-invasive diagnosis of CAD. The authors enhanced AFCM by utilizing advanced state equations, applied SigmoidN as an activation function, which has been applied in different studies [24], and constructed the proposed model RE-AFCM. A total of 303 patient cases were included in this study, consisting of 116 healthy cases and 187 pathological cases. The authors decided to integrate thirty input concepts that best demonstrate patient status concerning CAD diagnosis along with one output concept. The dataset attributes correspond to the factors influencing the diagnosis of CAD. They concluded that the advanced methodologies increased the performance of the model by 7%. RE-AFCM outperformed traditional FCM and ML algorithms, and SigmoidN performs better than Sigmoid. Regarding the results, RE-AFCM attained 85.47% accuracy, with 89.3% sensitivity, 79.3% specificity, and 87.43% and 82.14% for PPV and NPV, respectively.

Apostolopoulos et al. [14] developed an Medical Decision Support System (MDSS), where FCM was applied for the prediction of CAD. This study was an enhancement of previous work, and a total of 303 patient cases were included, including 116 healthy cases and 187 pathological cases. The stenosis of the coronary artery was the only criterion for the labeling of instances. The authors decided to include thirty input concepts that represent patients regarding CAD diagnosis, along with one output concept. The proposed model achieved 78.2% accuracy, 83.96% sensitivity, 68.97% specificity, 81.34% Positive Predictive Value (PPV), and 72.73% Negative Predictive Value (NPV) and outperformed traditional classification algorithms by at least 2%. Based on the fact that the MDSS was not trained on the corresponding dataset, the extracted results are efficient.

1.3. Contribution of this Study

The existing body of literature in the realm of CAD diagnosis through Machine Learning (ML) techniques, particularly CNNs and FCMs, has made noteworthy advancements. However, these contributions often fall short of addressing the critical issue of explainability, thereby limiting their practical utility in clinical settings. The “black-box” nature of such models poses a significant barrier to their adoption by healthcare professionals who require transparent decision-making processes for ethical and practical reasons.

In light of these limitations, the present study introduces a groundbreaking methodology, termed DeepFCM, which aims to bridge this gap by offering a truly transparent and explainable model for CAD diagnosis. Unlike existing models, DeepFCM is designed to integrate both imaging data, specifically MPI polar maps, and tabular clinical data. This multi-modal approach not only enhances the model’s diagnostic accuracy but also its interpretability. The cornerstone of DeepFCM’s transparency lies in its three-pronged approach to explainability: (i) Visual explainability: The model incorporates an integrated

Gradient Class Activation Mapping (Grad-CAM) algorithm, which illuminates the significant regions within the MPI polar maps. This feature provides clinicians with a visual guide to the areas of interest that influenced the model's decision, thereby enhancing its transparency. (ii) Weight disclosure: DeepFCM goes a step further by revealing the internal weights and their corresponding influence on the diagnostic outcome. This level of transparency is instrumental in fostering trust and facilitating the model's adoption in clinical practice. (iii) Textual explanation: To bridge the gap between machine-based reasoning and human comprehension, the model employs the state-of-the-art Generative Pre-trained Transformer (GPT) 3.5 to generate coherent, meaningful, and human-readable explanations. This feature serves as a valuable tool for medical professionals, aiding them in making well-informed clinical decisions.

2. Materials and Methods

2.1. Coronary Artery Disease Dataset

2.1.1. Data Acquisition

Between 16 February 2018 and 28 February 2022, the Department of Nuclear Medicine at the University Hospital of Patras conducted a study involving 2036 consecutive patients who underwent gated-SPECT MPI using ^{99}mTc -tetrofosmin. They employed two-hybrid SPECT/CT gamma-camera systems (Varicam, Hawkeye, Infinia, Hawkeye-4, GE Healthcare) for MPI. Attenuation correction (AC) based on computed tomography was applied to stress and rest images for all patients. Among these patients, 506 individuals underwent ICA within sixty days of MPI for further examination. After excluding twenty patients due to inconclusive MPI results or missing ICA reports, the final study population consisted of 594 patients, with 43.82% showing CAD-positive results.

The ethical committee of the University General Hospital of Patras approved the data collection process (Ethical and Research Committee of the University Hospital of Patras, protocol number 108/10-3-2022). As this study was retrospective in nature, informed consent from the participants was not required. All data-related procedures were conducted anonymously and in accordance with the Declaration of Helsinki. The diagnostic results of MPI SPECT were provided by three experienced nuclear medicine physicians, who independently inspected the polar maps and resolved any discrepancies through consensus.

The raw image data were tomographically reconstructed on a dedicated workstation (Xeleris 3, GE Healthcare, Chicago, IL, USA) using the Ordered Subset Expectation-Maximization (OSEM) algorithm with two iterations and ten subsets. Subsequently, a low-pass filter (Butterworth) was applied with specific parameters for stress and rest images. The dedicated software (Xeleris 3.0513) automatically generated polar maps, which are 2D circular representations summarizing the results of the 3D tomographic slices. These polar maps were saved in Digital Imaging and Communications in Medicine (DICOM) format for further processing.

2.1.2. Image Data Preprocessing

In the domain of medical image analysis, image preprocessing holds immense significance as it serves to enhance the quality of DICOM images before embarking on further analysis [25]. Among the essential transformations applied, the utilization of color maps stands out as a key technique to convert grayscale images into visually informative colored representations. This enables the visualization of distinct structures within the image, such as bones and soft tissue, by assigning different colors to various anatomical elements. In the realm of medical imaging, where accurate identification and differentiation of structures are paramount, color maps prove to be an invaluable aid [7].

To execute the application of color maps, widely used libraries like Matplotlib offer a diverse selection of color maps to suit specific requirements. In the context of this study, each patient's data yield four polar maps, each corresponding to different conditions, capturing information with and without attenuation correction during rest and stress states. These four polar maps are then thoughtfully consolidated into a single Joint Photographic

Experts Group (JPEG) image format, preparing them for input into the model, enabling more in-depth analysis and interpretation. The transformation of the image via color maps serves an additional purpose, converting the visual representation into a numerical array format. This numerical representation unlocks the potential for conducting various numerical operations, such as filtering, segmentation, and registration, on the image data. It facilitates advanced analytical techniques and computational methodologies to extract meaningful insights from medical images [7].

In light of DL models' prevalence in modern medical image analysis, proper normalization of the image data becomes paramount. This step ensures that the image data are brought within a consistent range, avoiding issues arising from variations in scales and intensity. Given that DL models are highly sensitive to the input data's scale, normalization becomes an indispensable preprocessing step [18]. Two commonly employed normalization techniques include dividing the image values by their maximum value or performing subtraction of the mean and division by the standard deviation [2].

2.1.3. Clinical Data Preprocessing

High-dimensional data analysis is a challenge for researchers and engineers in the fields of ML and data mining algorithms. Feature selection provides an efficient approach to dealing with this problem by removing redundant features, which uses less computation time, enhances learning accuracy, and provides a more comprehensive dataset [26]. In feature selection, a subset of an original dataset is acquired that includes the most relevant features of the dataset [27].

Feature selection was conducted in this paper based on a previous study of the research team at EMERALD [26], where it was applied for CAD classification using clinical data and expert diagnosis. More specifically, five ML algorithms were applied to the dataset, and each ML algorithm generated a subset. This procedure was conducted both with input from expert assessments and doctors and without. To determine the optimal feature set for each algorithm, three feature selection algorithms have been developed: forward sequential feature selection, backward sequential feature selection, and genetic algorithms. An assessment of performance was conducted using standard metrics to detect the most effective feature set. In our study, the subset with the best performance metrics from the research study [26] was utilized and was named the optimal subset.

2.2. Deep Fuzzy Cognitive Map Model

Our proposed DeepFCM model includes the combination of the FCM, CNN, PSO, Grad-CAM, and NLP, along with expert knowledge. The FCM handles the clinical data, and PSO is responsible for the calculation of the weight matrix, which includes the interconnections among concepts. The CNN's role involves handling image data by extracting predictions for each case study and providing extra input to the FCM model. Grad-CAM interprets CNN predictions, and NLP evaluates the total process. The DeepFCM model combines the clinical data with the CNN's output, and the whole process is demonstrated in Figure 1.

2.2.1. Fuzzy Cognitive Maps

FCMs are soft computing tools that are a combination of fuzzy logic and neural networks [13]. FCMs were introduced by Kosko [10] as an advanced version of cognitive maps with the application of fuzzy causal functions with real numbers to the connections. The FCM is a fuzzy diagram that transforms a system into concepts, where each concept represents a variable, a state, or a characteristic of a system. Between concepts, there are weight values/interconnections that demonstrate how the concepts interact with each other. There are three types of weight values. More specifically, $w_{ij} = 0$ means that there is no causality between concepts, $w_{ij} > 0$ indicates a positive relationship, and $w_{ij} < 0$ determines a negative causality. Expert knowledge defines the number of nodes and the initial values of the interconnections among concepts. An FCM can be described by a weight matrix that

includes all the interconnections among concepts and the state vector, which has the values of concepts [10]. The value of each concept is influenced by the values of the connected concepts with the corresponding causal weights and by its previous state. The sum of the concept values of nodes together demonstrates the state vector of the system.

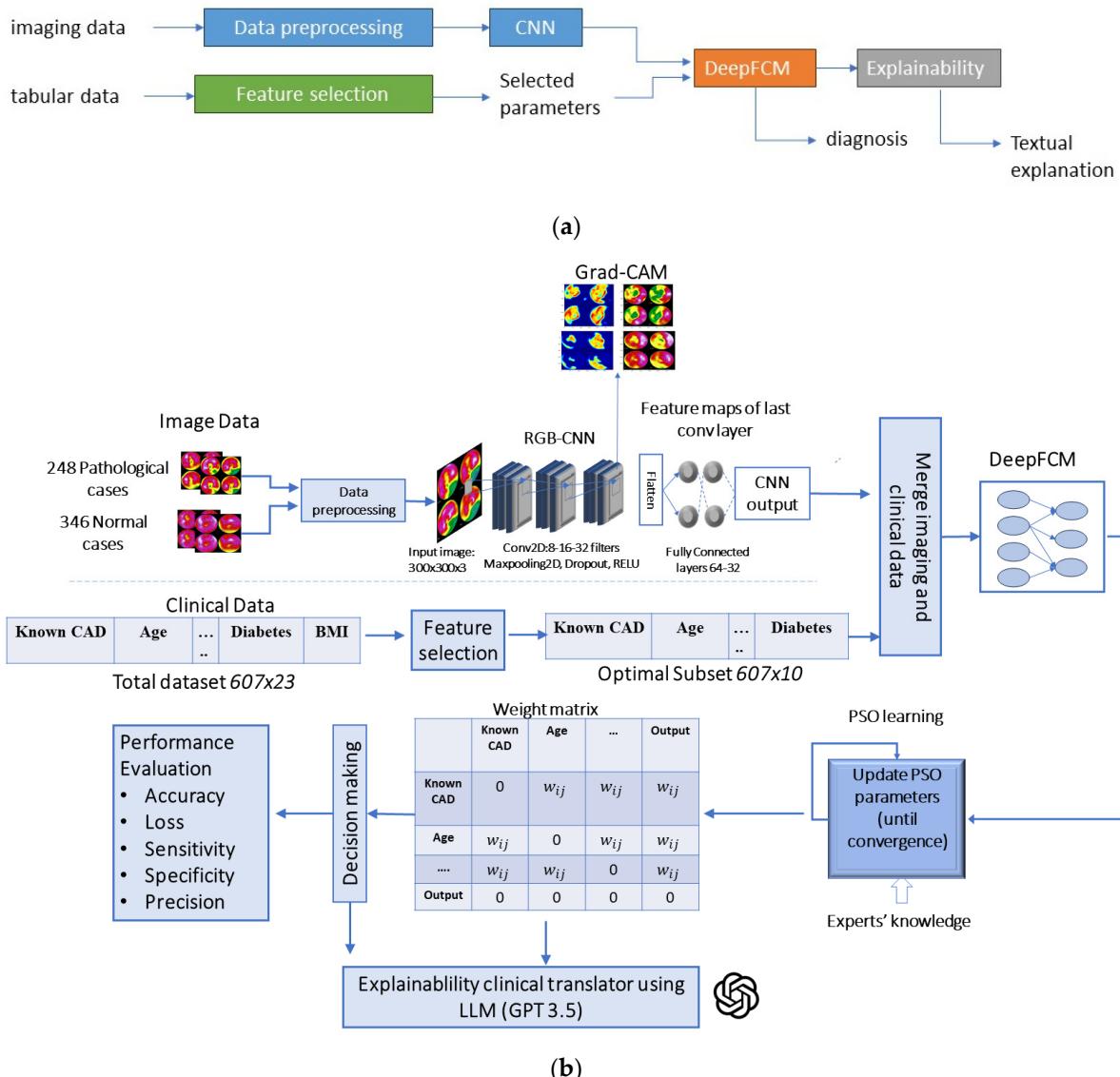


Figure 1. Figure of proposed methodology for DeepFCM: (a) high-level flowchart and (b) detailed flowchart.

Regarding the FCM inference and the evolution of the system, FCM calculates iteratively the state of concepts until it reaches the equilibrium point by multiplying the vector of concepts' values with the weight matrix [14,28,29]. To normalize the FCM-predicted values of concepts into a specific range after multiplication, a transfer function is used, where a sigmoid, bivalent, or trivalent function is applied. The sigmoid function is demonstrated in type (2). The equation is presented below.

$$A_i^{(K+1)} = f \left(A_i^{(K)} + \sum_{i,j}^N w_{ij} A_j^{(K)} \right) \quad (1)$$

where $A_i^{(K+1)}$ is the value of the concept iteration $(K + 1)$, $A_j^{(K)}$ is the concept at the iteration (K) , and f is the transfer function.

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

In the generated weight matrix, the diagonal has zero values since every concept influences other concepts but not itself [14].

After multiple iterations, the FCM could lead to one of the following scenarios:

- i. The equilibrium point is where the current states of the FCM have converged to steady values.
- ii. Limit cycle behavior, where the final state of the FCM, which indicates the outputs, in each iteration takes specific values.
- iii. Chaotic behavior, where each concept takes random and unstable values.

2.2.2. RGB-CNN

A CNN indicates an algorithm that mimics humans' decision making. A CNN consists of input, hidden, and output layers. The hidden layer incorporates convolutional, pooling, dropout, and fully connected layers and aims to extract patterns from image data. A CNN can automatically extract features by employing a variety of filters on the input images, and via an advanced learning process, the most significant pixel values are retained. CNNs have extracted remarkable results from previous medical studies [2,7,18,19].

The convolutional layer is the first layer of a CNN, and its primary functionality is the creation of a feature map, which contains an abstract representation of the input image [2]. Pooling layers are applied after every convolutional layer to down-sample the data and reduce the computational complexity of the network. Pooling layers select the maximum or average value within a local window to retain the most important features [2]. Concerning the dropout layer, it helps to reduce unnecessary pixels and prevents overfitting. The dropout layer nullifies random pixel values to decrease the computational time of the training process. A flattening layer is applied next to convert multi-dimensional data into a vector. The last layer of the CNN is a sequence of fully connected layers, where each node is connected to the preceding one leading to the network's final prediction [7]. In most CNN models, the Rectified Linear Unit (ReLU) is utilized in convolutional and fully connected layers as an activation function, and sigmoid or softmax are used for output layers for binary and multiclass classification problems, respectively [7,30,31].

After a thorough exploration process, we concluded with the ideal combination of parameters regarding pixel values, batch size, dropout rate, and number of nodes and layers of convolutional and fully connected layers. The proposed RGB-CNN rescales the input images to 300×300 pixels and consists of three convolutional layers with 8, 16, and 32 filters (kernels) accordingly. After each convolutional layer, there is a max pooling with a 2×2 kernel size and a dropout layer with a drop rate of 0.2. Next, the flattening layer transforms multi-dimensional data into vectors to prepare data for fully connected layers. Regarding the fully connected layers, we selected two layers with 64 and 32 nodes, accordingly. The output layer is a single-node layer with a sigmoid activation function since we dealt with a two-class classification problem. Data augmentation was also utilized to increase the dataset size by generating altered versions of the original images. In our research, we applied rescaling to all images as a normalization technique, and regarding data augmentation, we employed `width_shift_range = 0.1`, `height_shift_range = 0.1`, `shear_range = 0.1`, and `zoom_range = 0.1` to prevent overfitting and develop a generalizable model [2,7,32].

2.2.3. Integration of CNN Predictions and Clinical Data to Construct the DeepFCM Model

The FCM was initially designed using clinical characteristics as the input, regarding patient status. We strengthened the FCM's performance by introducing CNN predictions

as additional input and exploiting the automatic feature extraction process. This hybrid approach leverages the strengths of the clinical data, where feature selection was employed to preserve the most important features and CNN—derived insights from medical images while extracting high—dimensional features. This integration constructs an innovative approach that provides a more comprehensive and accurate CAD diagnosis and provides a holistic view of patients' conditions to nuclear experts. DeepFCM is more likely to detect risk factors at an early stage and reduce diagnostic errors.

2.2.4. Initialization of DeepFCM Weights by Experts

Regarding the initialization among interconnections of concepts in an FCM, linguistic values can be provided by nuclear experts in a fuzzy set format. In general, fuzzy sets deliver uncertainty and mimic human knowledge. Traditionally, in binary logic, a statement can be true or false, and in set theory, an element can belong to only one set [32]. Fuzzy sets introduce partial truth, which is defined by human language and decisions. Fuzzy sets were defined by Zadeh [33], and they have been useful in pattern recognition and medical diagnosis.

Membership functions (MFs) are the building blocks of fuzzy set theory since they introduce the degree of fuzziness in a fuzzy set [34]. MFs can be developed in various shapes and should be compatible with the problem since it affects a fuzzy inference system. The different shapes could be triangular, trapezoidal, Gaussian, etc. The membership values, regardless of the shape, should depend on the range $[0, 1]$. The membership function which represents a fuzzy set is defined as μ_A , and for an element x of set X , the term $\mu_A(x)$ is the membership degree of element x in set X [35]. In Figure 2, we can see a demonstration of a fuzzy set with a triangular membership function. The triangular membership function is characterized by a set of three parameters, denoted as $\{a, b, c\}$, where c represents the base of the triangle and a and b determine the height [35]. These parameter values are established based on experts' knowledge.

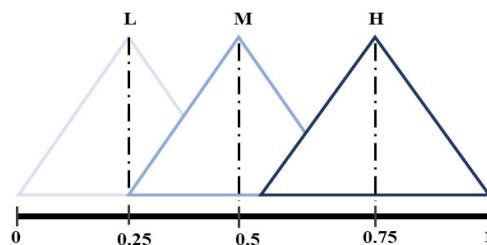


Figure 2. Demonstration of three fuzzy sets: low, medium, and high.

In this research study, the linguistic values provided by nuclear experts were as follows: Very Weak (VW) with a range $[0, 0.3]$, Weak (W) with a range $[0.15, 0.5]$, Medium (M) with a range $[0.35, 0.65]$, Strong (S) with a range $[0.5, 0.85]$, and Very Strong (VS) with a spectrum $[0.7, 1]$. The linguistic values were transformed into numerical ranges to be utilized in the algorithm.

2.2.5. DeepFCM Learning, Weight Initialization, and Update of Weights with PSO

In FCMs, the weight matrix includes all the interconnections among concepts [10]. The initial values of the weight matrix are randomly initialized or are based on linguistic values suggested by nuclear experts. In our case, the relationship values among meaningful concepts were initialized based on expert knowledge, as displayed in Table 1. The initial values of the rest of the interconnections were randomly selected from the range $[-1, 1]$.

Table 1. Representation of the suggested weights between meaningful input–input concepts and input–output concepts obtained from nuclear experts.

Relationship among Concepts	Linguistic Value Provided by Experts
Known CAD→Output	S
Previous PCI→Output	W
Diabetes→Output	S
Chronic Kidney Disease→Output	W
Angina Like→Output	S
ECG→Output	M
Male→ECG	W
Expert Diagnosis→Output	VS
CNN→Output	S

Particle Swarm Optimization (PSO) is a population-based algorithm that includes a collection of individuals to search for optimal regions in the search space. PSO utilizes a small number of parameters [36]. The population is called a swarm, and the individuals are called particles. Every particle in the system can move with a defined velocity within a search space, and it preserves the best position that has been encountered. Once the global best position has been calculated, it is shared with all the particles in the group [37].

In the present study, PSO was employed to handle the computation of the weight matrix and adapt the interconnections among concepts by minimizing the error of the objective function. For every particle, a weight matrix is generated, and the one that produces the minimum error when comparing the actual output with the predicted is forwarded to the testing phase. The weight matrix is critical for the FCM's performance, and the desirable characteristics of the weight matrix include stability and alignment with the dataset's characteristics while producing minimal error.

2.2.6. DeepFCM Inference—Natural Language

The concluding weight matrix, along with the concept vector and the clinical characteristics along with RGB–CNN's prediction, were inserted into a robust NLP system (GPT-3.5) to discuss the results and verbalize them in natural language. Grad–CAM was also employed in this research to interpret the RGB–CNN predictions and transform them from a black–box model to a more comprehensive model with transparent inner computations related to CAD diagnosis in polar map images.

GPT-3.5, as a Large Language Model, is trained with neural network methodologies on billions of words derived from articles, books, and internet content, and it learns the relationship between words and generates text by following patterns observed in sentences from its training data. In the context of inference, when a user inserts a prompt, GPT-3.5, drawing from its existing knowledge and the patterns that it has extracted, generates a response in human-like language [38–40].

2.3. Explainability-Enhancing Methods

2.3.1. Self-Explainable Aspects of DeepFCM

At the core of an FCM lies the weight table [10], a fundamental component that defines the strength of relationships between interconnected concepts. The weight table serves as a quantitative representation of domain knowledge and expertise, either contributed by domain experts or inferred from historical data [14]. Each element in the weight table specifies the degree of impact that one concept exerts on another, reflecting the causal influence in the cognitive mapping.

The weight table's explainable nature stems from its intuitive and comprehensible structure. Domain experts can easily comprehend the impact of each concept on others, as the weight values are interpretable and can be expressed in linguistic terms, such as “strong”, “weak”, “positive”, or “negative”. This transparency facilitates expert involvement in model development and validation, providing a valuable opportunity to refine and

fine-tune the model based on domain-specific insights. Furthermore, the weight table's transparency fosters the detection of influential concepts, as high weights indicate strong causal connections. This feature becomes particularly relevant in critical applications such as medical diagnosis, where identifying significant factors affecting the outcome is essential for effective decision making.

Another critical aspect contributing to the inherent explainable nature of FCMs lies in their representation of interconnections between concepts. The directed edges between nodes in an FCM indicate the causal relationships and the direction of influence from one concept to another. These connections represent the cause-and-effect dependencies that govern the dynamics of the system under consideration. By visually examining the graph structure of the FCM, domain experts can identify complex causal pathways, feedback loops, and interdependencies between concepts. This understanding not only enhances the model's transparency but also facilitates the identification of potential bottlenecks, vulnerabilities, or reinforcing factors within the system. Moreover, the causal relationships depicted by the interconnections in FCMs allow for "what-if" analyses, where experts can assess the impact of hypothetical changes to specific concepts on the overall system behavior. Such analyses promote risk assessment and strategic decision making in various applications, including policy planning, environmental management, and healthcare interventions.

In comparison to other complex ML models, FCMs offer several distinct advantages in terms of interpretability [28]. Neural networks, for example, are notorious for their black-box nature, as the intricate relationships within their hidden layers are challenging to unravel. In contrast, FCMs' explicit representation of causal connections fosters a holistic understanding of the decision-making process, enabling domain experts to assess model predictions with confidence and identify any potential biases or anomalies. Similarly, decision trees, although interpretable, may lack the expressive power to capture complex and uncertain relationships among variables. FCMs, however, excel at dealing with such complexities through the use of fuzzy logic, allowing for continuous and gradable influence between concepts [10].

2.3.2. Natural Language Processing Models

While FCMs offer transparency in the form of weight tables and concept interconnections, translating these numerical outputs into easily understandable language can be a formidable task for domain experts, especially in the medical domain. GPT-3.5 is a language model developed by OpenAI [40]. It is designed for natural language understanding and generation tasks. The model has been trained on a wide range of internet text to be able to provide informative and contextually relevant responses to user prompts [41]. GPT-3.5 is capable of understanding and generating human-like text, making it a versatile tool for various applications, including answering questions, assisting with writing, tutoring, and more [38,42].

To use GPT-3.5 for generating medical diagnoses from classification models, we integrated it into the DeepFCM framework, which incorporates both GPT-3.5 and the classification model. The process involves the following steps [38]:

- Train the classification model: Firstly, we developed a specialized classification model specifically for medical diagnosis, as explained. This model is trained on medical data with labeled diagnoses to accurately classify patients' conditions based on their symptoms and other relevant information.
- Integrate GPT-3.5: After training the classification model, we integrated GPT-3.5 into the medical information system. OpenAI provides an API that allows developers to access GPT-3.5 programmatically and send prompts for generating responses.
- Prompt generation: When a user enters medical information, such as symptoms or test results, into the system, the classification model generates a diagnosis that involves the final classification output, the final output vector, and the final weights of the DeepFCM model. These components are used as a prompt for GPT-3.5.

- Obtain response: GPT-3.5 will process the prompt and generate a response in natural language. This response could be a human-readable explanation of the diagnosis provided by the classification model, additional information about the condition, potential treatment options, or other relevant insights.
- Present results to users: We displayed the generated response to the user, helping healthcare professionals understand the reasoning behind the diagnosis and make informed decisions regarding patient care.

2.3.3. Gradient Class Activation Mapping (Grad-CAM)

Grad-CAM was introduced by Selvaraju et al. [43,44]. It is an interpretation method that provides insights into the decision-making process of CNNs by visualizing the important regions of an input image that contribute most significantly to a specific classification decision. In Grad-CAM, the gradients of the target class score with respect to the feature maps of the final convolutional layer are used to determine the importance of each spatial location within the feature maps [43]. These gradients are averaged to obtain the final class activation map, highlighting the regions that strongly influence the CNN's decision for a particular class. Although Grad-CAM provides valuable localization information, it has limitations when dealing with multiclass tasks, i.e., distinguishing between different object categories in a single image. It tends to emphasize only the most dominant object category, failing to capture intricate details in other classes [45,46]. In the present study, Grad-CAM was adapted to provide visual explanations of polar map images.

2.4. Experiment Setup

In terms of hardware and software specifications, the experiments were conducted on a Dell G15-5515 laptop with an AMD Ryzen™ 7 5800H Mobile (20 MB total cache, 8 cores, 16 threads), with an operating system Windows 11 Home Edition. The available laptop consists of 16 GB RAM with 2 x 8 GB, DDR4, 3200 MHz, and an NVIDIA GeForce RTX™ 3060, 6 GB GDDR6, 3 DP card. Regarding the coding process, we employed Python 3.9.0, utilizing TensorFlow 2.10.1 and Keras 2.10. To handle our dataset, the OpenCV library was employed, and for dataset splitting and result computation, scikit-learn was utilized. An investigation was applied through a series of experiments to conclude the suggested architectures. This involved evaluating various architectures for FCM and PSO parameters.

Commonly employed performance metrics were utilized in this study, like accuracy, loss, sensitivity, specificity, and precision, to assess the effectiveness of the proposed DeepFCM architecture. Accuracy denotes the ratio of correctly classified instances to the total number of instances. Loss indicates the error between predicted and actual values. Sensitivity and specificity indicate the true positives (TP) and true negatives (TN) accordingly. Precision is the ratio of the number of true positives to the total number of positive predictions. These metrics offer a well-rounded assessment of the model's capabilities, addressing both the classification accuracy and its ability to correctly identify CAD-positive and CAD-negative cases [2,19].

Regarding ensuring our model's robustness, k-fold cross-validation was performed, where k represents the number of partitions into which the dataset is divided [47]. In our case, we divided the dataset into 10 partitions, of which 9 were utilized as training and 1 as testing. This process was repeated until each partition had been utilized for testing. With the application of k-fold cross-validation, overfitting can be avoided and generalization can be ensured [47]. To reduce redundant training iterations, early stopping was applied. Early stopping is a regularization technique that is utilized in the CNN training process to prevent overfitting and improve generalization error and overall accuracy as well. Early stopping enhances the training process and minimizes computation time. Moreover, it inspects the generalization error that is calculated during the training process and stops the training [48].

3. Results

3.1. Classification Results

The DeepFCM methodology achieved an accuracy of 83.07%, with a sensitivity of 86.21% and a specificity of 79.99% (refer to Table 2). For the sake of comparison, we also provide the results of the standalone RBG–CNN model, which solely processes the images. Additionally, we showcase the performance of DeepFCM when it utilizes the full feature set, rather than just the selected features. We further present the diagnostic yield of doctors, which is based on their visual inspection of the polar maps and consideration of the clinical data. Readers need to note that the benchmarks for all these comparisons are the results from invasive coronary angiography.

Table 2. Comparison of results among expert diagnosis, RGB–CNN applied to images only, and DeepFCM applied to the optimal subset and total dataset.

Run	Accuracy	Loss	Sensitivity	Specificity	Precision
Expert diagnosis	78.91	0.21	77.9	79.7	75.09
RGB–CNN	75.42 ± 4.54	0.47	80.53	66.38	66.39
DeepFCM optimal subset	83.07 ± 4.72	0.17	86.21	79.99	81.78
DeepFCM all features	76.1 ± 5.52	0.24	77.79	72.78	71.67

It is demonstrated that DeepFCM applied to the optimal subset performed remarkably and achieved higher performance metrics. The integration of both imaging and clinical data surpassed the RGB–CNN model and expert diagnosis. Furthermore, feature selection enhanced the results by leading to a more effective representation of the dataset, as we can see by comparing the 83.07% accuracy with the optimal subset and 76.1% with the total dataset. The results showcase the promising potential of the hybrid DeepFCM model. The proposed model, DeepFCM, applied to the optimal subset attained average accuracy, with a mean of 83.07% and a small standard deviation of ±4.72%. The low loss value of 0.17 reflects the model’s robust convergence during training. The sensitivity of 86.21% and specificity of 79.99% signify its competence in identifying CAD–positive and CAD–negative case studies, respectively. Moreover, a precision of 81.78% highlights the model’s capability to accurately classify true positive cases.

In Table 3, we demonstrate the interconnection among concepts in the first column, the range of values provided by experts for the initialization of interconnection to be randomly selected in the second column, and the weight value that was produced from DeepFCM for the according interconnection in the third column. We can observe that the DeepFCM applied to the optimal subset generated efficient results that are close to the initial range with small deviations.

Table 3. Presentation of extracted ranges for the relationship between concepts produced from the DeepFCM model applied to the optimal subset and comparison with expert knowledge.

Relationship of Concepts	Transformed Linguistic Values to Ranges	Produced Values from DeepFCM
Known CAD→Output	[0.5, 0.85]	0.68 ± 0.08
Previous PCI→Output	[0.15, 0.5]	0.34 ± 0.09
Diabetes→Output	[0.5, 0.85]	0.63 ± 0.1
Chronic Kidney Disease→Output	[0.15, 0.5]	0.32 ± 0.009
Angina Like→Output	[0.5, 0.85]	0.67 ± 0.12
ECG→Output	[0.35, 0.65]	0.43 ± 0.08
Male→ECG	[0.15, 0.5]	0.17 ± 0.1
Expert Diagnosis→Output	[0.7, 1]	0.88 ± 0.1
CNN→Output	[0.5, 0.85]	0.75 ± 0.09

3.2. Interpretation Results

3.2.1. Grad-CAM

Regarding explainability, Grad-CAM was utilized in this research to provide interpretability to the RGB-CNN predictions. The Grad-CAM results were evaluated by a nuclear medicine specialist to strengthen the reliability of the results. For the implementation of Grad-CAM, the feature maps were extracted from the last convolutional layer of the RGB-CNN model and inserted into the Grad-CAM for the generation of heatmaps. The heatmaps demonstrate the impact of each pixel on the final prediction of RGB-CNN. In this study, the colormap entitled Jet was utilized, obtained from the OpenCV library, where low-impact pixel values are colored in blue and high-impact values are colored in red, as seen in Figure 3. Based on nuclear medicine experts, the Grad-CAM results offer interpretability and transparency of RGB-CNN's inner computations, regarding CAD diagnosis.

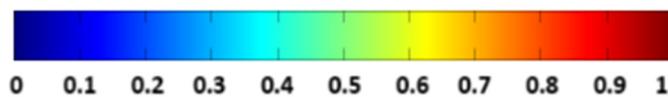


Figure 3. Jet colormap color range demonstration.

In Figure 4, we demonstrate the Grad-CAM application to four pathological case studies, and in Figure 5, four normal case studies, related to CAD diagnosis. Regarding examples a and b, they correspond to correctly predicted instances, while c and d refer to cases that were falsely predicted. The different colors specify the importance of pixels, demonstrating the impact of the CNN classifier on individual pixels. Grad-CAM identifies the regions of interest and colors them in red. We observe that the Grad-CAM has provided exceptional and comprehensive results, where it correctly highlighted and detected the regions of interest.

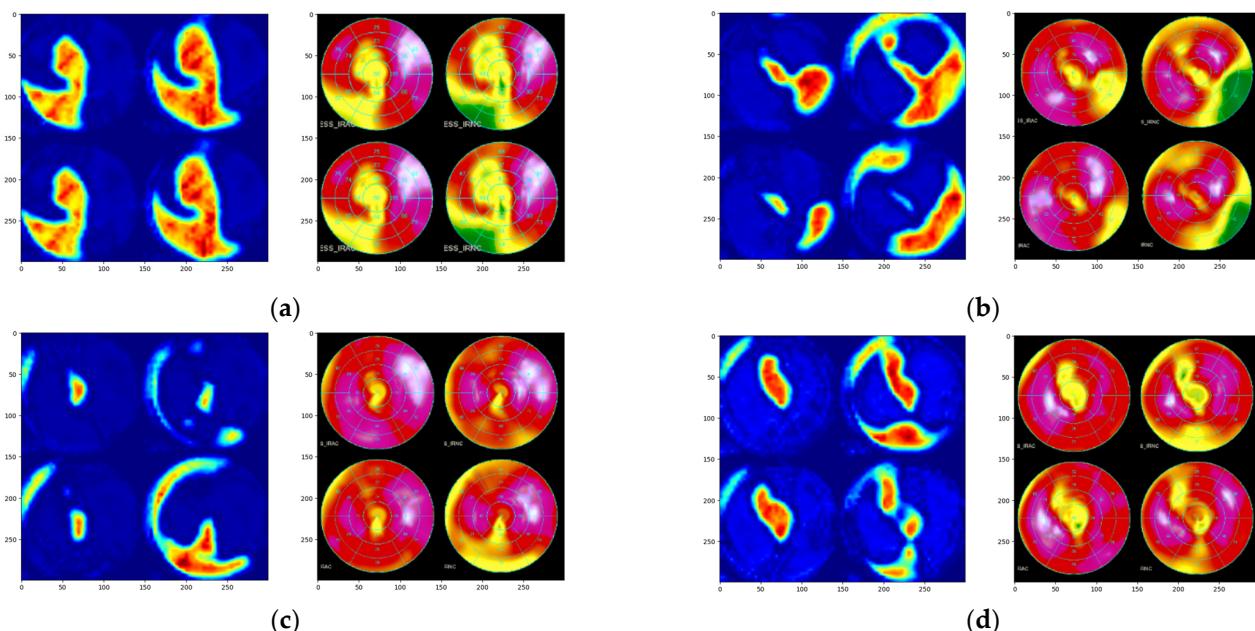


Figure 4. Representation of Grad-CAM application to pathological polar maps: (a) first TP case study, (b) second TP case study, (c) first False Positive (FP) case study, (d) second FP case study.

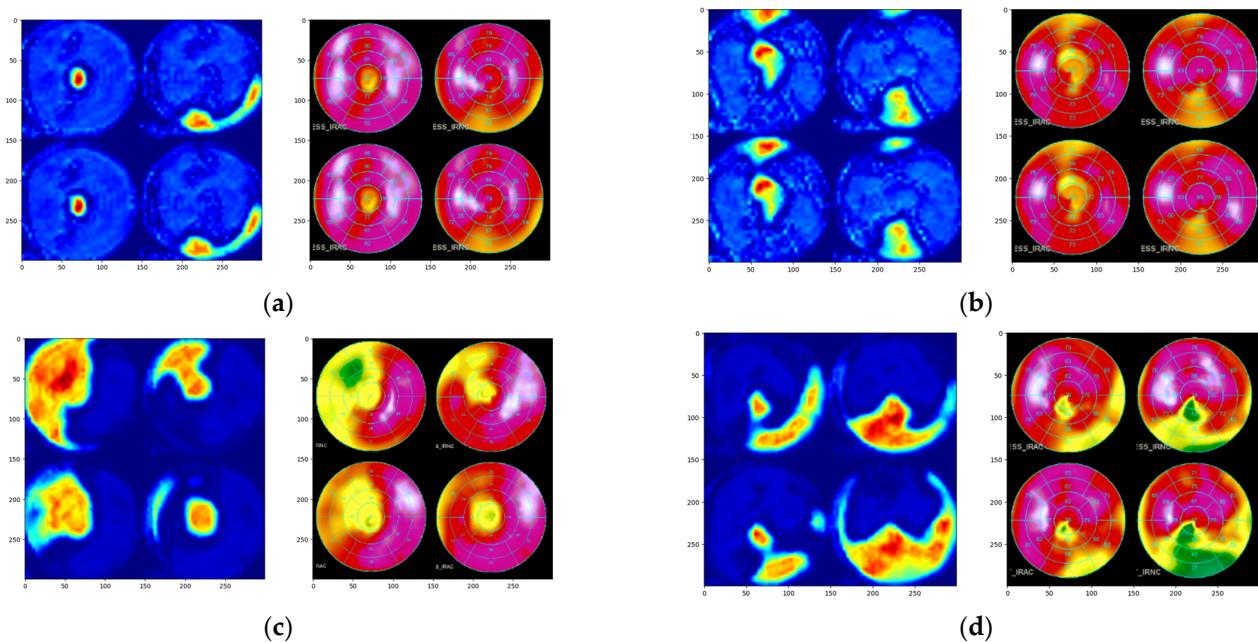


Figure 5. Representation of Grad-CAM application to normal polar maps: (a) first TN case study, (b) second TN case study, (c) first False Negative (FN) case study, (d) second FN case study.

3.2.2. NLP

In relation to NLP processing, we present two case studies with the following clinical characteristics. In the first case, the patient has no documented history of CAD, diabetes, or chronic kidney disease. There is no record of the patient having undergone Percutaneous Coronary Intervention (PCI), and the Electrocardiogram (ECG) results are normal. However, the patient exhibits symptoms related to angina. The patient is a male over 40. Both medical experts and the trained RGB-CNN correctly classified this instance as pathological. In the second case, the patient does not have any known history of chronic kidney disease or angina and has no record of having undergone an ECG procedure. Nevertheless, the patient displays symptoms associated with previous CAD, PCI, and diabetes. Additionally, the patient is a male over the age of 40. Both medical professionals and RGB-CNN accurately diagnosed the patient with CAD. The authors introduced the case studies into GPT-3.5 by creating a text prompt, as shown in Figure 6. This prompt includes a description of our proposed system and allows NLP to evaluate the DeepFCM's performance, the produced concept vector, and the clinical characteristics, regarding the first case study.

You are a Fuzzy Cognitive Map explainability-enhancing AI tool. Your task is to make doctor-friendly interpretations of FCM outputs. (Analysis of input concepts). Here is the final weight matrix, which shows the weights between each input concept and the output: 0.82, 0.57, 0.59, 0, 0.44, 0.43, -0.12, -0.42, 0.93, 0.91. A user will give you the initial values of each concept and the predicted class. You must give them a meaningful and informative explanation of the decision, based on the values of the input concepts and the weight matrix. Focus more on the concepts that have non-zero values. Avoid using the weight values in numbers.

Figure 6. Representation of text prompt applied to GPT-3.5 regarding the first case study.

In Figures 7 and 8, we present the texts generated by GPT-3.5 for the two case studies accordingly, where GPT-3.5 provided a detailed analysis of the relationship of each clinical characteristic with the output. We can observe that the produced texts offer nuclear experts an interpretable and thorough demonstration of patient status and provide the logic behind DeepFCM's predictive capabilities. GPT-3.5 facilitates the transformation of DeepFCM into a transparent and trustworthy tool suitable for integration into the decision-making process for CAD diagnosis.

Based on the input concept values you provided, it seems that the patient does not have a history of CAD or PCI. The patient does not have diabetes or chronic kidney disease. However, they do experience angina-like pain, which could be a symptom of CAD. The patient's ECG result is not mentioned, so we cannot determine its impact on the diagnosis. The patient is male, and their age is not specified to be below 40 years. The expert diagnosis and the polar map classification are both positive, indicating that the human expert and the CNN model predict the presence of CAD. This aligns with the output class of 1, which suggests that the patient is indeed suffering from CAD. It is important to note that the weight matrix values indicate the strength of the relationships between the input concepts and the output class, but they are not directly interpretable in verbal language. Instead, we focus on the presence or absence of certain input concepts and their overall impact on the diagnosis. In this case, the presence of angina-like pain, positive expert diagnosis, and positive polar map classification contribute to the diagnosis of CAD.

Figure 7. Demonstration of GPT-3.5 results regarding the first case study.

Based on the given input concept values and the weight matrix, the FCM predicts that the patient is suffering from CAD. Let's analyze the contributing factors. The presence of Known_CAD, Previous_PCI, and Diabetes all have positive weights, indicating that these factors increase the likelihood of CAD. Known_CAD suggests that the patient has a history of CAD, while Previous_PCI suggests they have undergone a Percutaneous Coronary Intervention. The presence of Diabetes is also known to be a risk factor for CAD. On the other hand, Chronic_kidney_disease and Angina_like have zero weights, implying that they do not significantly influence the prediction. Chronic_kidney_disease is not considered a strong indicator of CAD in this model, and Angina_like, although related to angina symptoms, does not have a strong impact on the prediction. The absence of ECG, Male concept, and Age < 40 also have zero weights, suggesting that they do not play a significant role in determining the presence of CAD in this particular case. The Expert_diagnosis and Polar Map classification are both positively weighted, indicating that they contribute to the prediction of CAD. The Expert_diagnosis provided by the human expert and the Polar Map classification from the CNN-based analysis of the Myocardial Perfusion Imaging test are both valuable inputs for identifying CAD. In summary, based on the input concept values and the weight matrix, the FCM predicts that the patient is suffering from CAD due to the presence of Known_CAD, Previous_PCI, Diabetes, Expert_diagnosis, and Polar Map classification. It is important to note that this interpretation is based on the specific weights and input values provided, and further analysis and medical evaluation may be necessary for a comprehensive diagnosis.

Figure 8. Demonstration of GPT-3.5 results regarding the second case study.

4. Discussion

In this research study, we developed an automatic FCM-based model to detect CAD in patients. The dataset included 346 normal and 248 pathological cases, which nuclear experts had initially characterized for classification purposes. Clinical and imaging data (polar maps) were included. Feature selection was applied to improve the results by retaining the features with a higher impact on the final diagnosis and discarding the features that introduced noise and redundancy into the model. For further enhancement of the FCM, we developed RGB-CNN, a lightweight model for CAD diagnosis exploiting CNN abilities in extracting high-dimensional features from the available images. This incorporation of CNN predictions along with historical data constructed the proposed model, DeepFCM, which demonstrated considerable promise in enhancing accuracy and interpretability to healthcare professionals.

The proposed model achieved $83.07 \pm 4.72\%$ accuracy, 0.17 loss, and 86.21%, 79.99%, and 81.78% for sensitivity, specificity, and precision, respectively. With the application Grad-CAM, RGB-CNN's computations became interpretable and transparent, enhancing the model's explainability for research applications. Furthermore, the results obtained from DeepFCM were integrated into the NLP model (GPT-3.5) to enhance the interpretability of the findings.

DeepFCM can elucidate intricate cause-and-effect relationships among symptoms, diseases, and treatments inherently. The key to understanding these relationships lies in the interpretation of the weight matrix within the FCM. The weight matrix captures the strength and direction of influences between different nodes or concepts, offering doctors a transparent and interpretable framework to analyze how various factors contribute to a patient's condition. For example, DeepFCM learned that the diagnosis of the expert, which was an integrated feature of the model, should affect the output by a weight of 0.88, which is the largest observed weight in Table 3. This fact indicates that DeepFCM distinguishes the human expert as the most vital contributor to the result, thereby maintaining the doctor-in-the-loop approach. The weight matrix implied an inner relationship between the ECG outcome and the gender of the patient. More specifically, it was found that when the patient is male, the weight of the ECG should be slightly strengthened (Table 3). This connection is indeed documented in the literature [49], where it was found that men exhibit higher sensitivity in the ECG test. Finally, an interesting observation is that the CNN's prediction on the polar maps strongly affects the output (0.75). To summarize, DeepFCM suggests that the following four factors constitute essential predictors: the human expert's diagnostic yield, the characterization of the polar maps, as provided by the RGB-CNN of the framework, the patient's history regarding CAD, and the presence of angina-like symptoms.

Grad-CAM provided insights into the decision-making process of RGB-CNN by highlighting the regions of the polar maps that are most influential in making a particular classification or diagnosis. It was observed that, in most of the cases, the produced heatmaps indicated decisive features in areas where the polar map was initially green, which is considered acceptable. In healthy subjects, Grad-CAM produced mainly blue heatmaps, which implies no decisive features toward the positive class. However, there were some cases where the produced heatmap pointed out irrelevant areas. Although such cases undermine the performance of DeepFCM, they can be exploited by the medical staff to guess a potential mistake in the framework. For example, a DeepFCM prediction that was based on an inconclusive heatmap can be considered non-reliable by human experts. On the other hand, a meaningful heatmap indicates that DeepFCM has discovered important features and may be considered more reliable.

The integration of the NLP model improved the informativeness of the system. It provided a verbal interpretation of the results. It went beyond raw numerical data and transformed them into meaningful and human-readable descriptions. By providing a verbal interpretation of the results, the NLP model bridges the gap between data and understanding. It offers clinicians or users the ability to grasp the significance of the numbers, translating statistical findings into comprehensible narratives or descriptions. However, the reader can observe its inherent limitations in performing an in-depth analysis of the provided data. For example, the model failed to analyze and discuss the effect of the patient's gender on the ECG concept.

In Table 4, we have gathered all the DL medical-related state-of-the-art studies for comparison reasons. We can observe that the rest of the studies have reached efficient results with only image data as input, where CAD is demonstrated in SPECT-MPI format or in polar map images, where CNN training was applied. However, our proposed DeepFCM framework, although it did not exceed the performance of the previous literature studies in terms of accuracy, offers explainability and transparency of results, with the application of the FCM and the integration of clinical and imaging data, along with expert knowledge. By incorporating both imaging and clinical data, we demonstrate a holistic view of a patient's status, taking into account not only the image but also the medical history. The proposed model, DeepFCM, leverages expert knowledge as well, which enhances reliability.

Table 4. Comparison of previous related studies with the proposed framework DeepFCM.

Study	Input Data	DL Methods	Classification Problem	Results
Proposed DeepFCM	Polar maps + clinical	DeepFCM	CAD\No-CAD	Accuracy: 0.83
Papandrianos et al. [18]	SPECT	RGB-CNN (hand-crafted)	CAD\No-CAD	Accuracy: 0.93 ± 0.28 AUC: 0.936
Papandrianos et al. [2]	Polar maps	RGB-CNN (hand-crafted)	CAD\No-CAD	Accuracy: 0.92
Apostolopoulos et al. [20]	Polar maps	VGG-16	CAD\No-CAD	Accuracy: 0.74 Sensitivity: 0.75 Specificity: 0.73
Apostolopoulos et al. [50]	Polar maps + clinical	CNN (Inception V3) + Random Forest	CAD\No-CAD	Accuracy: 0.78 Sensitivity: 0.77 Specificity: 0.79
Spier et al. [21]	Polar maps	Graph CNN (hand-crafted)	CAD\No-CAD	Agreement rating (Segment-by-segment): 0.83, Sensitivity: 0.47, Specificity: 0.7
Berkaya et al. [51]	SPECT	VGG-19	CAD\No-CAD	Accuracy: 0.93 Sensitivity: 1.0 Specificity: 0.86
Jui-Jen Chen et al. [24]	Gray SPECT images	3D-CNN (hand-crafted)	CAD\No-CAD	Accuracy: 0.87 Sensitivity: 0.81 Specificity: 0.92
Liu et al. [52]	Stress-only SPECT	ResNet-34	CAD\No-CAD	AUC: 0.872 ± 0.002
Zahiri et al. [47]	Polar maps	CNN	CAD\No-CAD	Accuracy: 0.7562 Sensitivity: 0.7856 Specificity: 0.7434 F1 score: 0.6646 AUC: 0.8450
Arvidsson et al. [53]	Polar maps + clinical (angina symptoms, age)	CNN	Probability of CAD in the left anterior artery, left circumflex artery, and right coronary artery	Per-vessel AUC: 0.89 Per-patient AUC: 0.95

The research presents some noteworthy limitations that warrant discussion. First and foremost, this study's dataset, while sufficiently large to facilitate the experiments conducted, is derived exclusively from a single hospital. This mono-centric data source inherently introduces limitations in terms of its representativeness and generalizability. The healthcare landscape can exhibit considerable regional variations in patient demographics, treatment protocols, and disease prevalence. Consequently, relying solely on data from one hospital can potentially skew the findings and limit their applicability to a broader population.

Another critical aspect that warrants further investigation pertains to the efficiency and effectiveness of the CNN model, a pivotal component of the DeepFCM framework. While the integration of CNNs in FCMs holds promise for a wide range of applications, including medical image analysis and decision support systems, there remains a need for in-depth research to assess and optimize the performance of this fusion.

The DeepFCM framework presents immediate potential for revolutionizing the routine diagnosis of CAD in clinical practice. Its unique combination of explainability and commendable classification accuracy makes it a compelling tool for healthcare professionals. DeepFCM not only offers robust classification accuracy but also provides transparent and understandable insights into its decision-making process, which is vital for building trust and facilitating the adoption of AI-driven solutions in healthcare. With the ability to pinpoint the critical factors contributing to CAD diagnosis, doctors can make more informed decisions and tailor treatment plans more effectively. This immediate potential signifies a

transformative step forward in enhancing the accuracy and reliability of CAD diagnosis, ultimately improving patient outcomes and streamlining everyday clinical routines.

5. Conclusions

Our research introduced the DeepFCM model, a pioneering approach that seamlessly integrates imaging and clinical data to enhance the diagnosis of CAD. Beyond its diagnostic capabilities, a defining feature of DeepFCM is its emphasis on explainability and transparency. By incorporating feature selection, we ensured that only the most pertinent clinical data influenced the outcomes. Additionally, the development of the RGB–CNN model showcased the potential of CNNs in this domain, with the integration of Grad–CAM further boosting the interpretability of these networks. In terms of quantitative results, DeepFCM achieved an accuracy of 83.07%, a sensitivity of 86.21%, and a specificity of 79.99%. These metrics not only matched but exceeded the diagnostic accuracy of expert evaluations and the standalone RGB–CNN model. Recognizing the importance of clear communication in healthcare, we incorporated the GPT–3.5 NLP model to translate DeepFCM’s predictions into comprehensible explanations for medical professionals. This step is crucial in fostering trust and understanding between computational models and the medical community. This work represents a significant stride forward in CAD diagnosis, merging clinical and imaging data with advanced computational techniques. The added layer of explainability serves to bridge the gap between complex algorithms and clinical practice, enhancing trust among nuclear medicine experts.

Author Contributions: Conceptualization, A.F., I.D.A., S.M., and N.P. (Nikolaos Papathanasiou); methodology, I.D.A., A.F., N.P. (Nikolaos Papandrianos), S.M., and E.I.P.; software, A.F. and I.D.A.; validation, N.P. (Nikolaos Papathanasiou), D.A., and N.P. (Nikolaos Papandrianos); formal analysis, N.P. (Nikolaos Papandrianos) and A.F.; investigation, I.D.A. and A.F.; resources, I.D.A., N.P. (Nikolaos Papathanasiou), and N.P. (Nikolaos Papathanasiou); data curation, I.D.A., N.P. (Nikolaos Papathanasiou), and N.P. (Nikolaos Papandrianos); writing—original draft preparation, A.F. and I.D.A.; writing—review and editing, N.P. (Nikolaos Papathanasiou), E.I.P., and S.M.; visualization, A.F.; supervision, N.P. (Nikolaos Papandrianos), E.I.P. and S.M.; project administration, E.I.P. All authors have read and agreed to the published version of the manuscript.

Funding: The research project was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “2nd Call for H.F.R.I. Research Projects to support Faculty Members & Researchers” (Project Number: 3656).

Institutional Review Board Statement: This research does not report human experimentation; it does not involve human participants following experimentation. All procedures in this study were in accordance with the Declaration of Helsinki.

Informed Consent Statement: This study was approved on the 3rd of March 2022 by the ethical committee of the University General Hospital of Patras (Ethical & Research Committee of University Hospital of Patras—protocol number 108/10-3-2022). The requirement to obtain informed consent was waived by the director of the diagnostic center due to its retrospective nature.

Data Availability Statement: The datasets analyzed during the current study are available from the nuclear medicine physician upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Domingues, I.; Pereira, G.; Martins, P.; Duarte, H.; Santos, J.; Abreu, P.H. Using Deep Learning Techniques in Medical Imaging: A Systematic Review of Applications on CT and PET. *Artif. Intell. Rev.* **2020**, *53*, 4093–4160. [[CrossRef](#)]
2. Papandrianos, N.I.; Apostolopoulos, I.D.; Feleki, A.; Apostolopoulos, D.J.; Papageorgiou, E.I. Deep Learning Exploration for SPECT MPI Polar Map Images Classification in Coronary Artery Disease. *Ann. Nucl. Med.* **2022**, *36*, 823–833. [[CrossRef](#)] [[PubMed](#)]
3. Poon, A.I.F.; Sung, J.J.Y. Opening the Black Box of AI-Medicine. *J. Gastroenterol. Hepatol.* **2021**, *36*, 581–584. [[CrossRef](#)] [[PubMed](#)]
4. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [[CrossRef](#)]

5. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.* **2023**, *55*, 1–33. [[CrossRef](#)]
6. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2020**, *23*, 18. [[CrossRef](#)]
7. Papandrianos, N.I.; Feleki, A.; Moustakidis, S.; Papageorgiou, E.I.; Apostolopoulos, I.D.; Apostolopoulos, D.J. An Explainable Classification Method of SPECT Myocardial Perfusion Images in Nuclear Cardiology Using Deep Learning and Grad-CAM. *Appl. Sci.* **2022**, *12*, 7592. [[CrossRef](#)]
8. Akella, A.; Akella, S. Machine Learning Algorithms for Predicting Coronary Artery Disease: Efforts toward an Open Source Solution. *Future Sci. OA* **2021**, *7*, FSO698. [[CrossRef](#)] [[PubMed](#)]
9. Teng, Q.; Liu, Z.; Song, Y.; Han, K.; Lu, Y. A Survey on the Interpretability of Deep Learning in Medical Diagnosis. *Multimed. Syst.* **2022**, *28*, 2335–2355. [[CrossRef](#)]
10. Kosko, B. Fuzzy Cognitive Maps. *Int. J. Man-Mach. Stud.* **1986**, *24*, 65–75. [[CrossRef](#)]
11. Kosko, B. *Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence*, Prentice-Hall International editions; Prentice-Hall: Englewood Cliffs, NJ, USA, 1992; ISBN 978-0-13-612334-7.
12. Khodadadi, M.; Shayanfar, H.; Maghooli, K.; Hooshang Mazinan, A. Fuzzy Cognitive Map Based Approach for Determining the Risk of Ischemic Stroke. *IET Syst. Biol.* **2019**, *13*, 297–304. [[CrossRef](#)] [[PubMed](#)]
13. Apostolopoulos, I.D.; Groumpas, P.P.; Apostolopoulos, D.I. State Space Advanced Fuzzy Cognitive Map Approach for Automatic and Non Invasive Diagnosis of Coronary Artery Disease. *Biomed. Phys. Eng. Express* **2021**, *7*, 045007. [[CrossRef](#)]
14. Apostolopoulos, I.D.; Groumpas, P.P. Non-Invasive Modelling Methodology for the Diagnosis of Coronary Artery Disease Using Fuzzy Cognitive Maps. *Comput. Methods Biomed. Biotech. Engin.* **2020**, *23*, 879–887. [[CrossRef](#)] [[PubMed](#)]
15. Sovatzidi, G.; Vasilakakis, M.D.; Iakovidis, D.K. Fuzzy Cognitive Maps for Interpretable Image-Based Classification. In Proceedings of the 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Padua, Italy, 18–23 July 2022; pp. 1–6.
16. Sovatzidi, G.; Vasilakakis, M.D.; Iakovidis, D.K. IF3: An Interpretable Feature Fusion Framework for Lesion Risk Assessment Based on Auto-Constructed Fuzzy Cognitive Maps. In *Cancer Prevention Through Early Detection*; Ali, S., van der Sommen, F., Papiež, B.W., van Eijnatten, M., Jin, Y., Kolenbrander, I., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 77–86.
17. Sovatzidi, G.; Vasilakakis, M.D.; Iakovidis, D.K. Automatic Fuzzy Graph Construction For Interpretable Image Classification. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 3743–3747.
18. Papandrianos, N.; Papageorgiou, E. Automatic Diagnosis of Coronary Artery Disease in SPECT Myocardial Perfusion Imaging Employing Deep Learning. *Appl. Sci.* **2021**, *11*, 6362. [[CrossRef](#)]
19. Papandrianos, N.I.; Feleki, A.; Papageorgiou, E.I.; Martini, C. Deep Learning-Based Automated Diagnosis for Coronary Artery Disease Using SPECT-MPI Images. *J. Clin. Med.* **2022**, *11*, 3918. [[CrossRef](#)]
20. Apostolopoulos, I.D.; Papathanasiou, N.D.; Spyridonidis, T.; Apostolopoulos, D.J. Automatic Characterization of Myocardial Perfusion Imaging Polar Maps Employing Deep Learning and Data Augmentation. *Hell. J. Nucl. Med.* **2020**, *23*, 125–132. [[CrossRef](#)]
21. Spier, N.; Nekolla, S.; Rupprecht, C.; Mustafa, M.; Navab, N.; Baust, M. Classification of Polar Maps from Cardiac Perfusion Imaging with Graph-Convolutional Neural Networks. *Sci. Rep.* **2019**, *9*, 7569. [[CrossRef](#)] [[PubMed](#)]
22. Otaki, Y.; Tamarappoo, B.; Singh, A.; Sharir, T.; Hu, L.-H.; Gransar, H. Diagnostic accuracy of deep learning for myocardial perfusion imaging in men and women with a high-efficiency parallel-hole-collimated cadmium-zinc-telluride camera: Multicenter study. *J. Nucl. Med. Soc. Nucl. Med.* **2020**, *61*, 92.
23. Otaki, Y.; Singh, A.; Kavanagh, P.; Miller, R.J.H.; Parekh, T.; Tamarappoo, B.K.; Sharir, T.; Einstein, A.J.; Fish, M.B.; Ruddy, T.D.; et al. Clinical Deployment of Explainable Artificial Intelligence of SPECT for Diagnosis of Coronary Artery Disease. *JACC Cardiovasc. Imaging* **2022**, *15*, 1091–1102. [[CrossRef](#)] [[PubMed](#)]
24. Chen, J.J.; Su, T.Y.; Chen, W.S.; Chang, Y.H.; Lu, H.H.S. Convolutional Neural Network in the Evaluation of Myocardial Ischemia from CzT Spect Myocardial Perfusion Imaging: Comparison to Automated Quantification. *Appl. Sci. Switz.* **2021**, *11*, 514. [[CrossRef](#)]
25. Suganyadevi, S.; Seethalakshmi, V.; Balasamy, K. A Review on Deep Learning in Medical Image Analysis. *Int. J. Multimed. Inf. Retr.* **2022**, *11*, 19–38. [[CrossRef](#)]
26. Samaras, A.-D.; Moustakidis, S.; Apostolopoulos, I.D.; Papandrianos, N.; Papageorgiou, E. Classification Models for Assessing Coronary Artery Disease Instances Using Clinical and Biometric Data: An Explainable Man-in-the-Loop Approach. *Sci. Rep.* **2023**, *13*, 6668. [[CrossRef](#)]
27. Ghosh, P.; Azam, S.; Jonkman, M.; Karim, A.; Shamrat, F.M.; Ignatious, E.; Shultana, S.; Beeravolu, A.; De Boer, F. Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques. *IEEE Access* **2021**, *9*, 19304–19326. [[CrossRef](#)]
28. Nápoles, G.; Ranković, N.; Salgueiro, Y. On the Interpretability of Fuzzy Cognitive Maps. *Knowl.-Based Syst.* **2023**, *281*, 111078. [[CrossRef](#)]
29. Jastrzebska, A.; Nápoles, G.; Homenda, W.; Vanhoof, K. Fuzzy Cognitive Map-Driven Comprehensive Time-Series Classification. *IEEE Trans. Cybern.* **2023**, *53*, 1348–1359. [[CrossRef](#)]

30. Manimegalai, P.; Suresh Kumar, R.; Valsalan, P.; Dhanagopal, R.; Vasanth Raj, P.T.; Christhudass, J. 3D Convolutional Neural Network Framework with Deep Learning for Nuclear Medicine. *Scanning* **2022**, *2022*, 9640177. [CrossRef] [PubMed]
31. Wang, P.; Qiao, J.; Liu, N. An Improved Convolutional Neural Network-Based Scene Image Recognition Method. *Comput. Intell. Neurosci.* **2022**, *2022*, 3464984. [CrossRef] [PubMed]
32. Oh, J.W.; Jeong, J. Data Augmentation for Bearing Fault Detection with a Light Weight CNN. *Procedia Comput. Sci.* **2020**, *175*, 72–79. [CrossRef]
33. Zadeh, L.A. Fuzzy Sets. *Inf. Control* **1965**, *8*, 338–353. [CrossRef]
34. Raharja, M.A.; Darmawan, I.D.M.B.A.; Nilakusumawati, D.P.E.; Supriana, I.W. Analysis of Membership Function in Implementation of Adaptive Neuro Fuzzy Inference System (ANFIS) Method for Inflation Prediction. *J. Phys. Conf. Ser.* **2021**, *1722*, 012005. [CrossRef]
35. Kreinovich, V.; Kosheleva, O.; Shahbazova, S. Why Triangular and Trapezoid Membership Functions: A Simple Explanation. In *Recent Developments in Fuzzy Logic and Fuzzy Sets: Dedicated to Lotfi A*; Springer: Vienna, Austria, 2020; pp. 25–31. ISBN 978-3-030-38892-8.
36. Han, F.; Chen, W.-T.; Ling, Q.-H.; Han, H. Multi-Objective Particle Swarm Optimization with Adaptive Strategies for Feature Selection. *Swarm Evol. Comput.* **2021**, *62*, 100847. [CrossRef]
37. Yi, J.; Ran, Y.; Yang, G. Particle Swarm Optimization-Based Approach for Optic Disc Segmentation. *Entropy Basel Switz.* **2022**, *24*, 796. [CrossRef]
38. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [CrossRef]
39. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large Language Models in Medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [CrossRef] [PubMed]
40. van Dis, E.A.M.; Bollen, J.; Zuidema, W.; van Rooij, R.; Bockting, C.L. ChatGPT: Five Priorities for Research. *Nature* **2023**, *614*, 224–226. [CrossRef] [PubMed]
41. Nath, S.; Marie, A.; Ellershaw, S.; Korot, E.; Keane, P.A. New Meaning for NLP: The Trials and Tribulations of Natural Language Processing with GPT-3 in Ophthalmology. *Br. J. Ophthalmol.* **2022**, *106*, 889–892. [CrossRef]
42. Currie, G.; Robbie, S.; Tually, P. ChatGPT and Patient Information in Nuclear Medicine: GPT-3.5 Versus GPT-4. *J. Nucl. Med. Technol.* **2023**, *51*, 165–166. [CrossRef] [PubMed]
43. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
44. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
45. Zhang, Y.; Hong, D.; McClement, D.; Oladosu, O.; Pridham, G.; Slaney, G. Grad-CAM Helps Interpret the Deep Learning Models Trained to Classify Multiple Sclerosis Types Using Clinical Brain Magnetic Resonance Imaging. *J. Neurosci. Methods* **2021**, *353*, 109098. [CrossRef] [PubMed]
46. Jahmunah, V.; Ng, E.Y.K.; Tan, R.-S.; Oh, S.L.; Acharya, U.R. Explainable Detection of Myocardial Infarction Using Deep Learning Models with Grad-CAM Technique on ECG Signals. *Comput. Biol. Med.* **2022**, *146*, 105550. [CrossRef] [PubMed]
47. Zahiri, N.; Asgari, R.; Razavi-Ratki, S.-K.; Parach, A.-A. Deep Learning Analysis of Polar Maps from SPECT Myocardial Perfusion Imaging for Prediction of Coronary Artery Disease. *Res. Sq.* **2021**, preprint. [CrossRef]
48. Heckel, R.; Yilmaz, F.F. Early Stopping in Deep Networks: Double Descent and How to Eliminate It. *arXiv* **2020**, arXiv:2007.10099.
49. Nguyen, P.K.; Nag, D.; Wu, J.C. Sex Differences in the Diagnostic Evaluation of Coronary Artery Disease. *J. Nucl. Cardiol. Off. Publ. Am. Soc. Nucl. Cardiol.* **2011**, *18*, 144–152. [CrossRef] [PubMed]
50. Apostolopoulos, I.D.; Apostolopoulos, D.I.; Spyridonidis, T.I.; Papathanasiou, N.D.; Panayiotakis, G.S. Multi-Input Deep Learning Approach for Cardiovascular Disease Diagnosis Using Myocardial Perfusion Imaging and Clinical Data. *Phys. Medica PM Int. J. Devoted Appl. Phys. Med. Biol. Off. J. Ital. Assoc. Biomed. Phys. AIFB* **2021**, *84*, 168–177. [CrossRef] [PubMed]
51. Kaplan Berkaya, S.; Ak Sivrikoz, I.; Gunal, S. Classification Models for SPECT Myocardial Perfusion Imaging. *Comput. Biol. Med.* **2020**, *123*, 103893. [CrossRef]
52. Liu, H.; Wu, J.; Miller, E.J.; Liu, C.; Liu, Y.; Liu, C.; Liu, Y.-H. Diagnostic Accuracy of Stress-Only Myocardial Perfusion SPECT Improved by Deep Learning. *Eur. J. Nucl. Med. Mol. Imaging* **2021**, *48*, 2793–2800. [CrossRef] [PubMed]
53. Arvidsson, I.; Overgaard, N.C.; Aström, K.; Heyden, A.; Figueroa, M.O.; Rose, J.F.; Davidsson, A. Prediction of Obstructive Coronary Artery Disease from Myocardial Perfusion Scintigraphy Using Deep Neural Networks. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4442–4449. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.