Iraqi Journal for Computer Science and Mathematics

Manuscript 1276

A Novel Benchmarking Framework for Selecting the Best Deep Learning Model Diagnosing COVID-19 Based on New Development for Dual MCDM Methods

Mahmood M. Salih

Yousif Raad Muhsen

M.A. Ahmed

Reem D. Ismael

Moceheb Lazam Shuwandy

See next page for additional authors

A Novel Benchmarking Framework for Selecting the Best Deep Learning Model Diagnosing COVID-19 Based on New Development for Dual MCDM Methods

Authors

Mahmood M. Salih, Yousif Raad Muhsen, M.A. Ahmed, Reem D. Ismael, Moceheb Lazam Shuwandy, and Z.T. Al-qaysi

IRAQI JOURNAL FOR COMPUTER SCIENCE AND MATHEMATICS 2025;6:18–40

ORIGINAL STUDY

A Novel Benchmarking Framework for Selecting the Best Deep Learning Model Diagnosing COVID-19 Based on New Development for Dual MCDM Methods

Mahmood M. Salih [a,*], **Yousif Raad Muhsen** [b,c], **M.A. Ahmed** [a], **Reem D. Ismael** [a], **Moceheb Lazam Shuwandy** [a], **Z.T. Al-qaysi** [a]

[a] Computer Science Department, College of Computer Science and Mathematics, Tikrit University (TU), Tikrit, Iraq
[b] College of Computer Science and Information Technology, Wasit University, Wasit, Iraq
[c] Technical Engineering College, Al-Ayen University, Thi-Qar, Iraq

ABSTRACT

COVID-19 was diagnosed using deep learning models by a group of studies. Evaluating and benchmarking these models are essential to achieving the most suitable model for diagnosing coronavirus. **Objective:** In this investigation, we offer an inclusive valuation of several deep learning models to detect the maximum appropriate and active model which gratifies doctors' requirements and assessment criteria. **Method:** This study combines Fuzzy decision by the opinion score method (FDOSM) and Fuzzy-Weighted Zero-Inconsistency (FWZIC). According to the advantage of Trapezoidal Intuitionistic fuzzy, we developed FWZIC into Trapezoidal Intuitionistic fuzzy named (TrIF-FWZIC) for weighting criteria and FDOSM into Trapezoidal Intuitionistic fuzzy FDOSM (TrIF-FDOSM) to evaluate and benchmark the effectively deep learning models and tackle the issue of uncertainty. Fundamentally, the methodology of this study is presented in 2 phases; the 1st phase is related to identifying a new decision matrix containing 24 evaluation criteria to evaluate the ten deep learning models. Furthermore, the 2nd phase is related to the development of TrIF-FWZIC and TrIF-FDOSM in two main stages. **Result:** The findings of this study were: (1) For the individual decision-maker, the best one was Xception for the first decision-maker with a score (i.e., 0.267510407). The optimal algorithm for the 2nd and 3rd decision-makers was ResNet-101 with scores (i.e., 0.316710828, 0.457770263), respectively. (2) The best deep learning model, depending on the group decision-making, was ResNet-101 with a score (i.e., 0.32574743). **Conclusion:** The proposed methodology undergoes validation, sensitivity analysis, and comparative evaluation. This research enhances the selection of effective models for COVID-19 diagnosis, catering to individual and collective decision-making scenarios.

Keywords: Deep learning, COVID-19, Evaluation, MCDM, FDOSM, FWZIC, Trapezoidal Intuitionistic fuzzy

1. Introduction

1.1. Motivation

Amidst an unparalleled worldwide health emergency, the scientific community exhibits a collective resolve to confront the challenges posed by the COVID-19 pandemic [1]. Deep learning algorithms have arisen as a promising avenue in the quest to change the diagnosis of COVID-19 [2, 3]. The valuation of these algorithms is of immense importance since it has the potential to reshape the healthcare domain and have a profound impact on mortality rates. The assessment of deep learning algorithms for COVID-19 serves as a connection between advanced

Received 4 December 2024; accepted 17 May 2025.
Available online 2 July 2025

* Corresponding author.
E-mail addresses: mahmaher1989@gmail.com (M. M. Salih), yousif@uowasit.edu.iq (Y. R. Muhsen), mohamed.aktham3@gmail.com (M. A. Ahmed), reema.alnasery@gmail.com (R. D. Ismael), mocheb@gmail.com (M. L. Shuwandy).

https://doi.org/10.52866/2788-7421.1276

technological advancements and practical medical implementations [4]. The algorithms, propelled by the potential of artificial intelligence, have exhibited noteworthy proficiency in scrutinising X-rays, CT scans, and molecular information. The ability of these diagnostic methods to swiftly and precisely identify instances of COVID-19 has the potential to greatly accelerate the process of diagnosis, facilitate timely medical intervention, and diminish the spread of the virus. Nevertheless, it is imperative to thoroughly assess the efficacy and dependability of these algorithms, as their mere development alone is insufficient [5]. Finally, the evaluation of these algorithms for COVID-19 analysis represents a significant juncture at the intersection of technology and medicine.

1.2. Problem statement and challenges

In disease detection using deep learning models, making the appropriate choice of an algorithm and assessing its suitability raises many complicated issues. The process of finding the optimal model for precise disease detection is complicated by the existence of a multitude of different deep models [6, 7]. The fact that no model overrules all others makes the situation quite complicated for making cheese-related choices. Also, remember that the computational ability of processes in these models can be complex, resulting in diagnostic procedures taking longer [8]. The procedure of assessing and comparing these models is made more complex by different criteria for appraisal and the fundamental disagreements that arise between them [7, 9]. Evaluation of deep learning models demands much consideration because of their possible impact on patients, healthcare providers, and healthcare institutions [10]. The deep models for autonomous disease diagnosis evaluation require conforming to several parameters, which will encapsulate varied evaluation criteria [11]. The field of deep learning often uses a range of criteria, including accuracy, F1-score, error rate, precision, true positive (TP), false positive (FP), false negative (FN), true negative (TN), sensitivity, specificity, p-value, and the Matthew Correlation Coefficient (MCC) [12]. From this study, it is quite evident that there is a distinct limitation in which the best automated disease detecting systems are currently evaluated and selected. In this particular context, there are three critical areas in which detailed attention should be taken: the wide range of assessment factors, the varied weightiness of each factor, and the probability of disagreement in case the factors overlap [13–15]. The evaluation of deep learning algorithms adds other complications as mentioned by the author in the cited scholarly work [7]. Since automated diagnostic criteria possess complexity and diversity, a detailed review entailing several angles is needed. The application of MCDM techniques would deliver a more optimal solution to addressing these problems and the process of choosing deep learning models [16, 17]. Despite this, the vagueness that arises in the course of making decisions severely influences the validity and trustworthiness of MCDM approaches [18]. MCDM could face uncertainty due to poorly defined criterion weights, vague preferences of the decision-makers, and data or information drought [19–21]. Within the described impediments, choosing the most preferred option is challenging due to the unpredictability or imprecision of expected outcomes. Also, the selection of the proper MCDM method can affect the final decision, as different methods have different mechanisms of dealing with ambiguity [15, 22, 23]. Dependable and durable outputs of the doubt management process in MCDM techniques are contingent upon careful analysis of decision frameworks and multiple grounds for uncertainties.

1.3. Research gap and contribution

In the field of fuzzy systems and their applications, FWZIC [22, 24] has taken on a wide range of fuzzy set (FS) related contexts. The FS environments include interval type-2 trapezoidal FSs, q-restricted Order Fuzzy Sets (q-ROFSs), Pythagorean Fuzzy Sets (PFSs), T-spherical FSs, and neutrosophic FSs [19, 25–27]. Previous researchers have made significant attempts to iron out ambiguity and uncertainty with these extensions, but several questions remain outstanding. This paper highlights a key limitation, one common in the theoretical approach and the practice of effectiveness. Up to date, none of the FDOSM nor the FWZIC approach has been implemented in the field of trapezoidal Intuitionistic fuzzy (TrIF). By combining FDOSM and FWZIC techniques, we introduce a process to make an effective integration of collective expertise brought together by the domain experts in working out the complex decisions. By making integration of the suggested kind, we enhance our capabilities in analyzing complex healthcare problems and provide sturdily thinking solutions, especially in detecting the most and least productive deep learning models in terms of diagnostics of COVID-19. This integration allows us to transcend the theory and concentrate on the possibility of implementing these methods in practice. Embedding TrIF in the FDOSM-FWZIC structure enables the analysis of various approaches with regard to uncertain decision scenarios, the final goal being deep. This research will develop a deeper understanding of the uses of deep learning models in diagnosing COVID-19. This

development can resolve long-standing issues and uncertainties of choosing the best-performing models, as well as show patterns in less effective models.

1.4. Objectives

The core goal line of this investigation is to achieve the following objectives:

1. Finding the decision matrix based on 24 criteria and 10 alternatives.
2. Providing a novel approach for enhancing the dependability of ranking deep learning modules in a fuzzy environment by integrating the TrIF fuzzy set into the existing FDOSM framework.
3. To enhance FWZIC by integrating the TrIF fuzzy set to offer a trustworthy weighting of criterion in a situation of ambiguity.
4. To enhance choosing the best deep learning module in a cooperative fuzzy setting, TrIF-FDOSM's capabilities are extended by incorporating a group of experts.
5. To verify the outcome of TrIF-FDOSM & TrIF-FWZIC.

1.5. Significance and implications

1.5.1. Why work with TrIF sets?

Our everyday lives are populated with many events and obstacles that contain high levels of uncertainty [28]. The implementation of TrIF sets for MCDM is high-impact with far-reaching positive consequences. The inclusion of TrIF sets in FDOSM and FWZIC allows for the processing of challenging problems and the best prediction of decision-making under conditions of uncertainty. Next, the discussion is engaged about the role and implications of the integration of TrIF with MCDM.

First, TrIFs provide a systematic method for describing the uncertainty and vagueness that surround a decision-making environment. TrIF improves the structure of standard fuzzy sets by adding membership, non-membership, and hesitancy measures [29]. In so doing, it gives a richer representation of authoritative perspectives and relevant situations, which reflects the inherent uncertainty in judging algorithms for diagnosing COVID-19. Furthermore, MCDM approaches are great for addressing such decisions, which are hampered by several conflicting factors [30]. The ability of TrIF to express uncertainty in a fine-grained sense enables better decision-making. With the help of correct estimates of the importance of criteria, decision-makers can make sound and dispassionate choices in their selection of the best algorithm for COVID-19 diagnosis. Additionally, the TrIF system allows for effortless utilization of expert insights and opinions. Specialists offer assessments that represent varied understandings of confidence, prudence, and disbelief [31]. The TrIF approach fosters the consolidation of diverse opinions, enabling the experts to present better-defined and wider analyses of model performance. Further, introducing TrIF in the process with CDM can mitigate the bias that is made during decision-making. The existing methods of evaluating data do not often cover the complex uncertainties that occur in real datasets [32]. The evaluation of deep learning models for COVID-19 diagnosis requires attention to a number of baroque criteria, mostly conflicting with one another. Having both membership and non-membership degrees, TrIF is outstandingly suitable for challenging assessment settings, providing a comprehensive and suitable evaluation schema. The research thus applied TrIF in attempting to clarify ambiguity.

1.5.2. Why utilise FWZIC?

There are two main classes of MCDM methods: The weighing methods, namely AHP, WASPA, and FWZIC [33]. In addition, multiple methods for ranking options, for example EDAS, FDOSM, and VIKOR [34]. The FWZIC method, which is new and advanced, is applied to allocate weights to criteria without causing inconsistency [22]. Applying FWZIC to assign weights to criteria is both effective and provides important benefits, making it attractive to decision-makers [35]. FWZIC is particularly effective at controlling inconsistencies that commonly occur while making decisions, a problem that can greatly undermine decision accuracy and reliability. Moreover, FWZIC represents a means to efficiently handle the task of assigning weightings to criteria [36]. FWZIC is unique in that it avoids asking decision-makers to compare criteria directly, making the entire decision process more time-efficient. Accordingly, decision-makers may allocate their resources efficiently and emphasize additional important steps in the decision-making flow. Because the stability of criteria in FWZIC is usually low, recalculations are often not required when the criteria set changes [22]. The way input is collected from experts in FWZIC is also relatively unproblematic. Besides, the repeated nature of the FWZIC method contributes to the coherence among final weights, which fundamentally improves the quality and dependability of decision outcomes [13]. This suggests that decision-makers can rely more on the final selection, given the precise and consistent assessment of all components.

1.5.3. Why choose FDOSM?

Considering the conditions of consistency, comparisons, uncertainty, and ambiguity embedded in

most MCDM methods, the academic literature has proposed several approaches to relax these limits [37–39]. Some of these methods named FDOSM and FWZIC method. While the FDOSM framework was provided by Salih in 2020 [25], it is an intelligent MCDM procedure that is sequential, fit of three stages of data input, transformation, as well as processing. Through the FDOSM approach, one can see a reasonable and effective methodology in an unclear context. The core benefit of the FDOSM is that it is able to support effective control over fuzzy data and ambiguous information. When it comes to complex decision-making environments with imperfect or inaccurate data, this trait turns decidedly beneficial [17]. Moreover, the approach to FDOSM ensures that the decision-makers do not have contradicting or competing preferences during the evaluation of the options [40]. The FDOSM method is the concept of ideal solutions on which decision-makers can determine the best option to determine and then compare its relation with other values using the same parameters. This methodology supports the decision-makers to make informed decisions and prioritize possibilities well. Furthermore, FDOSM proves to be applicable in individual and group decision-making situations, thus providing an opportunity for it to be applied to varied decision-making situations. The integration of FDOSM with uncertain methodologies and conflicting criteria has shown inspiring results in the previous studies [20, 26, 40–44].

2. Literature review

The literature review has 2 parts. In Section 2.1, we present a summary of work done on FDOSM and FWZIC. The MCDM studies of covid-19 are summarised in Section 2.2.

2.1. Previous research about FDOSM and FWZIC

FDOSM and FWZIC are recent MCDM methodologies used for the purposes of ranking and weighting, respectively [25]. Several significant academic publications have utilised them as a viable substitute for conventional MCDM methodologies, resulting in their adoption in several research investigations. FDOSM and FWZIC were developed using fuzzy triangular sets as their foundation. One limitation of TFN is its ability to effectively handle situations characterised by ambiguity and uncertainty [41]. When individuals are faced with practical difficulties, complexity arises from the inherent lack of clarity, unpredictability, and multiple interpretations, which in turn intensify the complexities of making decisions [35, 45]. There is a need for further development and enhancement of FDOSM and FWZIC in order to tackle the complexities associated with uncertainty effectively. It is also vital to extend such models to a new fuzzy version that will be able to consider uncertain and ambiguous cases, whilst making it possible to gain additional valuable information. As an assessment of the benchmarking process of the active queue management methodologies, the authors [46] carried out an assessment in their study. As a part of this evaluation, the construct of a multi-criteria decision-making problem that addressed multidimensional criteria was analysed. To upgrade FDOSM, the authors included four diverse strategies of aggregation, which are widely used in the direct agglomeration methodology. Moreover, the researchers have extended the FDOSM and FWZIC methodologies in T-spherical domain while considering the constructiveness of T-spherical fuzzy sets in dealing with information uncertainty. This adaptation was done with the objective of using these techniques in the distribution of COVID-19 vaccinations [26]. In addition, the choice of extending the usage of FDOSM and FWZIC within the framework of q-ROFS was based on its advantages and applicability prospects. The main aim was to integrate the use of q-RO_FDOSM and q-RO_FWZIC approaches so as to accomplish equitable distributions of COVID-19 immunization doses [40]. Additionally, in the year 2022, a study carried out by Alqaysi et al. [35] introduced the method of the FDOSM and FWZIC as a primary evaluation methodology for determining the ideal hybrid diagnostic models that could be used in classifying patients with autism. Subsequently, in light of the advantages associated with Cubic Pythagorean fuzzy sets, a recently developed sophisticated fuzzy framework has been introduced to address the challenge of uncertainty. Alamoodi et al. [13] proposed the incorporation of FWZIC and FDOSM as potential solutions for addressing the challenges associated with Sign language. In addition, the authors in reference [47] combined Spherical fuzzy rough sets with FDOSM and FWZIC to effectively identify the ideal smart e-tourism application using several factors. Moreover, in reference [48], the authors used Fermatean probabilistic hesitant fuzzy sets and FDOSM techniques to handle uncertainty successfully and assess the effectiveness of supply chains in the agri-food sector. The palm oil industry is now facing intense competitiveness and sustainability issues. To address these concerns, the industry needs to follow established standards and incorporate sector 4.0 methodologies. This should be done within a framework of circular economies and sustainable practices, which is referred to as I4.0-in-a-CE-and-SPs. In order to find the best way to

use Trade 4.0 in the context of Circular Economics and Green Manufacturing, we employ the FWZIC approach along with an interval-valued Pythagorean fuzzy rough set. The applications are then evaluated and ranked using the EDAS approach [49].

The topic of ambiguity has been suggested by previous scholars to remain an ongoing concern that warrants more consideration. Furthermore, the authors suggested integrating several fuzzy set methodologies with FDOSM and FWZIC in order to investigate the efficacy of these approaches in addressing the challenge of uncertainty. However, it is noteworthy that none of these iterations has taken into account the TrIF fuzzy set despite its potential to handle uncertainty.

2.2. Studies on covid-19 with MCDM methods

Previous studies attempted to provide a solution for evaluating and benchmarking deep learning modules in health care to select the optimal system. The study is an attempt to tackle the rise of medical waste caused by the COVID-19 pandemic through a MCDM framework, which is supposed to identify optimal treatment techniques. Using Fuzzy Preference Selection Index, it assesses sixteen criteria spread across economic, technology, environmental, and social aspects, and it classifies nine disposal technologies using the Fuzzy CRADIS method. It was found that the most important criterion is disinfection efficiency, autoclaving has been identified as the best treatment method. A sensitivity analysis substantiated the approach's practicality and reliability [50]. The study [7] developed a method for evaluating medical diagnostic systems. The MCDM method was utilised to choose the optimal traditional classification model for COVID-19 diagnosis. Two approaches, the TOPSIS and the Entropy techniques, were utilised to benchmark the distinct COVID-19 diagnostic models. In an integrated MCDM method, TOPSIS was used for benchmarking and ranking, while the weights of the criterion were established using Entropy. The initial studies partially evaluated limited evaluation criteria to solve the automatic diagnosis system benchmarking problem. Hence, increasing evaluation criteria may be necessary to get the correct diagnosis systems ranking [51]. Previous studies did not consider the issue of developing a deep decision matrix including all performance metrics in terms of evaluation criteria. Providing a full solution to address these issues is necessary.

3. Methodology

This sector delivers a detailed explanation of the benchmarking solution for the COVID-19 deep learning techniques based on the proposed TrIF-FDOSM and TrIF-FWZIC. The first phase of the deep learning COVID-19 techniques decision matrix creation is defined in Section 3.1. The second-phase methodology is presented in Section 3.2, which describes the proposed TrIF-FWZIC and TrIF-FDOSM steps used to benchmark the deep learning COVID-19 methods. Fig. 1 explain the benchmarking approach for the COVID-19 deep learning algorithms.

3.1. Phase one: Construction decision matrix

In this section, explain the achievement of the decision matrix (DM) used in assessing and benchmarking the deep learning COVID-19 model configuration.

3.1.1. Identification of deep COVID-19 diagnosis models

There are many COVID-19 diagnosis models (deep learning) to be ranked based on experts and MCDM approaches. The present study examined many alternatives, namely COVID-19 diagnostic models, that were derived from detection models that were regularly employed in previous studies. These models utilised Artificial Intelligence to evaluate and analyse COVID-19 diagnoses [52–55]. Therefore, as a proof of concept, ten deep COVID-19 diagnosis models were chosen to be evaluated and benchmarked.

3.1.2. Defining the criterion for evaluation

The word "criteria" denotes the numerous metrics used to compare and contrast different options (i.e., deep COVID-19 diagnosis models). As part of this research, we looked at 24 different criteria [56, 57] as shown in Table 1. Table 1 shows the description of the criteria.

3.1.3. Diagnosis model development

The first stage in developing diagnostic models involves gathering X-ray images from reliable sources and pre-processing the data to be ready for deep models [61]. Seven datasets of Chest X-ray (CXR) images are openly available and used as our key data source. The dataset employed in this work is made up of CXR medical images for both COVID-infected and healthy subjects. The data shown in Table 2 was gathered by Dr. Joseph Cohan and placed on GitHub for researchers. The materials were created for patients who have acute respiratory distress syndrome, severe acute respiratory syndrome, or Middle East respiratory syndrome [62]. 340 images were obtained, consisting of AI-generated CT scans, frontal X-rays, and non-frontal CXRs. The research also used another public dataset of chest X-rays containing 55 images from infected COVID-19 patients. The third

Table 1. Clearly description of criteria used in this study.

Criteria	Formula	Description
Accuracy	$ACC = \frac{TP+TN}{TP+FP+FN+TN}$	This formula is used to determine the accuracy of classifiers by comparing their percentage accuracy to the real accuracy. The formula is based on the percentage of accurate predictions to the total number of input samples [58].
Recall	$Recall = \frac{TP}{TP+FN}$	This formula is used to compute the rate of True Positive, which is also called Recall and represents the total number of correct samples that have been identified with respect to all the positive representations. In this regard, recall represents the capability of the model to identify the infected patients correctly [59].
Precision	$Precision = \frac{TP}{TP+FP}$	The precision formula computes the quotient of accurately identified samples divided by the total number of samples. It functions as a measure of the classifier's efficacy in accurately removing unwanted topics.
F_Score	$F\_\text{score} = \frac{Two*TP}{2*TP+FP+FN}$	The F1 score formula is used to compute the weighted rate of sensitivity and precision. The optimal F1 value represented by the values reaches one, and the worst value is represented by 0. Comparatively, sensitivity and precision have similar effects on the rate of F1 score.
Specificity	$Specificity = \frac{TN}{TN+FP}$	This formula is used to compute the rate of TN samples while evaluating the classifier's performance in predicting true negative samples. In this regard for COVID-19, Specificity can classify un-infected people correctly [60].
False Positive Rate (FPR), or Fall-out	$FPR = \frac{FP}{(FP+TN)}$	This formula is used to compute the rate of incorrectly positively classified samples while evaluating the classifier with respect to the proportion of negative samples.
The False Negative	$FNR = \frac{FN}{TP+FN}$	This formula is used to compute the rate of miss classification while evaluating the classifier with respect to the positive rate missing.
The False Discovery Rate (FPR), or P value	$FDR = \frac{FP}{TP+FP}$	If (P >0.05), then the test hypothesis is false. If a P value is larger than 0.05, it suggests that there was no statistical significance to the study. A p-value is a statistical metric that quantifies the ratio of false-positive discoveries to the overall number of positive test outcomes.
Geometric Mean (GM)	$GM = \sqrt{(\frac{TP}{TP+FN}) \times (\frac{TN}{FP+TN})}$	GM is determined by calculating the product of the values in a set of numbers, representing the central tendency of the group.
Balanced Accuracy (BC)	$BC = \frac{((\frac{TP}{TP+FN})+(\frac{TN}{FP+TN}))}{2}$	It is calculated by taking the average of recollection obtained in each class. In this case, the poorest value is 0, and the best value is 1. Classification issues with unbalanced datasets are handled with balanced accuracy. A class's average recall is calculated by averaging the results of all students. False adjustment causes the best value to be 1, and the worst value to be zero.
Matthew Correlation Coefficient MCC	$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$	The MCC is a statistical method for evaluating models. MCC= 0 is the predicted result for the coin-tossing classifier, with t ranging from [1 + 1] to extreme values of –1 and +1 for perfect misclassification and excellent classification, respectively.
False omission rate (FOR)	$\text{FOR} = \frac{\text{FN}}{\text{FN}+\text{TN}} = 1 - \text{NPV}$	FOR is a statistical measure used to assess the accuracy of a binary classification model, particularly in the context of medical testing or diagnostic systems.
predicted positive condition rate (PPCR)	$\text{PPCR} = \frac{tp+fp}{tp+fp+tn+fn}$	Percentage of total population highlighted.
Positive likelihood ratio (LR+)	$\text{LR}+ = \frac{\Pr(T+\|D+)}{\Pr(T+\|D-)}$ $\text{LR}+ = \frac{\text{sensitivity}}{1-\text{ specificity}}$	Positive Likelihood Ratio or Sensitivity Index, is a statistical measure used to evaluate the performance of a diagnostic test or a binary classification model.
Negative likelihood ratio (LR−)	$\text{LR}- = \frac{\Pr(T-\|D+)}{\Pr(T-\|D-)}$ $\text{LR}- = \frac{1-\text{ sensitivity}}{\text{specificity}}$	LR-is also known as the negative likelihood ratio or specificity index.

(Continued)

Table 1. Continued

Criteria	Formula	Description
Diagnostic odds ratio	$\text{DOR} = \frac{TP/FN}{FP/TN} = \frac{TP/FP}{FN/TN} = \frac{TP \cdot TN}{FP \cdot FN}$	Measures the diagnostic test's efficacy. For example, if a test is positive, it means that the patient has some kind of illness, and it means that it's more likely than not that they don't.
Log loss	$\text{Loss} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \cdot \log_e(\hat{y}_i) + (1-y_i) \cdot \log_e(1-\hat{y}_i)]$	It is a critical probabilistic classification metric. Predict proba is based on probabilistic estimations and may be used to assess a classifier's probability outputs rather than discrete predictions. The smaller the log-loss number, the more accurate the forecasts will be.
Negative Predictive Value (NPV)	$\text{NPV} = \frac{\text{TN}}{\text{TN}+\text{FN}} = 1 - \text{FOR}$	You may find out how frequently a negative test is really a false negative using the negative predictive value. If a test comes out negative, how certain can we be that the patient does not have the disease? The extra sensitive a test is, the fewer likely a person with a negative test is to have the condition, and hence the larger the negative predictive value of the test.' More specialised tests have a lower negative predictive value and a higher positive predictive value because a person with a confident test is less likely to be disease-free.
kappa	$\kappa = \frac{2\times(TP\times TN - FN\times FP)}{(TP+FP)\times(FP+TN)+(TP+FN)\times(FN+TN)}$	Kappa is an intriguing measure to consider. Its roots may be traced back to psychology, where it is used to measure the agreement between two human raters (e.g., psychologists) while assessing topics (patients). In more recent times, the machine-learning community has "appropriated" it to assess categorisation performance.

Table 2. Description of selected datasets.

Dataset	Repo	Samples Samples	#Total	Selected Samples	#Selected	Dataset URL
X-Ray (Dr Joseph Cohan)	GitHub	COVID-19	340	COVID-19	260	"https://github.com/ieee8023/covid-chestxray-dataset"
X-Ray (Pneumonia)	Kaggle	PNEUMONIA NORMAL	4273 1406	NORMAL	669	"https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia"
Radiography Database	Kaggle	COVID-19 NORMAL Viral Pneumonia	219 1341 1345	COVID-19	166	"https://www.kaggle.com/tawsifurrahman/covid19-radiography-database"
X-Ray	Kaggle	COVID-19 NORMAL	174 174		128	"https://www.kaggle.com/fusicfenta/chest-xray-for-covid19-detection"
COVID-19 & Normal-poster anterior (PA) X-rays	Kaggle	COVID-19 NORMAL	140 140	COVID-19 NORMAL	90	"https://www.kaggle.com/tarandeep97/covid19-normal-posteroanteriorpa-xrays"
X-ray Dataset	GitHub	COVID-19	55	COVID-19		"https://github.com/agchung/Figure1-COVID-chestxray-dataset"
COVID-19 and Pneumonia Scans Dataset	Robofow	COVID-19 Healthy Viral Pneumonia	199 1965 3723	COVID-19 NORMAL Viral Pneumonia	114	"https://public.roboflow.ai/classification/covid-19-and-pneumonia-scans"
Xray Dataset	Kaggle	Normal COVID-19	94 94		94	"https://www.kaggle.com/khoongweihao/covid19-xray-dataset-train-test-sets/notebooks"

dataset belongs to the Kaggle repository and consists of 5679 two-class CXR images, which include healthy subjects and subjects infected with COVID-19. The first category in this dataset consists of 669 images, while the second category (for the infected patients) consists of 2905 images. The fifth COVID-19 medical dataset also belongs to the Kaggle repository and consists of 348 CXR images for infected and non-infected patients, with 174 CXR images for each category. The sixth medical dataset consists of 280 two-class CXR images for the infected and normal people. The final dataset belongs to the Robofow repository and

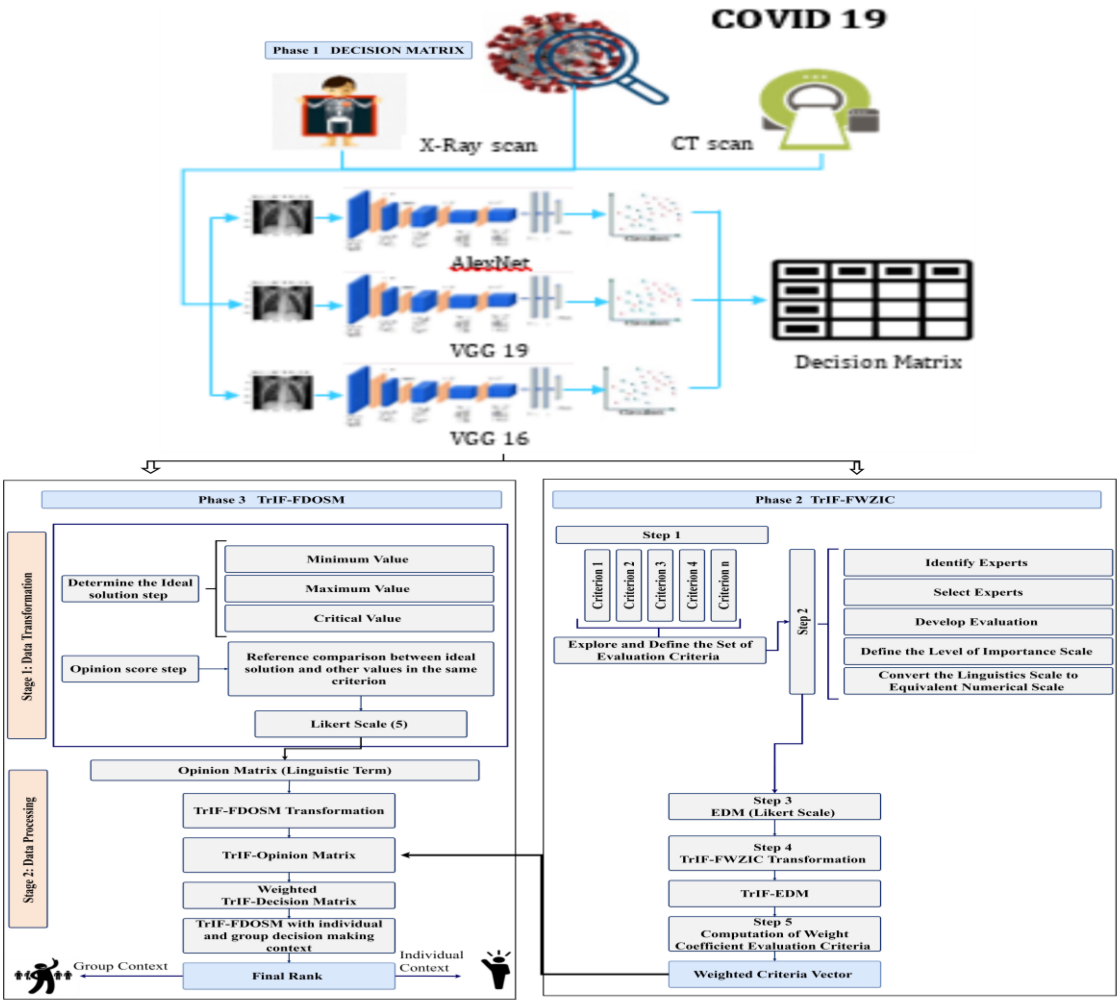

Fig. 1. Research methodology.

consists of 199 COVID-19 images and 1965 Healthy images. In Fig. 2 shows the samples of medical images for infected and normal people. In this study, we selected the frontal part of the X-ray images for the normal and the infected cases. CT-scan for normal and infected people was also selected. Additionally, both class labels are distributed equivalently over the whole dataset to solve the imbalance problems. As a result, the experimental dataset that has been used in this study consists of 669 for the normal and the abnormal cases, which compromises 1,338 for the whole dataset.

COVID-19 CXR pictures can be found on GitHub and Kaggle with resolutions between 508 * 500 and 4248 * 3480. Because of the constraints of the experimental setup, we reduced the image size to 150 * 150 pixels. Keras's "preprocess input" function performs a transformation on the input image to make it compatible with the model. We resize the input image using the Keras "preprocess input" function so that it meets the standard requirement for categorisation. We detailed 75% for training and 25% for testing.

Overfitting models were alleviated by using a data augmentation strategy in this stage. Therefore, augmenting the data by creating additional images helps reduce the risk of an overfitting defect that may occur due to the in-depth nature of the model. Data augmentation offers a significant benefit by enhancing the model's ability to generalise data, particularly for X-ray datasets. Data augmentation employs several approaches to augment the training samples, hence improving the effectiveness of the model. The Keras API, namely the "Image Data Generator," is employed for this objective. The experiment utilised two methods, specifically "in place" and "on the fly", for data augmentation. These methods involved randomly transforming the photos. Every image is subjected to a 20-degree rotation throughout the

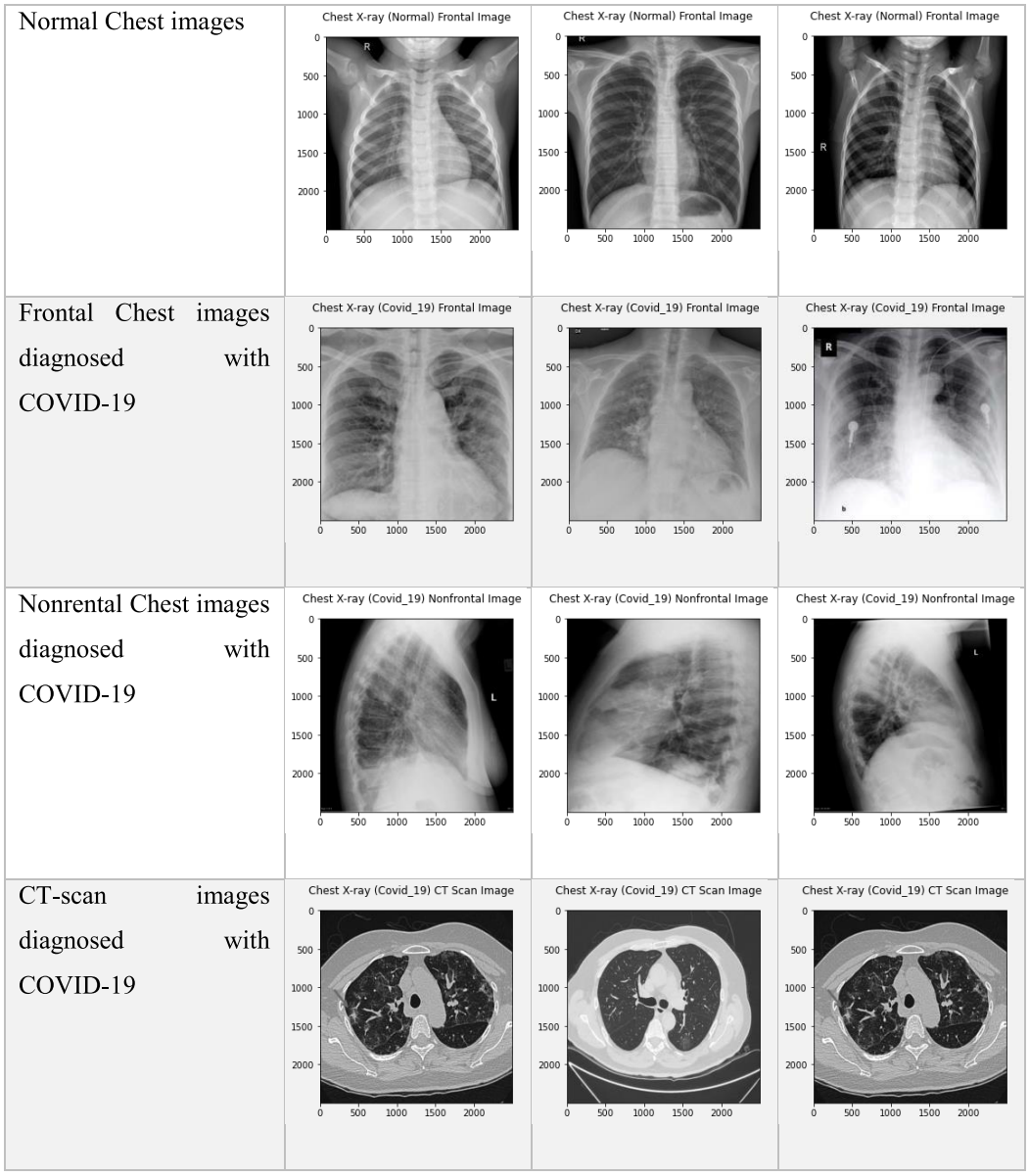

Fig. 2. Chest image samples.

augmentation process. In order to achieve a precise horizontal flip, the shear and zoom parameters have been precisely adjusted to a value of 20 percent. The implemented augmentation techniques aimed to enhance the generalizability of the intended model. The data augmentations were only done once to the X-ray training dataset. The training data for augmented X-ray images are inputted into a deep-learning model to achieve the ultimate forecast.

By exploiting pre-trained Convolutional Neural Network (CNN) models, this research aims to design a robust system for diagnosing COVID-19. The process of correctly classifying data is an integral part of machine learning. In this setting, deep transfer learning (DTL), which uses pretrained CNN models, is the technique used. In cases where training data is minimal, DTL has presented impressive levels of effectiveness. DTL is data-driven and is based on the transfer of knowledge from a source domain that uses a wide available dataset to learn the model to a target domain that has a relatively small dataset. This knowledge transfer enables accurate image classification in situations where the availability of training samples is limited. DTL applied to CNN involves the transfer of specific layers from a pre-trained model to the target domain. The initial model underwent millions of images of training. By utilising the pre-trained model's extensive learning on a wide range

Table 3. Decision matrix.

Criteria \ Network	AlexNet	VGG-16	VGG-19	SqueezeNet	GoogleNet	MobileNet-V2	ResNet-18	ResNet-50	ResNet-101	Xception
TP	472	411	481	407	427	496	480	486	509	503
FN	38	99	29	103	83	14	30	24	1	7
FP	147	68	106	73	71	54	53	1	4	2
TN	363	442	404	437	439	456	457	509	506	508
BACC	417.5	426.5	442.5	422	433	476	468.5	497.5	507.5	505.5
ACC	0.818627451	0.836275	0.86764706	0.827451	0.84902	0.933333	0.918627	0.97549	0.995098	0.991176
Recall	0.925490196	0.805882	0.94313725	0.798039	0.837255	0.972549	0.941176	0.952941	0.998039	0.986275
PPV	0.762520194	0.858038	0.81942078	0.847917	0.85743	0.901818	0.900563	0.997947	0.992203	0.99604
F Score	0.718417047	0.711073	0.78084416	0.698113	0.73494	0.879433	0.852575	0.951076	0.990272	0.982422
Specificity	0.711764706	0.866667	0.79215686	0.856863	0.860784	0.894118	0.896078	0.998039	0.992157	0.996078
FPR	0.288235294	0.133333	0.20784314	0.143137	0.139216	0.105882	0.103922	0.001961	0.007843	0.003922
FNR	0.074509804	0.194118	0.05686275	0.201961	0.162745	0.027451	0.058824	0.047059	0.001961	0.013725
FDR- P value	0.237479806	0.141962	0.18057922	0.152083	0.14257	0.098182	0.099437	0.002053	0.007797	0.00396
FNR	0.074509804	0.194118	0.05686275	0.201961	0.162745	0.027451	0.058824	0.047059	0.001961	0.013725
NPV	0.905236908	0.817006	0.9330254	0.809259	0.840996	0.970213	0.938398	0.954972	0.998028	0.986408
GM	0.811622608	0.835722	0.86435678	0.826928	0.848938	0.932509	0.918351	0.97523	0.995094	0.991164
Kappa	0.818627451	0.836275	0.86764706	0.827451	0.84902	0.933333	0.918627	0.97549	0.995098	0.991176
FDR	0.237479806	0.141962	0.18057922	0.152083	0.14257	0.098182	0.099437	0.002053	0.007797	0.00396
MCC	0.652327745	0.673795	0.74382071	0.656038	0.698233	0.869345	0.838108	0.951949	0.990213	0.9824
FOR	0.094763092	0.182994	0.0669746	0.190741	0.159004	0.029787	0.061602	0.045028	0.001972	0.013592
PPCR	0.606862745	0.469608	0.5754902	0.470588	0.488235	0.539216	0.522549	0.477451	0.502941	0.495098
LR+	3.210884354	6.044118	4.53773585	5.575342	6.014085	9.185185	9.056604	486	127.25	251.5
LR−	0.104683196	0.223982	0.07178218	0.235698	0.189066	0.030702	0.065646	0.047151	0.001976	0.01378
DOR	30.67239527	26.98485	63.2153546	23.65461	31.80943	299.1746	137.9623	10307.25	64388.5	18251.71

of datasets, this methodology improves the model's performance on tasks that require a smaller amount of training data. According to the studies [63, 64], the CNN model's task-dependent layers, layers that aren't currently being used for classification, are kept separate from the rest of the network design, such as the output classification layer. Accordingly, ten well-known transfer-learning models that were used in prior COVID-19 diagnostic studies and had good results were chosen to fulfil the goal of this study. Models of DTL and deep diagnosis are trained using the training set. The test dataset is fed to the 10 trained deep diagnosis models to evaluate the ability of deep diagnosis models to detect COVID-19 cases and normal. X-ray pictures are shown in Table 3, to demonstrate the effectiveness of 10 deep COVID-19 diagnostic models.

Based on Table 3 above, we cannot predict the best algorithm due to the variability of results among the ten algorithms used. Therefore, MCDM methods (i.e., TrIF-FWZIC and TrIF-FDOSM) present as an MCDM solution in the current study and are elaborated on in more detail in the next section.

3.2. Phase two: MCDM methods development

In this phase, the development of dual MCDM methods (FWZIC and FDOSM) using TrIFNs is presented.

3.2.1. TrIF-FWZIC

The determined distribution criteria will be used in the following TiIF-FWZIC procedures to develop each weight. The five phases of FWZIC and all the details involved are laid out in full below.

Step 1: The first stage is to specify the criteria that will be used to rank the alternatives.

There are two procedures at this stage. The first step involves identifying and presenting a set of assessment criteria that has already been established, and the second involves organising and categorising the criteria that have been gathered. In addition, the previously mentioned panel of experts evaluated the created and selected criteria, as outlined in Section 3.1, which will be further detailed in the subsequent steps [65].

Step 2: Structured expert judgment (SEJ)

Three experts were engaged to evaluate the importance of the criteria established in the previous phase and to identify their significance. The SEJ panel was established subsequent to the compilation of a list of potential experts through thorough investigation and identification. Subsequently, a comprehensive evaluation form was created to solicit feedback from all the SEJ panellists, and the linguistic rating system was transformed into a numerical scale [65].

Identify experts: Anyone who has experience with a subject could be considered an expert in a

Table 4. Linguistic likert scale.

Numerical scale	Linguistic scale
1	Not_important
2	Slight_important
3	Moderately_important
4	Important
5	Very_important

specific subject. The present research used a bibliometric analysis of all authors and co-authors of papers that included deep learning model criteria as the basis for the expert selection technique.

***Select experts*:** Experts were chosen to participate in the research once the group of experts was identified. In most cases, it's preferable to employ the maximum number of qualified individuals allowed. Three professionals were selected as specialists in this research. Emails were sent to everyone suggested as an expert during the expert identification process to gauge their level of interest and whether or not they saw themselves as qualified to serve on the panel. After a pool of potential experts was narrowed down to three, they worked together as a panel of judges.

Develop the evaluation form: The creation of an assessment form is an important stage since it serves as a tool for gathering agreement amongst experts. All three of the experts chosen in the previous round assessed the questionnaire before it was put through reliability and validity testing and finalised as part of the assessment form.

Define the level of importance scale: At this point, the three experts will assess the importance of each criterion using a 5-point Likert scale. Employing a response scale that incorporates varying lengths does not provide any theoretical issues [66].

Convert linguistic scale to numerical scale: The preference values are recognised subjectively; however, they cannot be used for further research unless the language scale is converted into a uniform numerical scale. Consequently, the experts' reported importance/significance on the linguistic Likert scale was transformed to a uniform numerical scale at this point. The conversion is illustrated in Table 4.

Experts are expected to use a Likert scale to give different weights to the various deep-learning criteria. With the use of a linguistic scale, the assessment criteria can be easily applied to the given significance. There are a wide variety of significance levels, from "not important" to "very important." To further analyse the scores given by experts, however, linguistic scores must be transformed into numerical values to be of any use.

Step 3. Construct the expert decision matrix (EDM)

Table 5. EDM.

Experts \ Criteria	C1	C2	...	Cn
EX1	Imp (EX1/C1)	Imp (EX1/C2)	...	Imp (EX1/Cn)
EX2	Imp (EX2/C1)	Imp (EX2/C2)	...	Imp (EX2/Cn)
...	...	...	...	...
EXm	Imp (EXn/C1)	Imp (EXn/C2)	...	Imp (EXm/Cn)

Imp **represents the importance level.

Table 6. Linguistic terms and their equivalent TrIFNs [32].

linguistic scale	TrIFNs
Not important	(0,0.1,0.2,0.3;0,0.1,0.2,0.3)
Slight important	(0.1,0.2,0.3,0.4;0.05,0.2,0.3,0.5)
Moderately important	(0.3,0.4,0.5,0.6;0.2,0.4,0.5,0.7)
Important	(0.5,0.6,0.7,0.8;0.4,0.6,0.7,0.9)
Very important	(0.7,0.8,0.9,1;0.7,0.8,0.9,1)

How the experts were chosen and how their preferences were communicated was laid out in detail in the preceding stage. The EDM is built in this stage. Table 5 outlines the crucial elements of EDM.

Step 4: Application of Trapezoidal Intuitionistic fuzzy

Here, we apply the TrIF to the EDM data, and then we defuzzify the data so that it can be used more precisely and more easily in later analysis, yielding a Trapezoidal Intuitionistic EDM. In MCDM, however, the issue is nebulous and imprecise since providing a specific preference weight to each criterion is hard. Rather than employing exact numbers, as is done in classical approaches, the fuzzy technique uses approximate values for qualities (criteria) to account for the imprecision and uncertainty inherent in the problem-solving scenario (all definitions of TrIFNs will be explained in the next section) [32], where the value of each linguistic term with TrIFNs is shown in Table 6.

Table 6 proves the conversion of all linguistic variables into TrIFNs, with each requirement for Experts represented by a fuzzy integer. A specialist in deep learning and diseases was requested to assess the significance of the criteria.

Step 5: Computation of the final values of the weight coefficients of the evaluation criteria

In this stage, the weight numbers of the assessment criteria (w1, w2, ..., wn) are produced based on the fuzziness data got in the prior step.

The ratio of fuzzification information is intended using the formula (Eq. (1)). Table 7 shows that.

$$\frac{\widetilde{Imp\,(E1/C1)}}{\sum_{j=1}^{n} \widetilde{Imp(E1/C_{1j})}} \tag{1}$$

Table 7. TrIF-EDM.

Experts \ Criteria	$\widetilde{C1}$	$\widetilde{C2}$	...	$\widetilde{Cn}$
E1	$\widetilde{E_1 : C_1}$	$\widetilde{E_1 : C_2}$	...	$\widetilde{E_1 : C_n}$
E2	$\widetilde{E_2 : C_1}$	$\widetilde{E_2 : C_2}$	...	$\widetilde{E_2 : C_n}$
...	...	...	...	...
Em	$\widetilde{E_m : C_1}$	$\widetilde{E_m : C_2}$	...	$\widetilde{E_m : C_n}$

1. Using formula 2 to get the value of every criterion.$(\widetilde{w1}, \widetilde{w2}, \ , \widetilde{wn})^T$.

$$= \left(\begin{bmatrix} 1 - \prod_{i=1}^{n} (1 - a_i^L), & 1 - \prod_{i=1}^{n} (1 - a_i^{M1}), \\ 1 - \prod_{i=1}^{n} (1 - a_i^{M2}), & 1 - \prod_{i=1}^{n} (1 - a_i^U) \end{bmatrix}, \begin{bmatrix} \prod_{i=1}^{n} (b_i^L), & \prod_{i=1}^{n} (b_i^{M1}), \\ \prod_{i=1}^{n} (b_i^{M2}), & \prod_{i=1}^{n} (b_i^U) \end{bmatrix} \right) \tag{2}$$

2. To obtain the ultimate weight, employing formula 3. It is important to note that the total weight should be equal to 1.

$$\widetilde{W_J} = \left(\sum_{i=1}^{m} \frac{\mathrm{Imp}(\overline{E_{tj}}/C_{tj})}{\sum_{j=1}^{n} \mathrm{Imp}(\mathrm{E}(\mathrm{E_{tj}}/C_{tj})} \right) /m \tag{3}$$

3. Defuzzification is carried out to determine the ultimate weight, using (Eq. (4)) as the defuzzification technique. In order to find the ending values of the weight, it is necessary to assign a weight to each criterion based on the total sum of weights assigned to all criteria for the purpose of rescaling in this stage.

$$\widetilde{A} = \frac{(a_3' + a_4' - a_1' - a_2') + (a_3 + a_4 - a_1 - a_2)}{4} \tag{4}$$

3.3. TrIF-FDOSM

This part explains comprehensive outline of the procedural processes involved in the TrIF-FDOSM approach for evaluating and ranking DM alternatives, specifically focusing on deep learning algorithms. These steps are visually depicted in Fig. 3. The data transformation unit and data processing step of TrIF-FDOSM were chosen by the expert, and their description is provided below.

3.3.1. Stage 1: Transformation of Data: this stage included 2 steps

Step 1: In this step, Determine the optimal solution for each criterion employed in benchmarking the deep learning approaches for COVID-19. Therefore, the ideal solution is defined as follows [67]:

$$A^* = \left\{ \left[\left(\max_i v_{ij} | j \in J \right), \left(\min_i v_{ij} \ | j \in J \right), \right. \right. \\ \left. \left. (Op_{ij} \in I.J) \, | i = 1.2.3. \ldots . m \right] \right\} \tag{5}$$

Where max is the perfect solution for benefit criteria, min is the ideal solution for nonbenefit criteria, and Op i is the critical value when the optimal value is among maximum and minimum. As a result, the person making the choice is the one who determines what the important value is. Using the critical value in this study, however, does not need any criterion to establish the optimal option.

Step 2: Using the benchmarking criteria for deep learning COVID-19 approaches, in step 2, the optimum answer is compared to alternative values. It's the ideal solution selection phase that compares an ideal solution to the value of alternatives in the same criteria [15, 17].

$$Op_{Lang} = \left\{ \left(\left(\tilde{v} \otimes v_{ij} | j \in J \right) . | i = 1.2.3. \ldots . m \right) \right\}, \tag{6}$$

The ⊗ mean comparison between optimal solution and alternative values in similar criteria is represented by the reference comparison. Trapezoidal intuitionistic fuzzy may be used to generate fuzzy numbers from a linguistic term opinion matrix, which is the end product of this block's computations.

$$Op\_Lang = \begin{matrix} A_1 \\ \vdots \\ A_m \end{matrix} \begin{bmatrix} op_{11} & \ldots & op_{1n} \\ \vdots & \ddots & \vdots \\ op_{m1} & \ldots & op_{mn} \end{bmatrix}. \tag{7}$$

In this research, opinion matrices are created by having three experts do FDOSM comparisons. Different geographic regions are represented in the panel of experts (decision-makers). Moreover, the experts must have ten years of experience in artificial intelligence and deep learning methods.

3.3.2. Stage 2: Data-processing unit

Trapezoidal intuitionistic fuzzy transforms the opinion matrix into a fuzzy decision matrix in the last block. This was followed by the use of triangle membership and arithmetic means, which aggregated data via direct aggregation in two key contexts (individual and group decision-making). For the assessment

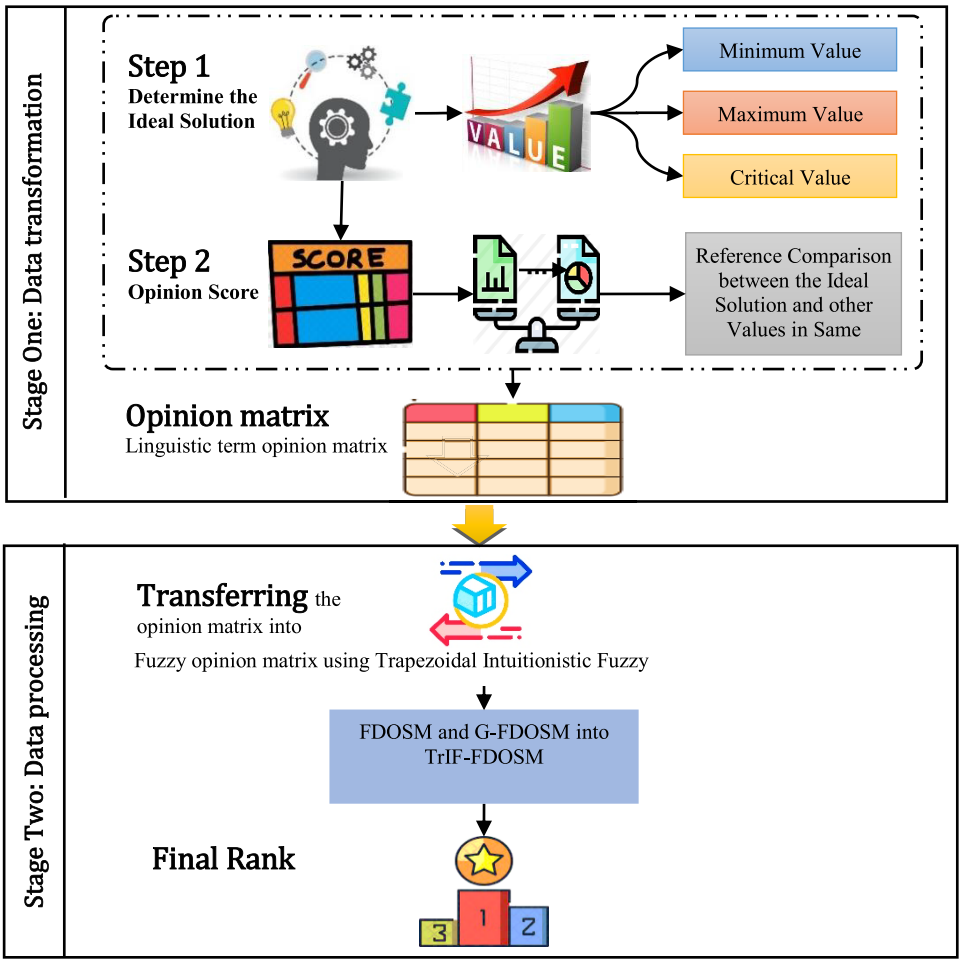

Fig. 3. TrIF-FDOSM stages.

and benchmarking of the COVID-19 algorithms, TrIF-FDOSM is extended from FDOSM.

3.3.2.1. Benchmarking deep learning COVID-19 models based on TrIF-FDOSM. In this paper, we applied TrIF-FDOSM. The steps are defined as follows:

Step 1: First, we present TrIF (see Fig. 4) and briefly review some definitions of it. Intuitionistic fuzzy sets have been developed by Atanassov and widely used. The concept of the intuitionistic fuzzy sets is to consider the membership value and the non-membership value to describe x in X. The summation of membership and non-membership is equal to 1.

Def 1: Let $X \ \emptyset$ be a given set. An intuitionistic fuzzy set in X is an object A given by [32]:

$$\widetilde{A} = \left\{\left\langle x,\ \mu_{\tilde{A}}(x), \nu_{\tilde{A}}(x)\right\rangle; x \in X\right\}, \tag{8}$$

Where $\mu_{\tilde{A}} : X \rightarrow [0,1]$ and $\nu_{\tilde{A}} : X \rightarrow [0,1]$ satisfy the condition $0 \leq \mu_{\tilde{A}}(x) + \nu_{\tilde{A}}(x) \leq 1$, for every $x \in X$. Hesitancy is equal to $1 - \mu_{\tilde{A}}(x) + \nu_{\tilde{A}}(x)$.

Def 2: (TrIFNs) An intuitionistic fuzzy subset in R called a TrIFN (A) has the following membership and non-membership functions [32]:

$$\mu_{\tilde{A}}(x) = \begin{cases} \frac{x - a_1}{a_2 - a_1}, & for\ a_1 \leq x \leq a_2 \\ 1, & for\ a_2 \leq x \leq a_3 \\ \frac{a_4 - x}{a_4 - a_3}, & for\ a_3 \leq x \leq a_4 \\ 0, & otherwise \end{cases}, \tag{9}$$

And

$$\nu_{\tilde{A}}(x) = \begin{cases} \frac{a_2 - x}{a_2 - a_1'}, & for\ a_1' \leq x \leq a_2 \\ 0, & for\ a_2 \leq x \leq a_3 \\ \frac{x - a_3}{a_4' - a_3}, & for\ a_3 \leq x \leq a_4' \\ 1, & otherwise \end{cases}, \tag{10}$$

Where $a_1' \leq a_1 \leq a_2 \leq a_3 \leq a_4 \leq a_4', 0 \leq \mu_{\tilde{A}}(x) + \nu_{\tilde{A}}(x) \leq 1$ and TrIFN is denoted by $\tilde{A}_{\text{TrIFN}} = (a_1, a_2, a_3, a_4; a_1', a_2, a_3, a_4')$ (see Fig. 5).

Def 3: Let $\tilde{A}$ and $\tilde{B}$ be two Atanassov's IFSs in set X. The intersection of $\tilde{A}$ and $\tilde{B}$ is defined as the following

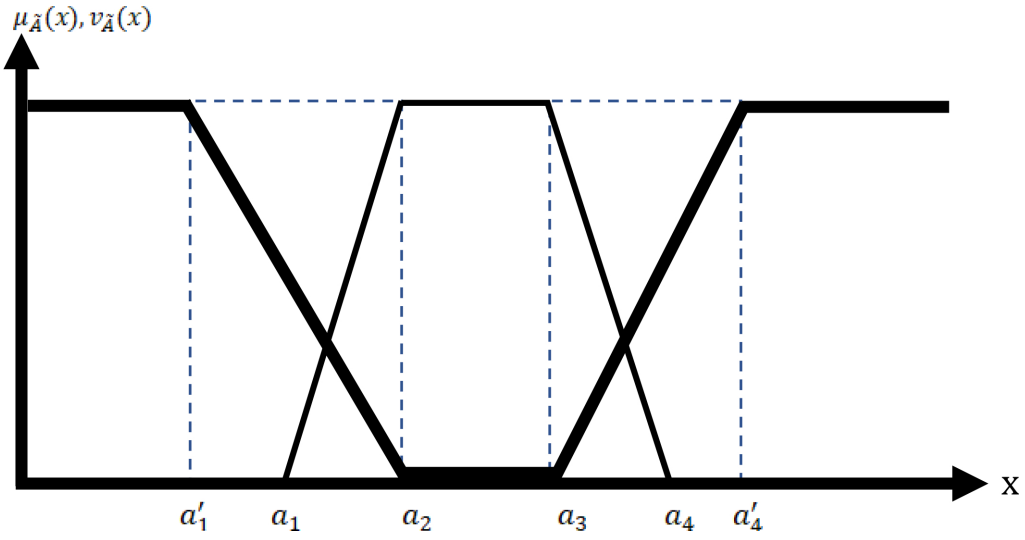

Fig. 4. A trapezoidal intuitionistic fuzzy number.

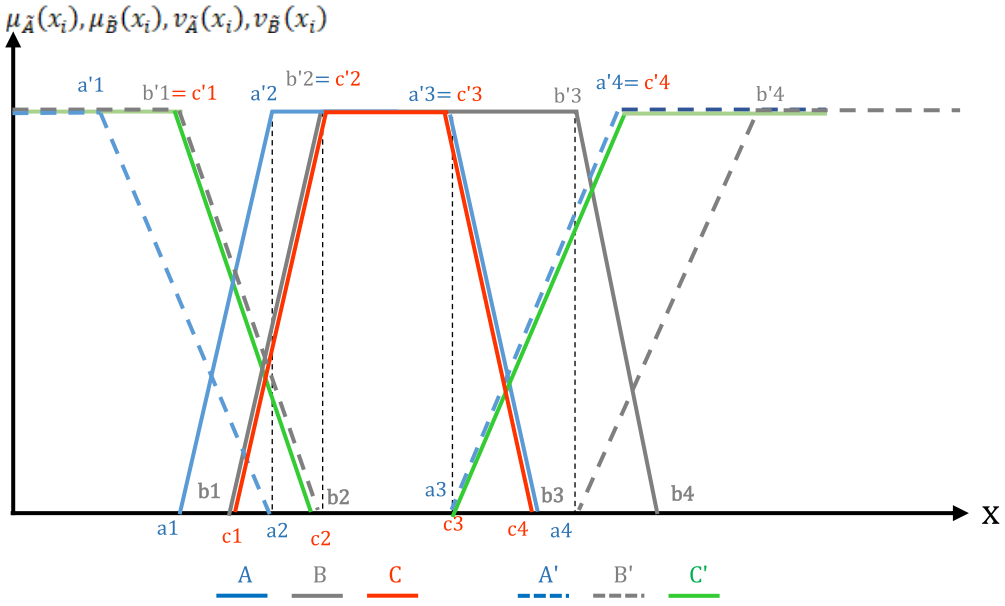

Fig. 5. Intersection of two TrIFNs.

equation [68]. (See Fig. 5).

$$\tilde{A} \cap \tilde{B} = \left\{ x_i,\ \min(\mu_{\tilde{A}}(x_i), \mu_{\tilde{B}}(x_i)),\ \max(v_{\tilde{A}}(x_i),\ v_{\tilde{B}}(x_i)) | x_i \in X \right\} \quad (11)$$

Let $\tilde{A}_{\text{TrIFN}} = (a_1,\ a_2,\ a_3, a_4;\ a_1',\ a_2',\ a_3',\ a_4')$ and $\tilde{B}_{\text{TrIFN}} = (b_1,\ b_2,\ b_3, b_4;\ b_1',\ b_2',\ b_3',\ b_4')$ Assume that the magnitudes of these parameters are as in Fig. 6. In this figure $\mu_{\tilde{A}}(x_i)$ and $\mu_{\tilde{B}}(x_i)$ represent the membership functions of the fuzzy sets $\tilde{A}$ and $\tilde{B}$ respectively. And, $v_{\tilde{A}}(x_i)$ and $v_{\tilde{B}}(x_i)$ represent the non-membership functions of the fuzzy sets $\tilde{A}$ and $\tilde{B}$ respectively. The intersection $\tilde{A}_{\text{TrIFN}} \cap \tilde{B}_{\text{TrIFN}}$ is TrIFN denoted by $\tilde{C}_{\text{TrIFN}} = (c_1,\ c_2,\ c_3, c_4;\ c_1',\ c_2',\ c_3',\ c_4')$ Fig. 6 illustrates this intersection process. The intersection of (A) and B membership functions, shown by the red line, is indicated as $\tilde{C}_{\text{TrIFN}}$.

Definition 4: TrIFNs have aggregation operators. In decision-making situations, the aggregation of preferences is an essential issue [68]. Intuitionistic fuzzy information may be aggregated to account for experts' scepticism. Trapezoidal intuitionistic fuzzy numbers might be aggregated as follows:

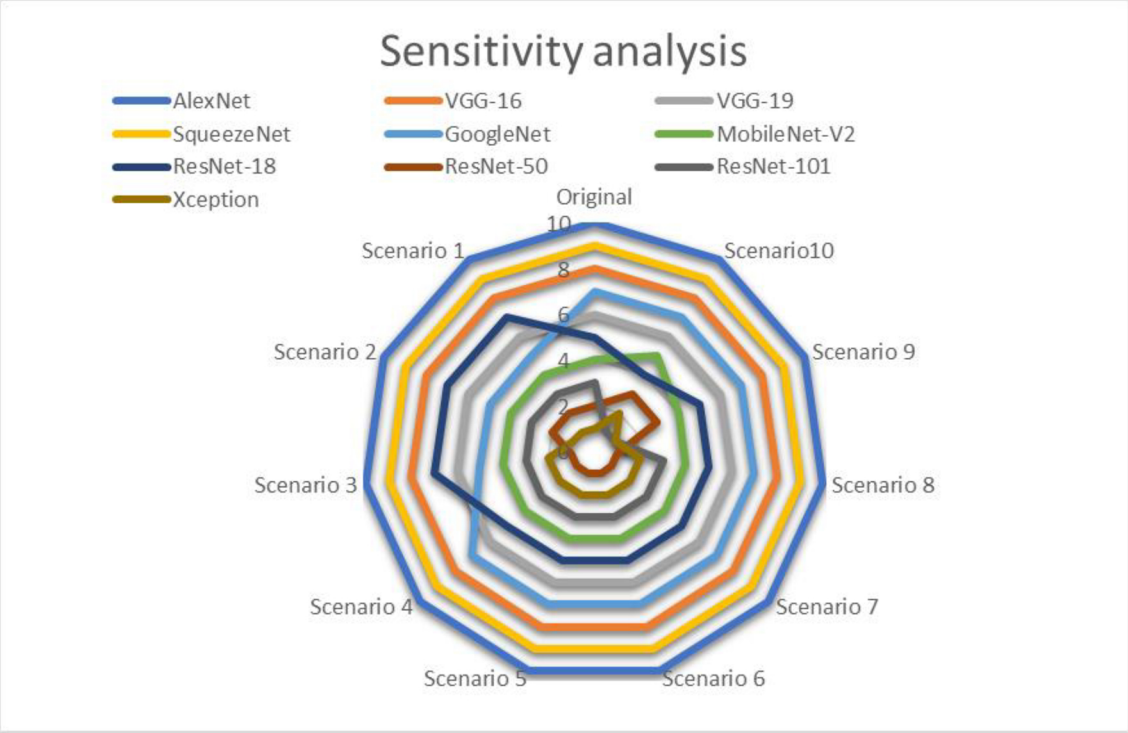

Fig. 6. Sensitivity analysis.

Table 8. Convert the linguistic terms into Trapezoidal Intuitionistic Fuzzy Numbers (TrIFNs).

Linguistic Terms	TrIFNs
No Difference (NO.D)	(0,0.1,0.2,0.3;0,0.1,0.2,0.3)
Slight Difference (S.D)	(0.1,0.2,0.3,0.4;0.05,0.2,0.3,0.5)
Difference (D.I)	(0.3,0.4,0.5,0.6;0.2,0.4,0.5,0.7)
Big Difference (B.D)	(0.5,0.6,0.7,0.8;0.4,0.6,0.7,0.9)
Huge Difference (H.D)	(0.7,0.8,0.9,1;0.7,0.8,0.9,1)

Let $I_i = ([a_i^L,\ a_i^{M1},\ a_i^{M2},\ a_i^U],\ [b_i^L, b,\ b_i^{M2},\ b_i^U])$, and $(i = 1,\ 2,\ ,\ n)$ is a set of TrIFNs, then the result is a TrIFNs aggregate by the following equation, and $f_m = (I_1,\ I_2,\ ,\ I_n)$

$$= \left(\begin{bmatrix} 1 - \prod_{i=1}^{n} (1 - a_i^L)\, w,\ 1 - \prod_{i=1}^{n} (1 - a_i^{M1})\, w, \\ 1 - \prod_{i=1}^{n} (1 - a_i^{M2})\, w,\ 1 - \prod_{i=1}^{n} (1 - a_i^U)\, w \end{bmatrix}, \begin{bmatrix} \prod_{i=1}^{n} (b_i^L)\, w,\ \prod_{i=1}^{n} (b_i^{M1})\, w, \\ \prod_{i=1}^{n} (b_i^{M2})\, w,\ \prod_{i=1}^{n} (b_i^U)\, w \end{bmatrix} \right) \tag{12}$$

According to this, the linguistic terms of the opinion matrix will be denoted in trapezoidal intuitionistic fuzzy numbers (see Table 8) to create the fuzzy opinion decision matrix.

Step2: Use Definition 4 to sum up the results from the previous step for each possibility (Eq. (12)). Upon completion of the fuzzy decision matrix, the aggregation method is used to select the optimal choice.

3.3.3. Benchmarking deep learning COVID-19 methods based on group decision-making context

An aggregated conclusion from many evaluators is required to unify the benchmarking outcome because of differences in decision-makers usage of the benchmarking deep learning COVID-19 models. A group decision-making process was required in order to combine all benchmarking of the decision-makers into COVID-19 models. Consequently, the group decision-making context will be adopted and used with TrIF-FDOSM. Also, the total mark is calculated by using the arithmetic mean. Also, remember that the highest score shows the best option. Group decision-making context is applied. Only after the last ranking is determined will experts share their ideas. [14, 17].

$$\begin{aligned} &Group\ TrIF - FDOSM = \ \oplus\ F^* \\ &\oplus = \text{Arithmetic mean.} \\ &F^* = \text{The Final result for each expert.} \end{aligned} \tag{13}$$

Table 9. Final the weight for each criterion.

criteria	weight	criteria	weight	criteria	weight
TP	0.0497	F Score	0.0434	BCAC	0.0378
FN	0.0502	SPC, TNR	0.0497	FDR	0.0312
FP	0.0497	FPR	0.0439	MCC	0.0374
TN	0.0497	FNR	0.0374	FOR	0.0378
BACC	0.0374	FDR - P value	0.0378	PPCR	0.0378
ACC	0.0559	FNR	0.0374	LR +	0.0374
TPR	0.0374	NPV	0.0374	LR−	0.0378
PPV	0.0502	GM	0.0378	Diagnostic odds ratio	0.0378

4. Results and discussion

Findings related to applying deep learning algorithms to COVID-19 are shown here, with special emphasis on analyzing various outcomes in terms of weighting and ranking them. This part has been organized into two different sections. Here, you will find the outcomes of the TrIF-FWZIC approach as regards how criteria are weighted and put into practice. In addition, this section shows the rank of different routing algorithms using both individual decision-making and group decision-making.

4.1. Weighting result

The TrIF-FWZIC technique is introduced in Section 3.2; the section describes the weight impacts of each criterion. The TrIF-FWZIC method, with group weights, was produced using these five phases with no inconsistency issues. Table 9 displays the criterion-weighted results.

A variety of metrics used in evaluating deep learning models for COVID-19 diagnosis was analyzed using a TrIF-FWZIC. Table 9 summarizes the rating process by showing which criteria are most significant. When measuring a model's performance, ACC was given the most importance, with a weight of 0.0559. Metrics measuring sensitivity, including TP, FN, and PPV, were given strong weights (0.0502–0.0497), meaning specialists look to accurately find people with COVID-19. Measures used less often, for example, the FDR and FOR had the lowest weights (about 0.0312–0.0378), showing that they have a lower significance. Specifically, MCC, GM, and LR + and LR− were equally important, showing that the model's power to distinguish groups and its consistency were considered in equal measure. This variety in importance reflects that the framework considers different features of diagnostic performance, picking the best-performing deep learning models for patients. These weights were decided to be entered in TrIF-FDOSM to evaluate the deep learning model.

4.2. Ranking result

This section compares deep-learning COVID-19 models using both individual and group decision-making scenarios. For the purpose of this benchmarking procedure, the authors provide the opinion matrix and fuzzy opinion matrix. Three decision-makers provide their viewpoints using a five-point scale, thereby transforming the decision matrix into the opinion matrix. The Appendix (Table A1) comprises the reported opinion matrices of all decision-makers. Following Eq. (5), decision-makers select the ideal answer for each criterion, and after building the opinion matrix, a Likert scale (Eq. (6)) is used to compare the ideal and other values for each choice. The subsequent procedure entails utilising Trapezoidal Intuitionistic Fuzzy Numbers (terms 1, 2, and 3) to convert the opinion matrix into a fuzzy number matrix. Table 5 converts each decision-maker's opinion matrix into a fuzzy opinion matrix. Following that, Tables A1 and A2 in the Appendix exhibit the fuzzy opinion matrices of other experts. The resultant fuzzy opinion matrices are then used to benchmark deep-learning COVID-19 models using TrIF-FDOSM. The results of each individual's decision-making context are detailed in the following section. Deep learning COVID-19 models were benchmarked against the three decision-makers' individual decision-making settings, as shown below (see Table 10).

A decision maker's viewpoint is critical in each criterion, according to the benchmarking findings. The best alternative is the highest score. Furthermore, the last score is the worst alternative. The best alternative for the 1st decision-maker was "Xception models", with "0.267510407". Extensive FDOSM findings based on the 3 decision-maker's viewpoints are shown in Table 10. Three decision-makers gave their opinion on extending FDOSM and achieved the final result for benchmarking deep learning COVID-19 models.

Furthermore, the best alternative for the 2nd and 3rd experts was the "ResNet-101 model" with the score "0.316710828, 0.457770263", respectively. On the one hand, the 1st DM to determine the

Table 10. Results of each individual expert.

	DM_1		DM_2		DM_3	
Alternatives	Score	Rank	Score	Rank	Score	Rank
AlexNet	1.12347E-07	8	3.43034E-08	10	1.36747E-07	10
VGG-16	4.05949E-08	10	5.78875E-06	8	1.29007E-06	8
VGG-19	8.12408E-06	6	4.40337E-05	6	4.13857E-05	6
SqueezeNet	5.82193E-08	9	1.16321E-06	9	6.50735E-07	9
GoogleNet	2.8068E-06	7	1.38122E-05	7	3.21983E-05	7
MobileNet-V2	0.003321091	5	0.009147748	4	0.017787815	4
ResNet-18	0.003587109	4	0.002682304	5	0.011058793	5
ResNet-50	0.152637676	3	0.071378892	3	0.154158218	3
ResNet-101	0.202761199	2	0.316710828	1	0.457770263	1
Xception	0.288187647	1	0.267510407	2	0.335221384	2

worst alternative was “VGG-16” with a score “of 4.05949E-08”. For the 2nd and 3rd DM, the worst alternative was “AlexNet,” with the score “3.43034E-08, 1.36747E-07”, respectively. When the opinion matrix is turned into a fuzzy opinion matrix, this expansion allows additional flexibility in dealing with the uncertainty of the opinion matrix. When comparing the final findings to the opinion matrix of each decision-maker, we get more accurate results because of this flexibility. However, Table 10 cannot determine the optimal choice due to the influence of decision-maker’s opinions on the variability of ranking scores. Group decision-making can be employed to prioritise the alternatives based on the collective opinions of all specialists. Furthermore, group decision-making is essential for resolving the issue of discrepancies in the ultimate ranking. The subsequent part presents the outcome of TrIF-FDOSM in the context of group decision-making.

4.3. Results of group expert

In this sector, we explain the results of the group decision-making context (GDM). As mentioned in Section (3.3.3), and according to Eq. (13), the three decision-makers’ final results were aggregated using the arithmetic mean operator, and the final GDM ranking for benchmarking deep learning COVID-19 models was achieved. Table 11 shows the final GDM results, with the highest score as the best alternative.

The assessment and ranking of ten deep learning models according to several criteria led to interesting and useful outcomes for determining the best deep learning model. From Table 11, ResNet-101 was ranked first with a score of 0.32574743; it was the most preferred alternative of the evaluated models. Coming hot on its heels in second place was Xception with a score of 0.296973146, while third in line was ResNet-50 with a score of 0.126058262. From these results, we discover that the ResNet-s family of models is distinguished by the advanced architecture

Table 11. Group TrIF-FDOSM context.

Alternatives	Score	Rank
AlexNet	9.44658E-08	10
VGG-16	2.37314E-06	8
VGG-19	3.11812E-05	6
SqueezeNet	6.24056E-07	9
GoogleNet	1.62724E-05	7
MobileNet-V2	0.010085551	4
ResNet-18	0.005776069	5
ResNet-50	0.126058262	3
ResNet-101	0.32574743	1
Xception	0.296973146	2

of ResNet-101 and ResNet-50 (which helps to maintain the problem of vanishing gradients through skip connections). Similarly, Xception’s high position has the ability to be attributed to its depth wise separable convolutions, which increase the efficiency and accuracy of learning while being computationally feasible.

In the mid-level ranking, MobileNet-V2 (fourth place, score: 0.010085551) and ResNet-18 (5, score: 0.005776069) showed moderate performance. Having a lightweight architecture and relatively quick inference time, these models are quite appealing for the release in limited-resource environments, such as mobile and embedded systems. Following this up, we have VGG-19 (6th ranked, score: 3.11812E-05) and GoogleNet (in seventh place, score 1.62724E-05) and 0.0000312) provided rather modest results as well. Both of these models are heritage models that have played an important role in the evolution of CNN architecture. VGG-16 (ranked eighth, score: 2.37314E-06) and SqueezeNet (rank ninth, score 6.24056E-07), proved to be even worse (both models are limited in their model capacities or adaptability to the criteria evaluated in this study). AlexNet, also part of this category, achieving the lowest score of 9.44658E-08 and the tenth position, can also provide proof.

The advantage of the ranking of deep learning COVID-19 models is the fact that it can direct

practitioners, researchers, and system developers to choose the best deep learning model for their particular case. This is particularly important in areas like healthcare, finance, and autonomous systems, where speed-accuracy-interpretability trade-offs will largely shape the success of a solution. Moreover, the TrIF-FDOSM approach enhances the robustness of the results by the consideration of the factors of uncertainty and vagueness in the expert judgment as well as the performance data. This is especially helpful when we cannot measure criteria exactly or when the evaluation of the qualitative characteristics of the models is based on expert opinion. All in all, such a ranking, besides revealing the most appropriate models, defines a replicable and flexible framework that could be adjusted to future model assessments or to other application domains. It's still possible to use the ranking of group decision-making context as the foundation for objective validation procedures. The following part goes into great depth about the outcomes of the validation.

5. Validation

Here, it discusses the proof procedure for the final result of TrIF-FDOSM to substantiate the deep learning COVID-19 models benchmarking group decision-making results obtained by the TrIF-FDOSM. Objective validation is applied by dividing the benchmarked deep-learning COVID-19 models into different groups. Several MCDM researchers have used this method to guide their findings. The outcome of the validation is unaffected by the number of groups or alternatives within each group. Several procedures are performed to validate the group benchmarking deep learning COVID-19 model results. (1) The deep learning COVID-19 models are sorted/ordered according to group decision-making results. (2) After sorting, the deep learning COVID-19 models have been separated into 2 equal groups. (3) The mean ($\bar{x}$) for each group in group decision-making result is calculated [69, 70]. Table 12 presents the precise validation for the TrIF-FDOSM method in group benchmarking. This validation result was attained by the minimum mean value for group benchmarking because of the philosophy of the TrIF-FDOSM in getting the ideal solution by the minimum linguistic terms. This method is developed by scientists and experts in the MCDM domain. Therefore, the lowest mean is supposed to get the minimum mean by the main group, and their value is compared with the second group to check their validity. The result is assumed to be valid if the assumption is consistent with the evaluation result and if the result of the second group is equal to or higher than that of the first group.

Table 12. Validation of Group Benchmarking Results of Deep Learning COVID-19 Models.

Group	Deep Learning models	Mean
1st Group	ResNet-101 Xception ResNet-50 MobileNet-V2 ResNet-18 VGG-19 GoogleNet	1.8194444
2nd Group	VGG-16 SqueezeNet AlexNet	3.6166666

Table 12 demonstrates the validation results of benchmarking deep-learning COVID-19 models using the proposed TrIF-FDOSM. The mean value of the first group (1.8194444) is lower than the mean value of the second group (3.6166666). Thus, the group expanded the reliable TrIF-FDOSM findings of the comparative evaluation of deep learning COVID-19 techniques and carried out a logical ranking.

In this second round of testing, the sensitivity analysis foresees how adjusting the relative importance of different criteria would affect the deep learning findings' systematic ranking [26]. Before diving into a sensitivity analysis, it's crucial to determine which criteria are most important. Table 9 shows that, out of the total of 24 criteria, accuracy was the most important one (0.0559). Using Eq. (14), we developed 10 scenarios to analyse the impact of varying criterion weights. Fig. 6 depicts 10 potential results based on the 0.3 rise and the 24 criteria utilised. The relative importance of each criterion was calculated using the following formula.

$$w_{n\beta} = (1 - w_{n\alpha}) \frac{w_\beta}{(1 - w_n)} \tag{14}$$

Fig. 6 shows how the ranking of deep learning models was affected when the criteria weights were varied in a sensitivity analysis. Using an MCDM technique, this method helps explain the impact of modifying the role of various evaluation metrics on model rankings for COVID-19 diagnosis. Ten well-known deep learning models were tested using the original weight configuration and against ten other alternatives that changed the weighting of the criteria. The results reveal that rank stability is high in almost all of the models. The rankings of AlexNet, VGG-16, VGG-19, and SqueezeNet do not change in any case, confirming that their relative performance stays the same. These models are not significantly affected by the way the weights are distributed and

seem to have stable results. Several models are a little bit more sensitive. GoogleNet shifts from 7th to 5th position in scenarios 1-3, but returns to 6th place in the following ones. The fluctuation in this scenario made this network switch from the 4th to the 5th-best network. Similarly, ResNet-18 changes position between fifth and seventh based on variation in scenarios, demonstrating that small changes in criteria do not have a big impact on its ranking. Although Xception, ResNet-50, and ResNet-101 are on top, they perform almost similarly. Xception is first most of the time but moves to 2nd place in Scenario 3 and 4, back and forth with ResNet-50 or ResNet-101. The fact that the slight variance in performance among the leading models at the top does not significantly alter the underlying truth that these models perform much better than others. Following the sensitivity analysis, it is confirmed that relying on this way to evaluate the model is reliable and can handle changes. With the changes in criteria importance, the ranking of the models didn't shift much, ensuring the effectiveness of the approach. This makes it more likely that healthcare professionals use the approach to pick the best deep learning models when addressing COVID-19 or other urgent medical situations.

In our last part of our investigation, we deployed the Spearman correlation coefficient (SCC) in order to determine correlations among the ten different scenarios. Spearman's rank is a statistic that is used quantitatively to determine the degree and the direction of a monotonic correlation between two variables. In this case, the given SCC approach was used to measure the level of similarity or dissimilarity of the given scenarios. The results of the correlation analysis have been depicted graphically in Fig. 7, which reveals much about the overall relationships among the 10 scenarios.

All Spearman Correlation Coefficient values found between the original evaluation and the alternatives show a high level of rank-order stability throughout. The relationship between the original ranking and those created under other settings was very strong in all cases, ranging from 0.93 (Scenario 3) up to 0.98 (Scenarios 4 to 8). This shows that the model rankings do not change much when the decision-making environment changes (weights of criteria), with just small differences observed. With a correlation of 0.93 indicating few changes in model order, even lower values in the results can still rely on the chosen method, and the similar 0.98 repeatedly shows that the evaluation method is reliable. In general, the data suggest that the chosen approaches result in dependable assessments.

6. Implications

Using Fuzzy MCDM strategies to assess Deep Learning models for detecting COVID-19 with CXR radiographs provides important practical value with the potential to shape healthcare and model development in many positive ways.

1. MCDM techniques facilitate a more refined decision-making approach by adding a clear and organised scheme to model evaluation and prioritisation.
2. By incorporating customisation and prioritisation functions, healthcare providers are able to customise the model selection process to their needs by initiating values for different parameters. Accuracy or sensitivity may be preferred, as some situations dictate, the characteristics of the patient population and clinical situation are considered. For this reason, they can choose a model that best suits their individual needs.
3. The relevance and flexibility of the Model-based group MCDM method are exemplified by its simplicity in adjusting to the new data or changes in the criteria. The ability to update and change models ensures that their rankings are current and reliable, and thus bolster ongoing advances in COVID-19 diagnostics.
4. The clarity of the established 24 criteria and rankings guarantees choosing the selected model as the most appropriate to the task at hand, thus supporting trust in services in healthcare.
5. The findings from MCDM provide direction to the experts conducting work in data science and model development. Through the evaluation of how their models behave with respect to a number of important factors, researchers will also be able to identify shortcomings and possibly focus their activities for the benefit of further developing more robust diagnostic solutions.
6. Reviewing many parameters allows healthcare providers to achieve maximum results by choosing the most effective approach's model for patient care. When COVID cases are correctly and reliably distinguished, timely action and better disease control can be suggested, and can reduce the mortality rate.
7. Resource optimisation at healthcare institutions can be achieved through the identification and implementation of the most efficient and effective methods for resource allocation. This practice guarantees the efficient utilisation of computational resources, server space, and budgetary allocations, resulting in cost

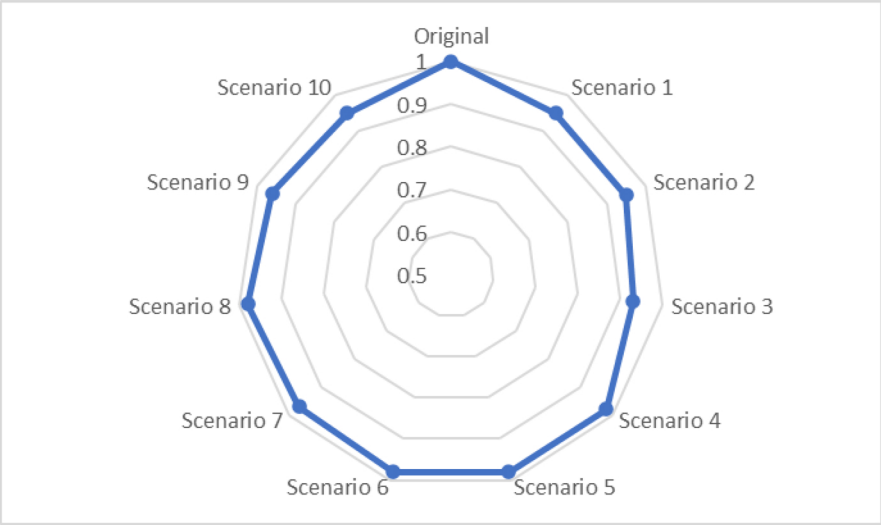

Fig. 7. Spearman correlation coefficient.

reductions and enhanced healthcare service as a whole.

7. Conclusion

COVID-19 may be diagnosed using deep learning models, which are one of the most advanced AI methods. The evaluation and benchmarking of the deep learning models concede multi-criteria decision problems according to different issues (i.e., multi-evaluation criteria, criteria importance, and data variation). This study presents a new decision matrix to evaluate and benchmark the deep learning models. This decision matrix contains 24 evaluation criteria. The goal of the decision matrix is to be the framework for any researcher who wants to evaluate and benchmark any deep learning models. On the other hand, in this study, FDOSM and FWZIC were extended into TrIF to address the uncertainty and ambiguity problems that methods suffered from. Therefore, the methodology of this work was designed with two main phases. The 1st phase is related to constructing the decision matrix (i.e., explaining how to extract the evaluation criteria and apply the deep learning models to the data set to achieve the value of each criterion). The 2nd phase is related to the development of TrIF-FWZIC and TrIF-FDOSM. The result of TrIF-FWZIC shows that the accuracy seemed to have the greater weight value (0.0559), whereas the FDR seemed to have the smallest weight value (0.0312). In addition, the individual decision-maker, the final result for the first decision-maker presents Xception as the best deep learning model with a score (i.e., 0.2881876). The best deep learning model for the second and third decision-makers was ResNet-101 with scores (i.e., 0.3167108, 0.4577702), respectively. Many variations are observed in the individual benchmarking results of deep learning models depending on each decision-maker. Therefore, group decision-making is applied. The best deep learning model, depending on the group decision-making, was “ResNet-101,” with a score (i.e., 0.325747). The worst deep learning model in group decision-making was “AlexNet” with a score (i.e., 9.44658E-08). The validation of the group decision-making result shows the first group is the lowest than the second group with scores (i.e., 1.8194444, 3.6166666) respectively. Despite the fact that this research makes significant discoveries, the study has many restrictions. While 10 deep learning models represent a substantial sample for analysis, their small number limits the applicability of the framework apart from these cases. The static nature of the data collection is unable to reflect dynamics and changes in performance for these models over time. In addition, a number of recent studies indicated that one could not identify the best deep learning model based on a given metric. Therefore, for future research direction, use the Fuzzy Delphi method to further standardise the assessment criteria for deep learning models. Rely on the FWZIC–FDOSM framework for the new domains (e.g., healthcare, MPSoC, IoT) and on a larger model pool to test generalizability. Proceed from type-1 fuzzy sets to interval and type-2 fuzzy MCDM in order to better reflect greater-order uncertainties in expert judgments. Finally, scalability of the FDOSM becomes highly problematic when the number of alternatives increases, and it essentially blocks computation.

In order to overcome these barriers, the innovation of improved versions of FDOSM with attention to computational effectiveness and robustness is crucial.

Funding

None.

Conflicts of interest

None.

Acknowledgement

None.

References

1. D. D. Miller and E. W. Brown, "Artificial intelligence in medical practice: the question to the answer?," *Am. J. Med.*, vol. 131, no. 2, pp. 129–133, 2018.
2. Y. L. Khaleel, M. A. Habeeb, and T. O. C. Edoh, "Limitations of deep learning vs. human intelligence: Training data, interpretability, bias, and ethics," *Appl. Data Sci. Anal.*, vol. 2025, pp. 3–6, 2025.
3. T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," *Chaos, Solitons & Fractals*, vol. 140, p. 110120, 2020.
4. I. Rahimi, F. Chen, and A. H. Gandomi, "A review on COVID-19 forecasting models," *Neural Comput. Appl.*, pp. 1–11, 2021.
5. Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imaging*, vol. 39, no. 8, pp. 2688–2700, 2020.
6. M. T. Ali, Y. R. Muhsen, R. F. Chisab, and S. N. Abed, "Evaluation study of radio frequency radiation effects from cell phone towers on human health," *Radioelectron. Commun. Syst.*, vol. 64, no. 3, pp. 155–164, 2021.
7. M. A. Mohammed *et al.*, "Benchmarking methodology for selection of optimal COVID-19 diagnostic model based on entropy and TOPSIS methods," *IEEE Access*, vol. 8, pp. 99115–99131, 2020, doi: 10.1109/ACCESS.2020.2995597.
8. Z. T. Al-Qaysi *et al.*, "Systematic review of training environments with motor imagery brain–computer interface: Coherent taxonomy, open issues and recommendation pathway solution," *Health Technol. (Berl).*, vol. 11, no. 4, pp. 783–801, 2021.
9. B. M. Albaker, "A deep reinforcement learning-based adaptive control strategy for UAVs in dynamic and complex environments," *Al-Iraqia J. Sci. Eng. Res.*, vol. 4, no. 1, pp. 77–88, 2025.
10. X. Liu *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis," *lancet Digit. Heal.*, vol. 1, no. 6, pp. e271–e297, 2019.
11. L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Sci. Rep.*, vol. 9, no. 1, p. 12495, 2019.
12. X. Xu *et al.*, "A deep learning system to screen novel coronavirus disease 2019 pneumonia," *Engineering*, vol. 6, no. 10, pp. 1122–1129, 2020.
13. A. H. Alamoodi *et al.*, "New extension of fuzzy-weighted zero-inconsistency and fuzzy decision by opinion score method based on cubic pythagorean fuzzy environment: A benchmarking case study of sign language recognition systems," *Int. J. Fuzzy Syst.*, pp. 1–18, 2022.
14. A. A. J. Al-Hchaimi, N. Bin Sulaiman, M. A. Bin Mustafa, M. N. Bin Mohtar, S. L. B. M. Hassan, and Y. R. Muhsen, "A comprehensive evaluation approach for efficient countermeasure techniques against timing side-channel attack on MPSoC-based IoT using multi-criteria decision-making methods," *Egypt. Informatics J.*, vol. 24, no. 2, pp. 351–364, 2023.
15. Y. R. Muhsen, N. A. Husin, M. B. Zolkepli, and N. Manshor, "A systematic literature review of fuzzy-weighted zero-inconsistency and fuzzy-decision-by-opinion-score-methods: assessment of the past to inform the future," *J. Intell. Fuzzy Syst.*, no. Preprint, pp. 1–22, 2023.
16. A. A. J. Al-Hchaimi, N. Bin Sulaiman, M. A. Bin Mustafa, M. N. Bin Mohtar, S. L. B. Mohd, and Y. R. Muhsen, "Evaluation approach for efficient countermeasure techniques against denial-of-service attack on MPSoC-based IoT using multi-criteria decision-making," *IEEE Access*, 2022.
17. X. Chew, K. W. Khaw, A. Alnoor, M. Ferasso, H. Al Halbusi, and Y. R. Muhsen, "Circular economy of medical waste: novel intelligent medical waste management framework based on extension linear diophantine fuzzy FDOSM and neural network approach," *Environ. Sci. Pollut. Res.*, pp. 1–27, 2023.
18. N. A. Husin, A. A. Abdulsaeed, Y. R. Muhsen, A. S. Zaidan, A. Alnoor, and Z. R. Al-mawla, "Evaluation of Metaverse tools based on privacy model using fuzzy MCDM approach," in *International Multi-Disciplinary Conference-Integrated Sciences and Technologies*, Springer, 2023, pp. 1–20.
19. M. S. Al-Samarraay *et al.*, "A new extension of FDOSM based on pythagorean fuzzy environment for evaluating and benchmarking sign language recognition systems," *Neural Comput. Appl.*, vol. 34, no. 6, pp. 4937–4955, 2022, doi: 10.1007/s00521-021-06683-3.
20. M. M. Salih, O. S. Albahri, A. A. Zaidan, B. B. Zaidan, F. M. Jumaah, and A. S. Albahri, "Benchmarking of AQM methods of network congestion control based on extension of interval type-2 trapezoidal fuzzy decision by opinion score method," *Telecommun. Syst.*, vol. 77, no. 3, pp. 493–522, 2021.
21. S. S. Joudar, A. S. Albahri, and R. A. Hamid, "Intelligent triage method for early diagnosis autism spectrum disorder (ASD) based on integrated fuzzy multi-criteria decision-making methods," *Informatics Med. Unlocked*, vol. 36, p. 101131, 2023.
22. R. T. Mohammed *et al.*, "Determining importance of many-objective optimisation competitive algorithms evaluation criteria based on a novel fuzzy-weighted zero-inconsistency method," *Int. J. Inf. Technol. Decis. Mak.*, vol. 21, no. 01, pp. 195–241, 2022.
23. M. M. Salih and R. A. Yousif, "Fuzzy-weightedzero-inconsistency, FWZIC, multi-criteria decision making, MCDM, weighting methods, fuzzy set," *Iraqi J. Comput. Sci. Math.*, vol. 5, no. 3, p. 3, 2024.
24. A. Kaya, D. Pamucar, H. E. Gürler, and M. Ozcalici, "Determining the financial performance of the firms in the Borsa Istanbul sustainability index: integrating multi criteria decision making methods with simulation," *Financ. Innov.*, vol. 10, no. 1, p. 21, 2024.

25. M. M. Salih, B. B. Zaidan, and A. A. Zaidan, "Fuzzy decision by opinion score method," *Appl. Soft Comput.*, vol. 96, p. 106595, 2020, doi: https://doi.org/10.1016/j.asoc.2020.106595.
26. M. A. Alsalem *et al.*, "Based on T-spherical fuzzy environment: A combination of FWZIC and FDOSM for prioritising COVID-19 vaccine dose recipients," *J. Infect. Public Health*, vol. 14, no. 10, pp. 1513–1559, 2021, doi: https://doi.org/10.1016/j.jiph.2021.08.026.
27. M. Deveci, I. Gokasar, and P. R. Brito-Parada, "A comprehensive model for socially responsible rehabilitation of mining sites using Q-rung orthopair fuzzy sets and combinative distance-based assessment," *Expert Syst. Appl.*, vol. 200, p. 117155, 2022, doi: https://doi.org/10.1016/j.eswa.2022.117155.
28. F. Al-Sharqi, A. Al-Quran, and M. U. Romdhini, "Decision-making techniques based on similarity measures of possibility interval fuzzy soft environment," *Iraqi J. Comput. Sci. Math.*, vol. 4, no. 4, pp. 18–29, 2023.
29. S. Wan, "Power average operators of trapezoidal intuitionistic fuzzy numbers and application to multi-attribute group decision making," *Appl. Math. Model.*, vol. 37, no. 6, pp. 4112–4126, 2013.
30. S. Rezvani, "Ranking method of trapezoidal intuitionistic fuzzy numbers," *Ann. Fuzzy Math. Informatics*, vol. 5, no. 3, pp. 515–523, 2013.
31. X. Li and X. Chen, "Multi-criteria group decision making based on trapezoidal intuitionistic fuzzy information," *Appl. Soft Comput.*, vol. 30, pp. 454–461, 2015.
32. C. Kahraman, S. Cebi, S. C. Onar, and B. Oztaysi, "A novel trapezoidal intuitionistic fuzzy information axiom approach: An application to multicriteria landfill site selection," *Eng. Appl. Artif. Intell.*, vol. 67, pp. 157–172, 2018.
33. Z. Alam, Y. Ali, and D. Pamucar, "Elevating Pakistan's flood preparedness: a fuzzy multi-criteria decision making approach," *Financ. Innov.*, vol. 10, no. 1, p. 140, 2024.
34. M. Akram, W. A. Dudek, and F. Ilyas, "Group decision-making based on pythagorean fuzzy TOPSIS method," *Int. J. Intell. Syst.*, vol. 34, no. 7, pp. 1455–1475, 2019.
35. M. E. Alqaysi, A. S. Albahri, and R. A. Hamid, "Hybrid diagnosis models for autism patients based on medical and sociodemographic features using machine learning and multicriteria decision-making (MCDM) techniques: An evaluation and benchmarking framework," vol. 2022, no. ii, 2022.
36. W. G. Ahmed Abbas Jasim Al-Hchaimi, Yousif Raad Muhsen and A. A. Entisar Soliman Alkayal, Riyadh Rahef Nuiaa Al Ogaili, and Zaid Abdi Alkareem Alyasseri, "Prioritizing network-on-chip routers for countermeasure techniques against flooding denial-of-service attacks: A fuzzy multi-criteria decision-making approach," *Comput Model Eng Sci*, vol. 142, no. 3, 2025.
37. A. Alnoor, Y. R. Muhsen, N. A. Husin, X. Chew, M. B. Zolkepli, and N. Manshor, "Z-cloud rough fuzzy-based PIPRECIA and CoCoSo integration to assess agriculture decision support tools," *Int. J. Fuzzy Syst.*, vol. 27, no. 1, pp. 190–203, 2025, doi: 10.1007/s40815-024-01771-7.
38. M. Baydaş, M. Yılmaz, Ž. Jović, Ž. Stević, S. E. G. Özuyar, and A. Özçil, "A comprehensive MCDM assessment for economic data: success analysis of maximum normalization, CODAS, and fuzzy approaches," *Financ. Innov.*, vol. 10, no. 1, p. 105, 2024.
39. A. K. Oleiwi, H. M. Saleh, A. M. Mahmood, and I. AVCI, "A survey of MCDM-based software engineering method," *Babylonian J. Math.*, vol. 2024, pp. 13–18, 2024.
40. A. S. Albahri *et al.*, "Integration of fuzzy-weighted zero-inconsistency and fuzzy decision by opinion score methods under a q-rung orthopair environment: A distribution case study of COVID-19 vaccine doses," *Comput. Stand. Interfaces*, vol. 80, no. March 2021, p. 103572, 2022, doi: 10.1016/j.csi.2021.103572.
41. M. S. Al-Samarraay *et al.*, "A new extension of FDOSM based on Pythagorean fuzzy environment for evaluating and benchmarking sign language recognition systems," *Neural Comput. Appl.*, vol. 34, no. 6, pp. 4937–4955, 2022, doi: 10.1007/s00521-021-06683-3.
42. A. H. Alamoodi *et al.*, "Based on neutrosophic fuzzy environment: a new development of FWZIC and FDOSM for benchmarking smart e-tourism applications," *Complex Intell. Syst.*, vol. 8, no. 4, pp. 3479–3503, 2022, doi: 10.1007/s40747-022-00689-7.
43. U. S. Mahmoud *et al.*, "A methodology of DASs benchmarking to support industrial community characteristics in designing and implementing advanced driver assistance systems within vehicles," 2021.
44. M. S. Al-samarraay *et al.*, "Extension of interval-valued pythagorean FDOSM for evaluating and benchmarking real-time SLRSs based on multidimensional criteria of hand gesture recognition and sensor glove perspectives," *Appl. Soft Comput.*, vol. 116, p. 108284, 2022, doi: 10.1016/j.asoc.2021.108284.
45. A. Wadhawan and P. Kumar, "Deep learning-based sign language recognition system for static signs," *Neural Comput. Appl.*, vol. 32, no. 12, pp. 7957–7968, 2020.
46. O. S. Albahri *et al.*, "Multidimensional benchmarking of the active queue management methods of network congestion control based on extension of fuzzy decision by opinion score method," *Int. J. Intell. Syst.*, vol. 36, no. 2, pp. 796–831, 2021.
47. R. T. Mohammed *et al.*, "A decision modeling approach for smart e-tourism data management applications based on spherical fuzzy rough environment," *Appl. Soft Comput.*, vol. 143, p. 110297, 2023.
48. S. Qahtan *et al.*, "Evaluation of agriculture-food 4.0 supply chain approaches using fermatean probabilistic hesitant-fuzzy sets based decision making model," *Appl. Soft Comput.*, p. 110170, 2023.
49. H. A. Ibrahim, A. A. Zaidan, S. Qahtan, and B. B. Zaidan, "Sustainability assessment of palm oil industry 4.0 technologies in a circular economy applications based on interval-valued pythagorean fuzzy rough set-FWZIC and EDAS methods," *Appl. Soft Comput.*, vol. 136, p. 110073, 2023.
50. A. T. Demir and S. Moslem, "Evaluating the effect of the COVID-19 pandemic on medical waste disposal using preference selection index with CRADIS in a fuzzy environment," *Heliyon*, vol. 10, no. 5, 2024.
51. S. Seker and N. Aydin, "Hydrogen production facility location selection for Black Sea using entropy based TOPSIS under IVPF environment," *Int. J. Hydrogen Energy*, vol. 45, no. 32, pp. 15855–15868, 2020.
52. A. U. Ibrahim, M. Ozsoz, S. Serte, F. Al-Turjman, and P. S. Yakoi, "Pneumonia classification using deep learning from chest X-ray images during COVID-19," *Cognit. Comput.*, pp. 1–13, 2021.
53. I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: Automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, pp. 635–640, 2020.
54. Y. Pathak, P. K. Shukla, A. Tiwari, S. Stalin, and S. Singh, "Deep transfer learning based classification model for COVID-19 disease," *Irbm*, vol. 43, no. 2, pp. 87–92, 2022.
55. M. J. Horry *et al.*, "COVID-19 detection through transfer learning using multimodal imaging data," *Ieee Access*, vol. 8, pp. 149808–149824, 2020.

56. P. Sukumar and R. K. Gnanamurthy, "Computer aided detection of cervical cancer using pap smear images based on adaptive neuro fuzzy inference system classifier," *J. Med. Imaging Heal. Informatics*, vol. 6, no. 2, pp. 312–319, 2016.
57. Y. Xie, F. Xing, X. Kong, H. Su, and L. Yang, "Beyond classification: structured regression for robust cell detection using convolutional neural network," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 358–365.
58. A. A. Nafea, S. A. Alameri, R. R. Majeed, M. A. Khalaf, and M. M. AL-Ani, "A short review on supervised machine learning and deep learning techniques in computer vision," *Babylonian J. Mach. Learn.*, vol. 2024, pp. 48–55, 2024.
59. A. K. Abed, "Utilizing artificial intelligence in cybersecurity: A study of neural networks and support vector machines," *Babylonian J. Netw.*, vol. 2025, pp. 14–24, 2025.
60. A. I. Gide and A. A. Mu'azu, "A real-time intrusion detection system for dos/ddos attack classification in IoT networks using KNN-neural network hybrid technique," *Babylonian J. Internet Things*, vol. 2024, pp. 60–69, 2024.
61. S. Marina, "Improving diagnostic accuracy of brain tumor MRI classification using generative AI and deep learning techniques," *Babylonian J. Artif. Intell.*, vol. 2025, pp. 55–63, 2025.
62. J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv Prepr. arXiv2006.11988*, 2020.
63. E. Deniz, A. Şengür, Z. Kadiroğlu, Y. Guo, V. Bajaj, and Ü. Budak, "Transfer learning based histopathologic image classification for breast cancer detection," *Heal. Inf. Sci. Syst.*, vol. 6, pp. 1–7, 2018.
64. A. Abubakar, H. Ugail, A. M. Bukar, A. A. Aminu, and A. Musa, "Transfer learning based histopathologic image classification for burns recognition," in *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, IEEE, 2019, pp. 1–6.
65. A. Alnoor *et al.*, "Toward a sustainable transportation industry: Oil company benchmarking based on the extension of linear diophantine fuzzy rough sets and multicriteria decision-making methods," *IEEE Trans. Fuzzy Syst.*, pp. 1–11, 2022, doi: 10.1109/TFUZZ.2022.3182778.
66. M. A. Alsalem *et al.*, "Based on T-spherical fuzzy environment: A combination of FWZIC and FDOSM for prioritising COVID-19 vaccine dose recipients," *J. Infect. Public Health*, vol. 14, no. 10, pp. 1513–1559, 2021.
67. A. Hb. Alamoodi *et al.*, "New extension of fuzzy-weighted zero-inconsistency and fuzzy decision by opinion score method based on cubic pythagorean fuzzy environment: A case study of sign language recognition systems," *Int. J. Fuzzy Syst.*, pp. 1–18, 2022.
68. X. Zhang and P. Liu, "Method for aggregating triangular fuzzy intuitionistic fuzzy information and its application to decision making," *Technol. Econ. Dev. Econ.*, vol. 16, no. 2, pp. 280–290, 2010.
69. S. Qahtan, H. A. Alsattar, A. A. Zaidan, M. Deveci, D. Pamucar, and D. Delen, "Performance assessment of sustainable transportation in the shipping industry using a q-rung orthopair fuzzy rough sets-based decision making methodology," *Expert Syst. Appl.*, vol. 223, p. 119958, 2023.
70. S. Qahtan, H. A. Alsattar, A. A. Zaidan, M. Deveci, D. Pamucar, and L. Martinez, "A comparative study of evaluating and benchmarking sign language recognition system-based wearable sensory devices using a single fuzzy set," *Knowledge-Based Syst.*, vol. 269, p. 110519, 2023.