Journal Pre-proof

Systematic Review of Artificial Intelligence Techniques in the Detection and Classification of COVID-19 Medical Images in Terms of Evaluation and Benchmarking: Taxonomy Analysis, Challenges, Future Solutions and Methodological Aspects

O.S. Albahri, A.A. Zaidan, A.S. Albahri, B.B. Zaidan, K.H. Abdulkareem, Z.T. Al-qaysi, A.H. Alamoodi, A.M. Aleesa, M.A. Chyad, R.M. Alesa, L.C. Kim, M.M. Lakulu, A.B. Ibrahim, N.A. Rashid

PII: S1876-0341(20)30558-X

DOI: https://doi.org/10.1016/j.jiph.2020.06.028

Reference: JIPH 1398

To appear in: *Journal of Infection and Public Health*

Received Date: 16 March 2020

Revised Date: 6 June 2020

Accepted Date: 25 June 2020

Please cite this article as: Albahri OS, Zaidan AA, Albahri AS, Zaidan BB, Abdulkareem KH, Al-qaysi ZT, Alamoodi AH, Aleesa AM, Chyad MA, Alesa RM, Kim LC, Lakulu MM, Ibrahim AB, Rashid NA, Systematic Review of Artificial Intelligence Techniques in the Detection and Classification of COVID-19 Medical Images in Terms of Evaluation and Benchmarking: Taxonomy Analysis, Challenges, Future Solutions and Methodological Aspects, *Journal of Infection and Public Health* (2020), doi: https://doi.org/10.1016/j.jiph.2020.06.028

Systematic Review of Artificial Intelligence Techniques in the Detection and Classification of COVID-19 Medical Images in Terms of Evaluation and Benchmarking: Taxonomy Analysis, Challenges, Future Solutions and Methodological Aspects

O.S.Albahri[1], A.A.Zaidan[1*], A.S.Albahri[1], B.B.Zaidan[1], K. H.Abdulkareem[2], Z.T.Al-qaysi[1], A.H.Alamoodi[1], A.M.Aleesa[3], M.A.Chyad[1], R.M.Alesa[3], L.C.Kim[1], M.M.Lakulu[1], A.B.Ibrahim[1] and N.A.Rashid[1]

[1]Department of Computing, FSKIK, Universiti Pendidikan, Tanjung Malim 35900, Malaysia
[2]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Malaysia
[3]Faculty of Electronic and Electrical Engineering, Universiti Tun Hussein Onn, 86400, Batu, Pahat, Johor, Malaysia

*Corresponding authors: aws.alaa@gmail.com/aws.alaa@fskik.upsi.edu.my

Abstract

This study presents a systematic review of artificial intelligence (AI) techniques used in the detection and classification of coronavirus disease 2019 (COVID-19) medical images in terms of evaluation and benchmarking. Five reliable databases, namely, IEEE Xplore, Web of Science, PubMed, ScienceDirect and Scopus were used to obtain relevant studies of the given topic. Several filtering and scanning stages were performed according to the inclusion/exclusion criteria to screen the 36 studies obtained; however, only 11 studies met the criteria. Taxonomy was performed, and the 11 studies were classified on the basis of two categories, namely, review and research studies. Then, a deep analysis and critical review were performed to highlight the challenges and critical gaps outlined in the academic literature of the given subject. Results showed that no relevant study evaluated and benchmarked AI techniques utilised in classification tasks (i.e. binary, multi-class, multi-labelled and hierarchical classifications) of COVID-19 medical images. In case evaluation and benchmarking will be conducted, three future challenges will be encountered, namely, multiple evaluation criteria within each classification task, trade-off amongst criteria and importance of these criteria. According to the discussed future challenges, the process of evaluation and benchmarking AI techniques used in the classification of COVID-19 medical images considered multi-complex attribute problems. Thus, adopting multi-criteria decision analysis (MCDA) is an essential and effective approach to tackle the problem complexity. Moreover, this study proposes a detailed methodology for the evaluation and benchmarking of AI techniques used in all classification tasks of COVID-19 medical images as future directions; such methodology is presented on the basis of three sequential phases. Firstly, the identification procedure for the construction of four decision matrices, namely, binary, multi-class, multi-labelled and hierarchical, is presented on the basis of the intersection of evaluation criteria of each classification task and AI classification techniques. Secondly, the development of the MCDA approach for benchmarking AI classification techniques is provided on the basis of the integrated analytic hierarchy process and VlseKriterijumska Optimizacija I Kompromisno Resenje methods. Lastly, objective and subjective validation procedures are described to validate the proposed benchmarking solutions.

Keywords: COVID-19, Medical Image, Artificial Intelligence, Evaluation, Benchmarking, Decision-making, MCDA.

1. Introduction

In the last days of 2019, a group of patients infected with a novel coronavirus disease (coronavirus disease 2019, COVID-19) was recognised in Wuhan, China. Since then, the contagions of COVID-19 have spread around the world. COVID-19 affects people in different ways. Most infected patients develop common symptoms (i.e. fever, fatigue and dry cough) [1], and others may experience additional symptoms (i.e. aches and pains, nasal congestion, runny nose, sore throat and diarrhoea) [2]. COVID-19 exposed weaknesses in the healthcare system of many countries, and the inability of healthcare systems to manage patients has caused anxiety. One of the important reasons behind the rapid spread of COVID-19 is the lack of specificity in clinical detection methods [3]. Molecular approaches such as quantitative real-time reverse transcription–polymerase chain reaction (rRT-PCR) [4] and other methods such as serologic tests [5] and viral throat swab testing [6] are necessary and widely utilised for the detection of COVID-19. However, studies have shown that chest radiographs (X-rays) [7] and chest computed tomography (CT) scans [8] can assist and reveal anomalies indicative of different lung diseases, including COVID-19. CT scan and X-ray tests could be utilised as a primary detection tool to evaluate the severity of COVID-19, monitor the emergency case of infected patients and predict COVID-19 progression [9]. However, time is often limited in such emergencies and does not allow these experiments to be performed using

existing traditional manual diagnosis [10]. These procedures require a specialist doctor and are susceptible to human error during testing or reading and interpreting findings, which are not acceptable in crucial cases. Given the recent spread of COVID-19, hospitals are filled with numerous patients who are either improving from the viral infection or becoming worse (dying) [11]. In this case, CT scan and X-ray tests should be performed with maximum speed and efficiency to save as many lives as possible [9]. The role of intelligent technologies would effectively help in the diagnosis and classification processes [7].

The use of artificial intelligence (AI) has increased in different fields, especially in medical detection [12]. AI has been widely used to gain more accurate detection results and decrease the burden on the healthcare system [13]. It can decrease the decision time associated with the detection process of traditional methods [14]. The development of AI techniques to recognise the risks of epidemic diseases is considered a key factor in the improvement of the prediction, prevention and detection of future global health risks [15]. Numerous types of AI classifiers have been reported by a few researchers with real COVID-19 datasets with different case studies and targets [9]. Although AI techniques can be beneficial in the diagnosis and classification of COVID-19, selecting the appropriate AI technique that can produce accurate results is challenging [16, 17]. The large diversity amongst available AI techniques creates difficulties in deciding which of them to use in the development of COVID-19 diagnosis and classification particularly when there is no dedicated AI technique that is far better than the other. In addition, the majority of these techniques suffer from low accuracy and computational efficiency [18]. On the other hand, the difficult part is associated with the evaluation and comparison because of the multiple evaluation criteria and conflict between them are increasing the challenge [19].

The evaluation and benchmarking procedures of AI techniques are critical in acquiring a technique that can produce the best results [17, 20]. A similar process is essential since there will be affected on the persons who suspected with COVID-19 and medical organisation due to this process could result in losing their life and spreading the virus amongst others. In order to evaluate and benchmark AI classification techniques that can be used in the detection of COVID-19 medical images, several requirements guarantee the reliability of these techniques given that they are associated with patients' lives. However, two main questions can be encountered in this process. Firstly, what are the appropriate criteria that could be used in the evaluation? Secondly, what is the correct benchmark procedure that could be used to select a suitable AI technique amongst others?

Therefore, the present study aims to (i) shed light and systematically review the research efforts of emerging and new technologies of COVID-19 medical image detection based on AI approach; (ii) map related studies into coherent taxonomy and highlight the AI techniques, datasets, case studies and AI classification types used; (iii) highlight and analyse different aspects such as research gabs and future challenges with respect to evaluation and benchmarking; and (iv) propose a potential pathway solution with detailed methodology to tackle the identified research gabs and future challenges of evaluation and benchmarking of AI classification techniques used in COVID-19 medical image detection. The remaining sections of this study are presented as follows. Section 2 presents the methods used in reviewing systematic literatures of the topic. Section 3 presents the taxonomy analysis highlight points of the included final set of studies. Section 4 presents a critical review and analysis of the identified studies. Section 5 presents the future challenges related to the evaluation and benchmarking of AI classification techniques used in COVID-19 medical image detection. Section 6 presents a proposal of potential future solutions for the identified research gabs. Section 7 provides the methodology of the proposed solutions. Finally, Section 8 presents the conclusion.

2. Methods

This study is based on a systematic literature review (SLR), which has been recognised for its role in acquiring a sufficient understanding with regard to a topic of interest [21, 22]. SLR has also been highlighted for its remarkable structured analytical methods for research synthesis and its ability to accommodate different types of studies from various scientific disciplines [24, 25]. During the process, different academic digital databases are utilised to extract relevant literature, including (1) ScienceDirect, which offers different scientific literature across all domains; (2) Scopus, which offers sufficient coverage of literature from all disciplines; (3) IEEE, which is recognised for its scientific reliability of covering multi-disciplinary sciences and engineering and computer science literature; (4) Web of Science, which demonstrates high coverage of different literature topics and studies across all domains; and (5) PubMed, which also covers a variety of disciplines with multi-disciplinary focus on medicine- and technology-related literature [26-28]. These databases are deemed sufficient to cover the latest and most reliable literature for COVID-19 diagnosis and detection. The studies extracted from these databases are relevant and reliable to understand the role of intelligent systems (i.e. AI) and their involvement in scientist's efforts with relation to COVID-19. The literature search was comprehensively conducted on the five major databases in a span of 10 years between 2010 and May 5, 2020. The selection of the databases was due to their scientific reliability, soundness and coverage for literature from various domains with regard to deep learning efforts and COVID-19. Boolean operators were utilised during the process to gather as much relevant literature as possible. The first group of keywords was meant for intelligent systems and their

relevant keywords, the second group was meant for COVID-19 relevant keywords, whilst the third group was meant for medical images with different relevant keywords to make sure all literature associated with the three groups are included. In this SLR, different criteria were enforced for the selection of related literature. All articles were selected if they were English and conducted between 2010 and May 5, 2020. For the publication types, all articles were selected if they were journal, conference or review papers [29-31]. As far as the topic of interest is concerned, this SLR only selected publications that discuss any form of AI and COVID-19 using medical images. The exact query is presented in Figure 1.

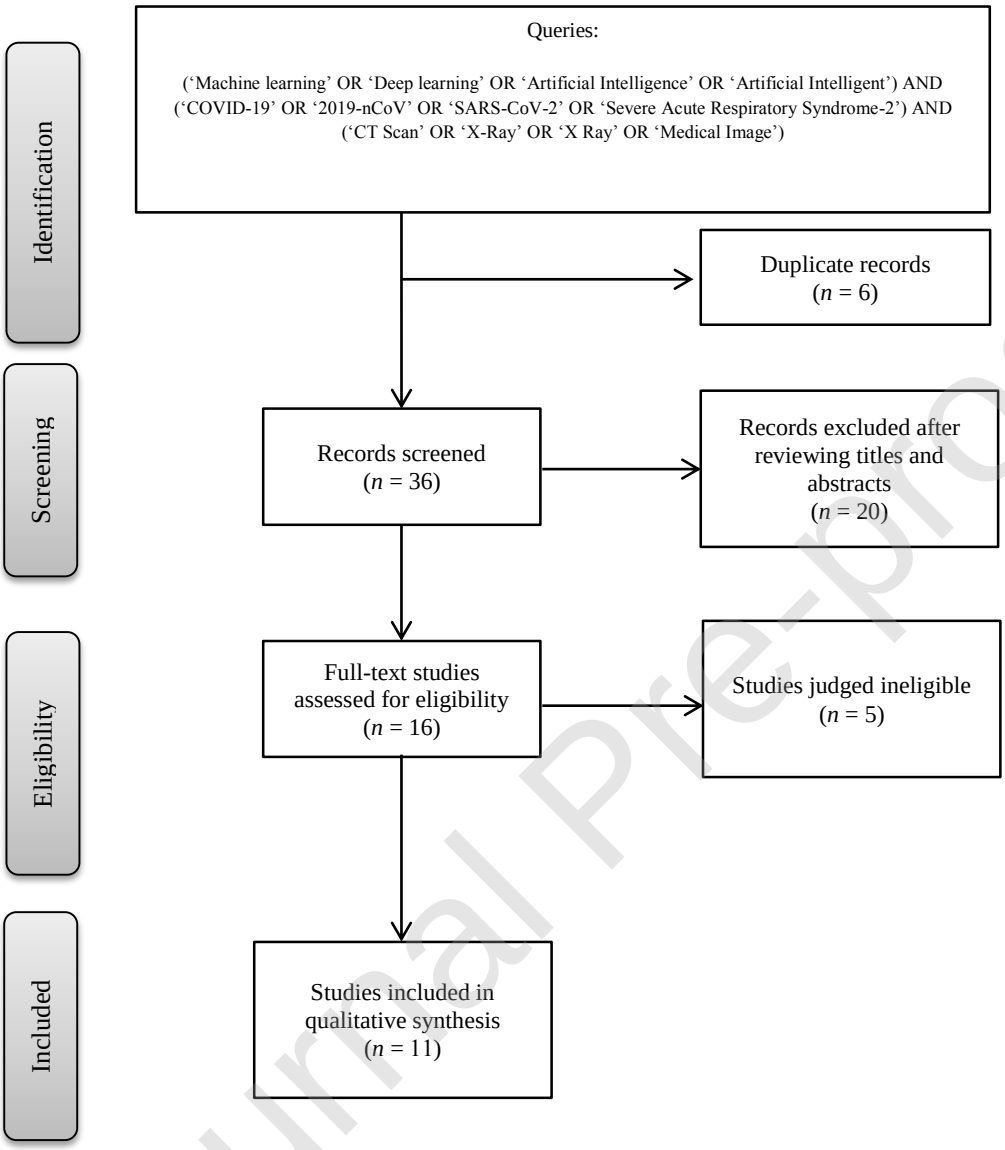

Figure 1. Method of SLR of the study topic

The search was conducted in the middle of May 2020 using the advanced search boxes of the five digital databases. The initial search yielded 36 publications after duplication process, which excluded a total of six duplicated records. Next, the titles and abstracts of the publications were scanned. Twenty articles were excluded as they failed to meet the criteria. The remaining articles underwent another round of screening through full-text reading to investigate the relevancy of the selected articles from the previous phase and determine whether they are suitable to be included in the final set. After this process, five articles were excluded, and only 11 articles met all criteria and were deemed suitable for inclusion in this review. Furthermore, the key demographic statistical findings from the articles are presented on the basis of two aspects, namely, database used and countries (Figure 2).

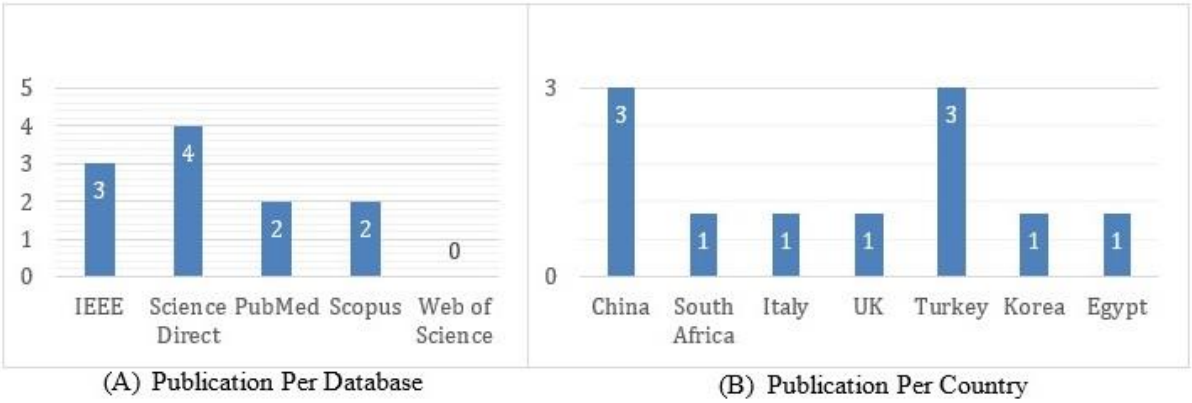

Figure 2. Statistics of the included studies by databases and countries

All these 11 articles identified from the literature are scattered over the databases and the countries. For the databases, four studies were obtained from ScienceDirect, three from IEEE, two from PubMed, and two from Scopus. The only database with no identified articles was Web of Science. As for the countries of the corresponding authors, three studies came from China, three from Turkey, one from South Africa, one from Italy, one from UK, one from Korea, and one from Egypt.

3. Results and Discussion

This section elaborates the final set of articles (11 articles) that have been collected in this systematic review regarding AI techniques used in the detection and classification of COVID-19 medical images. This final set was divided into two clusters, namely, review cluster and research cluster. The taxonomy and classification of the related articles are shown in Figure 3.

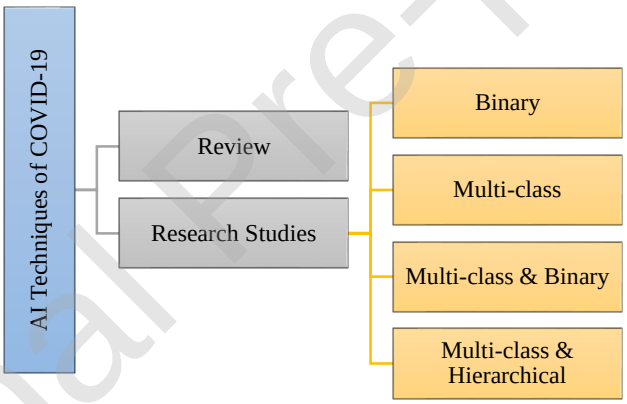

Figure 3. Taxonomy of research literature on AI techniques used in the detection and classification of COVID-19 medical images

3.1. Review

The primary aim of reviewing articles on AI techniques used in the detection and classification of COVID-19 medical images is to understand the current thinking in this field and justify the need for future research on related topics that have been overlooked or understudied. This cluster contained only one article. In [9], the study reviewed the rapid responses in the community of medical imaging (empowered by AI) towards COVID-19. The authors emphasised that AI-empowered image acquisition can significantly help automate the scanning procedure and reshape the workflow with minimal contact to patients, providing the best protection to the imaging technicians. They focused on the entire pipeline of medical imaging and analysis techniques involved with COVID-19, including image acquisition, segmentation, diagnosis and follow-up, using the integration of AI with X-ray and CT images.

3.2. Research Studies

The second cluster focused on research studies and contained 10 articles, which consist of four sub-clusters: binary, multi-class, integrated multi-class and binary, and integrated hierarchical and multi-class.

3.2.1. Binary Classification

The flat classification refers to binary classification problems with only two different classes. One article involved this sub-cluster. The study of [32] demonstrated the ability of deep learning method in the diagnosis of COVID-19 on the basis of medical images acquired by CT. Regarding the class labels that have been used in identifying the existence of the infection, this study relied on false-negative (FN) results, which jeopardise the epidemic from being prevented and controlled and affect decisions for health monitoring or discharging. The dataset utilised was made out of the information of 10 patients. Out of the 10 negative cases, two were positively identified for COVID-19 by utilising the rRT-PCR test. The previous clearly indicated and yielded almost 20% FN rate for rRT-PCR.

3.2.2. Multi-class Classification

There are numerous issues and challenges linked to multi-class classification. However, there is one output for a sample. This sub-cluster includes four different publication works identified. The first work [33] involved the development of a scoring tool aimed at COVID-19 severity. Such tool was proven to be important in assisting healthcare workers in identifying and determining which patients suspected or confirmed for COVID-19 are in high need for respiratory interventions. The research utilised a tool for assigning patients into categories according to severity in line with the classifications of the WHO, namely, severe and moderate/mild. In different terms, patients who are at the critical stage need ventilation, other patients in the severe stage need oxygen, whereas those patients in the moderate stage do not need oxygen despite having pneumonia. For patients in the mild stage, they only have upper respiratory tract disease. In addition, the dataset utilised was gathered from 13,500 COVID-19 patients. According to an early assessment, the tool developed correctly classified 93.6% of patients, underestimated 0.8% of patient severities and overestimated 5.7%. Another work [34] introduced COVIDiagnosis-Net, which is an AI detection approach for COVID-19. The approach is based on deep SqueezeNet with Bayes optimisation, which can help detect COVID-19. The deep learning technique exhibited 98.3% accuracy in detecting normal, pneumonia and COVID-19 cases. The technique also had a 100% accuracy for the single detection of COVID-19 amongst other classes. [35] proposed a patch-based technique with convolutional neural network. The technique makes use of a small number of training parameters for the diagnosis of COVID-19. The work was inspired by a statistical analysis for potential imaging biomarkers of chest X-rays. Their results indicated that pre-processing for normalisation of data helped in the processing of cross-database and significantly improved the accuracy of segmentation (Jaccard similarity coefficients from 0.932 to 0.943, $p < 0.001$). According to the results, pre-processing was a significant aspect in ensuring the performance of the segmentation in the cross-database. In [36], CoVID-19 was identified with the use of MobileNetV2 and SqueezeNet, a deep learning technique, in addition to feature sets obtained by the techniques. They were processed using the social mimic optimisation method. Fuzzy colour technique was used to restructure data classes as a pre-processing measure, and the structured images were stacked with the original images. Thereafter, efficient features were grouped and classified with the use of support vector machines (SVMs) with an overall classification rate of 99.27%.

3.2.3. Integrated Multi-class and Binary Classifications

This sub-cluster contains three articles that focused on integrated multi-class and binary classification problems. [37] emphasised the deployment of AI to support the work of a radiologist. They indicated that the application of AI in COVID-19 infection will allow monitoring the course of the disease. [38] indicated the importance of AI in maintaining the spread of COVID-19. [7] presented a model for detecting COVID-19 by using 125 X-ray images in accurately diagnosing binary classification (COVID vs. no findings), in addition to multi-class classification (COVID vs. no findings vs. pneumonia). The accuracy of the model was 98.08% for binary classes and 87.02% for multi-class cases.

3.2.4. Integrated Hierarchical and Multi-class Classifications

Another classification problem type is hierarchical classification where the learning output is identified over special class taxonomy. [39] defined hierarchical classification as follows: 'the input is to be classified into one, and only one, each class which are be divided into subclasses or grouped into superclasses. The hierarchy is defined and cannot be changed during classification. Hierarchical classification can be transformed into flat classification.' This sub-cluster contains two articles that focused on integrated hierarchical and multi-class classification problems. [40] identified COVID-19 pneumonia from various healthy lung types and developed a classification approach, which takes into consideration multi-class and hierarchical perspectives. In addition, resampling algorithms were used for re-balancing the distribution of the classes. The approach acquired a macro-average F1 score of 0.65 with the use of multi-class method and F1 score of 0.89 for the identification of COVID-19 in hierarchical classification scenario. [41] developed a model with hybrid capability for detecting COVID-19 with the use of improved marine predators algorithm (IMPA) and ranking-based diversity reduction strategy to acquire particle numbers that are not capable of finding suitable solution within a consecutive number of iterations. Nine chest X-ray images were utilised for the validation of IMPA performance. The

threshold levels were between 10 and 100 and compared with five algorithms: (1) whale optimisation algorithm, (2) salp swarm algorithm, (3) sine cosine algorithm, (4) equilibrium optimiser and (5) Harris hawks algorithm. Results showed that the hybrid model proposed based on the experiment outperforms all other algorithms for a range of metrics. Furthermore, on all threshold levels, the performance was convergent in Structural Similarity Index and Universal Quality Index metrics.

3.3. Highlight Points

The academic literature in the research cluster was further discussed from different points of view on the basis of three perspectives, namely, dataset, AI technique and case study used. The dataset that has been used or proposed to be used in the development and evaluation of the COVID-19 diagnosis system was categorised as primary or secondary. The primary dataset is the dataset that was collected during the research and approved by the ethical approval committee. Conversely, the secondary dataset was acquired online (public dataset) and published by researchers to help other researchers test their AI methods and techniques. Moreover, regarding the AI techniques, this study identified the AI algorithms into traditional machine learning algorithms, such as SVM and decision tree, or deep learning algorithms such as convolutional and deep neural networks. Table 1 presents a summary of the studies described in this cluster, focusing on their most important characteristics for COVID-19 diagnoses such as the type of datasets used and summarising AI techniques utilised to solve the case study problems for detecting the COVID-19 medical images.

Table 1. Summary of the perspectives of works described in research cluster studies

	Type of Datasets		**AI Techniques**		
Ref.	Primary Data	Secondary Data	Traditional Machine Learning Techniques	Deep learning Techniques	**Case Study**
[32]	χ	χ	χ	√	CT Scan
[42]	√	χ	√	χ	CT Scan
[37]	√	√	√	√	CT Scan
[38]	√	√	√	√	X-ray
[7]	χ	√	χ	√	X-ray
[43]	χ	√	χ	√	X-ray
[44]	χ	√	χ	√	X-ray
[41]	χ	√	√	χ	X-ray
[36]	χ	√	√	√	X-ray
[40]	√	√	√	√	X-ray

AI approaches for the detection of COVID-19 are considered one of the latest and most trending topics due to the growing pandemic. It is difficult to represent the true state-of-the-art for this purpose considering that new works are emerging every day. Nevertheless, we concluded that majority of the literature aimed to investigate hybrid AI techniques by combining deep learning and traditional machine learning, which is contributed by different types of datasets. In addition, the standard image diagnosis tests for pneumonia are chest X-ray and CT scan. X-ray is more useful amongst these studies because it is cheaper, faster and more widespread than CT. The primary aims of the studies are to identify pneumonia caused by COVID-19 from other types using either X-ray or CT scan. Given that pneumonia can be structured as a hierarchy, a classification scheme considering the multi-class and hierarchical perspectives requires attention and leads to the best COVID-19 recognition rate. The reason behind this is that there is a hierarchy between the pathogens that cause pneumonia. However, only one study [40] considered hierarchical classification approach in the literature.

4. Critical Review and Analysis

On the basis of previous literature, classification tasks for COVID-19 were different in terms of aspects related to the accuracy of results, in spite of the differences of the overall performance. Previous literature was solely focused on accuracy enhancement, time reduction or even overall performance improvements for the classification. Furthermore, differences exist in previous literature with respect to classification techniques, phases and classification procedures. On the one hand, the developed COVID-19 classification techniques in the analysed studies provide three COVID-19 classification tasks (i.e. binary classification, multi-class classification and hierarchical classification). On the other hand, [39] indicated that all relevant label distribution in a classification problem changes, which explains why four classification types can be performed in the AI techniques, namely, binary, multi-class, multi-labelled and hierarchical classifications. Multi-labelled classification is described in [39] as follows: 'the input is to be classified into several of non-overlapping classes. When the learning task is document topic classification, multi-labelling is often referred as multi-topic classification. In the multi-labelled classification problem, categories are isolated and their relations are not considered important.' However, no study has provided multi-labelled classification for the detection of

COVID-19 medical images. This is considered the first research gab identified in the literature reviewed. Furthermore, the growing number of classification techniques developed for COVID-19 is considered a major problem for health organisations and other treatment centres. The reason behind that these medical organisations that aim to adopt classification techniques for detection of COVID-19 will be encountered a challenge on how to select the best and an appropriate classification technique that would provide an accurate and rapid detection of COVID-19 medical images. Apart from the disparity in COVID-19 classification techniques in terms of their overall performance, all results confirm the difficulty of making a decision to choose a better technique amongst others. In the analysed studies, there is no evidence or proposed solution confirmed to be superior over the rest. Moreover, although multi-labelled classification AI techniques used in the detection of COVID-19 have not been developed, they might be developed in the near future. In the case of this development, another important question will arise: 'which classification technique is appropriate for such purpose?' According to the included final set of articles that met the search query used, no study has provided a comprehensive evaluation and benchmarking solution for AI classification techniques (i.e. binary, multi-class, multi-labelled and hierarchical classifications) used in the detection of COVID-19 medical images. This is considered the second research gab identified in the literature reviewed. [17] recommended that an evaluation and benchmarking solution for multi-labelled and/or hierarchical classification techniques could be beneficial and essential to determine which AI technique is appropriate amongst others. To explain the detailed solution for the identified gabs, two problems should be discussed: 'what are the evaluation criteria used in each classification type (i.e. binary, multi-class, multi-labelled and hierarchical classifications), and what are the calculation processes of these criteria? Each of these classification methods has its own evaluation criterion. The calculation procedure for each evaluation criteria is completely different from each classification type [39],[17]. Thus, the evaluation and benchmarking procedure will be different within each classification method (the evaluation criteria and calculation procedures are specified in detail in the methodology section). This study attempts to fill the gap in the evaluation and benchmarking of different classification types that will be used in the detection of COVID-19. The proposed solution shall assist the administrations of health organisations to evaluate and benchmark COVID-19 AI classification techniques. It can also ensure that the selected classification techniques meet all necessary requirements. To provide such a solution, three specific challenges need to be addressed in the process of evaluation and benchmarking classification techniques, which are described in the next section.

5. Future Challenges of the Evaluation and Benchmarking of AI Classification Techniques Used in the Detection of COVID-19 Medical Images

In this section, three future challenges will be encountered in the processes of evaluation and benchmarking AI classification techniques used in the detection of COVID-19 medical images as discussed in the following subsections.

5.1. Challenge of Multiple Evaluation Criteria

As stated in the previous section, four categories of classification tasks are identified. Each category is different in terms of criteria type, where the calculation procedure is different for each evaluation criterion. Furthermore, the number of criteria is different within each classification category, for example, six evaluation criteria for binary classification, eight criteria for multi-class classification, four criteria for multi-labelled classification and six criteria for hierarchical classification [39]. In general, most evaluation processes for COVID-19 classification techniques need to consider more than one criterion. For example, the reliability of classification techniques can be measured on the basis of a confusion matrix that contains four parameters: true positive (TP), false positive (FP), true negative (TN) and FN. In other words, the rate of correct and incorrect classified samples is compared between actual class and the predicted class. Thus, this status will affect the results if only one or a full set of parameters is considered in the evaluation process. However, in this regard, there are no suggested solutions to handle these particular issues in terms of evaluation and benchmarking of COVID-19 AI classification techniques. Furthermore, the recommended solution must consider the issue that the evaluation of COVID-19 classification techniques is based on multiple evaluation criteria and consider the difference amongst classification tasks in terms of type of criteria.

5.2. Challenge of Criteria Trade-off

The issue of trade-off is defined as a situation when a reliability or aspect of something decreases whilst the reliability or aspect of another increases. According to the nature of the evaluation criteria used in AI techniques, different types of trade-off utilised by researchers for different criteria were performed, which in turn were confusing for decision-makers. In addition, in the scope of this study, the different use ratio in different criteria demonstrated effect that explains the conflict on other criteria utilised by researchers. Thus, the evaluation criteria conflict for COVID-19 classification shows important challenges in our intention towards creating a COVID-19 classification approach. Fundamentally, these types of challenges are due to terms confliction, especially the one between the criteria and the data. Thus, it is crucial to realise the advantages and disadvantages of a particular choice whilst making a decision. The trade-off term is frequently used in the

context of evaluation, where the process of selection acts as a decision-maker [45-47]. The trade-off, also known as conflicting criteria problem, between the evaluation criteria concentrated on the application reliability, time complexity for the COVID-19 classification procedure and error rate within the dataset in the benchmarking and evaluation of AI classification techniques used in COVID-19. With the aim of evaluating the COVID-19 classification techniques, these sorts of criteria are considered main necessities. The reliability should possess a high rate; time complexity to conduct the output that also need to below. In addition, the apparent error rate from the training of the dataset has to be simultaneously low. The generated conflicting data are monitored because the matrix of parameter section contains TP, FP, TN and FN, which displays the rise in TP and TN when FP and FN are minimised [48, 49]. This phenomenon shows an apparent conflict amongst the probability criteria. These parameters have a considerable effect on some of the remaining criteria values because some of the criteria rely on the values of these four parameters. Therefore, the process of evaluation and benchmarking must take into consideration such requirements. As a result, a new approach for the evaluation that handles all conflict criteria and data problems should emerge, and this method should be flexible. However, in this regard, there are no suggested solutions to handle these particular issues.

5.3. Challenge of Criteria Importance

Another challenge that might be encountered is associated with the importance of the criteria through the evaluation and benchmarking phases despite their conflict. In addition, this conflict between the criteria poses a significant challenge during the evaluation stage [50]. A suitable procedure for this kind of objectives needs to be developed whilst boosting the significance of a certain evaluation criterion and minimising others [51]. Two major key points must be considered. The first one is to achieve a sufficient understanding of the COVID-19 classification technique behaviour whilst assigning certain significance to the design. The next point is the evaluation approach whilst bearing in mind the issue of trade-off. However, a conflict might exist between the opinions of the evaluator and the objective of the developer, which poses an effect over the last evaluation of the needed approach [52]. From a technical point of view, the COVID-19 classification technique by means of evaluation and benchmarking simultaneously considers multiple criteria and then assign a suitable weight for all evaluation criteria of the COVID-19 classification technique. After making a comparison for all scores of the approach, the approaches with the most balancing rate should be assigned with the highest priority level, whereas the approaches with the least balancing rate should be assigned with lowest priority level. In addition, because COVID-19 classification techniques have to consider multiple criteria, it considered as a difficult and challenging task in time and error rate in the dataset which also could be significantly important in the COVID-19 classification. In addition, each decision-maker assigns a different weight for all these previous criteria [53]. On the other hand, the experts who are in charge of assigning a score for the COVID-19 classification techniques could assign more weights to different features aside from the ones that acquire less interest than any other criteria. By contrast, experts who aim to make use of benchmarking method in order to address such problems would consider different criteria as the most significant ones.

6. Research Proposal for Potential Future Direction

This section describes the potential future direction of the process of evaluation and benchmarking the COVID-19 classification techniques used in medical image detection. According to the future challenges discussed, such process could face a multi-complex attribute problem; like that all the AI techniques are considered available alternatives to be a suitable technique. Therefore, adapting candid and structured techniques for decisions using multiple criteria could boost the decision-making quality. Aside from analysis, assessment and ranking, multi-criteria decision analysis (MCDA) is considered a solution that aids decision-makers to organise and solve any problem [54, 55].

6.1 Definition and Significance of MCDA

MCDA is defined as ‘an extension of decision theory that covers any decision with multiple objectives. MCDA is a methodology for assessing alternatives on individual, often conflicting criteria, and combining them into one overall appraisal' [56]. The techniques of decision-making are widely recognised, and amongst them, MCDA is the most significant. It is also considered as an important part of operation research that handles problems of decision-making with respect to decision criteria [57]. The technique involves various processes including structuring, planning and solving different decision problems with the use of many criteria [58]. MCDA is increasingly being used as it can promote the decision quality [59]. It is achieved by making the process of the decision more reasonable, efficient, clear and explicit compared with other traditional processes [60, 61]. The most significant goals of MCDA include the allocation of the data miner to choose the most suitable alternatives, assigning a rank to the alternatives in decreasing order with regard to the efficiency and classifying the applicable alternatives amongst groups of available alternatives [62, 63]. On this basis, the ranking will take place on the most suitable alternative(s) [64]. There is a need for fundamental terms in MCDA to be defined, in addition to containing the decision matrix (DM) and its associated criteria [65, 66]. An evaluation matrix

contains *n* attribute and *m* alternatives, which need identification [67, 68]. The intersection of both criteria and alternatives is defined as z_ij. Therefore, we have a matrix (z_ij) _ (m*n) explained as follows:

$$DM = \begin{array}{c} \\ Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \dots & X_n \end{array} \\ \begin{bmatrix} z_{11} & z_{12} & \dots & z_{1n} \\ z_{21} & z_{22} & \dots & z_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ z_{m1} & z_{m2} & \dots & z_{\mathrm{mn}} \end{bmatrix} \end{array},$$

where $Y_1, Y_2, \dots, Y_m$ are probable alternatives, which decision-makers need to rank (i.e. COVID-19 classification AI techniques). $X_1, \mathrm{X}, \dots, X_n$ are the criteria against which the performance of each alternative is evaluated. Finally, z_{ij} is the rating of alternative Y_i with respect to criterion X_j. There is an improvement possibility for the decision-making process by means of comprising decision-makers and stakeholders, which will enable the process with support and structure [69, 70]. With the use of candid, the structure of multi-criteria decision methods can aid towards improving the decision-making quality and set of techniques [71, 72]. These techniques could identify which of the criteria are relevant and provide information for evaluating the current alternatives [73]. By performing this process, they are able to improve transparency, consistency and decision validity [74]. MCDA can contribute to fair, transparent and rational priority-setting processes [75]. MCDA has been widely used in many areas for different applications [76]. MCDA works by means of ranking and finding the suitable solution to select appropriate alternatives in different domains [77-82], especially in healthcare domain [83-85].

6.2 MCDA Methods

Several MCDA methods can be found in the literature, including the analytic hierarchy process (AHP), weighted product method, hierarchical adaptive weighting, best–worst method, multiplicative exponential weighting, weighted sum model, simple additive weighting, analytic network process, VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR), technique for reorganisation of opinion order to interval levels and technique for order of preference by similarity to ideal solution (TOPSIS). Each technique uses different representations [86-89]. The diversity of MCDA techniques raises a challenge in terms of the selection of the most suitable method for a single scenario. Each technique has its own limitations and strengths [90, 91]. Therefore, selecting the most appropriate MCDA technique is highly important. According to our analysis, all the presented methods in the literature were not used for evaluation and benchmarking of COVID-19 medical image classification over AI techniques. These methods are challenged by non-adoption requirement-driven approach, which makes them unsuitable for measurement and scoring in decision-making [76, 89]. However, for cases that involve numerous alternatives and criteria, TOPSIS and VIKOR are applicable. VIKOR and TOPSIS are convenient to use when the given data are quantitative or objective. TOPSIS can create a shortest distance solution towards the ideal solution and also the largest distance away from the negative-ideal solution. Nevertheless, there is no consideration for the relative significance of these distances [92]. On the other hand, VIKOR has functional relationship to discrete-alternative problems. TOPSIS and VIKOR are considered the most practical techniques in solving real-world problems. The advantage of TOPSIS and VIKOR is that they can rapidly decide the best alternative. Furthermore, they are suitable techniques for cases where there are many alternatives and criteria situations [92]. Nevertheless, the major drawback of TOPSIS and VIKOR is the lack of provisioning for elicitation of weight and checking for judgment consistency [92]. Thus, TOPSIS and VIKOR need an effective technique to acquire the relative importance of various criteria with respect to the objective, and AHP is able to provide such a technique. However, AHP is utilised for setting objective weights on the preferences of the stakeholder [93], and it is restricted majorly by the human capacity for information processing. Therefore, 7 ± 2 would be the comparison ceiling [94]. The latest trend in MCDA techniques integrates two or more techniques to compensate for the drawbacks of single techniques. AHP and VIKOR are commonly used MCDA approaches in various studies and especially in the medical domain [58]. To evaluate and benchmark AI classification techniques used in the detection of COVID-19 medical images, the present study recommends to integrate AHP for assigning weights for the evaluation criteria of each classification type subjectively by relying on the judgment of experts, and VIKOR is needed to offer a comprehensive ranking of COVID-19 AI classification techniques.

7. Methodology

This section describes and explains the evaluation and benchmarking methodology of AI classification techniques used in COVID-19 medical image detection. Figure 4 illustrates all elements of our study in the overall proposed methodology.

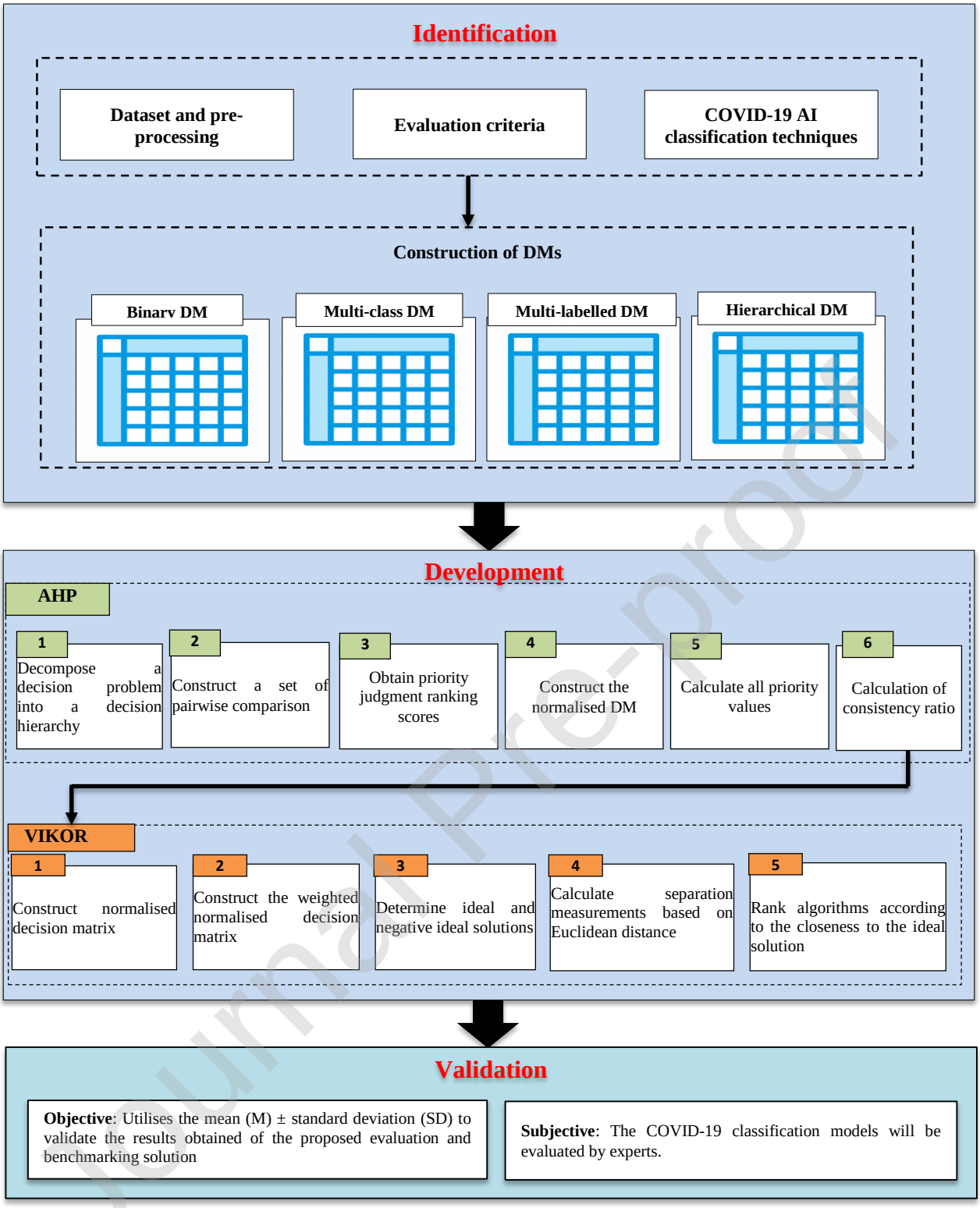

Figure 4. Proposed methodology for the evaluation and benchmarking of binary, multi-class, multi-labelled and hierarchical classification of COIVID-19 AI classification techniques

According to the proposed methodology, three phases have been performed for evaluating and benchmarking the COVID-19 AI classification techniques. The first phase is identification, which illustrates the datasets and required pre-processing and identifies the evaluation criteria used in the evaluation and benchmarking of COVID-19 AI classification techniques and the number and type of techniques. The output of this phase are four DMs, one for each classification type. In the second phase, integration of MCDA methods is presented. The AHP method is used to weigh the evaluation criteria subjectively, and the VIKOR method is used for

benchmarking AI classification techniques. In the third phase, objective and subjective validations are illustrated for ranking COVID-19 AI classification techniques. Further details are provided in the following subsections.

7.1 Identification Phase

In this phase, four main stages are conducted. First, the dataset and required pre-processing procedure (presented in Section 7.1.1) are identified. Second, the evaluation criteria within each type of classification (presented in Section 7.1.2) are identified. Third, the number and type of COVID-19 AI classification techniques are described in Section 7.1.3. Fourth, the construction of the four types of DMs based on identified elements is described in Section 7.1.4.

7.1.1 Dataset and Pre-processing

In this step, three main portions should be defined, namely, target dataset, required pre-processing technique for dataset and most suitable features for classification task [95-99]. Different COVID-19 datasets can be found in the literature. Some are based on X-ray images [34], whilst others are based on CT scan images [33]. Each dataset has some limitations. For example, the number of training samples is small, the provided images are of low quality, and the size of the images is not equal. Thus, pre-processing steps (e.g. using data augmentation [34] techniques to generate more medical image samples in order to provide a comprehensive training) are needed to tackle such issues. Furthermore, because COVID-19 can overlap with other pneumonia cases, image segmentation [35] can be used to define the region of interest as a pre-processing step for further analysis of COVID-19 cases. The features extracted from images have a great impact on classification [100] in terms of improving accuracy and minimising error rate, over-fitting and under-fitting issues [101, 102]. Thus, all mentioned scenarios will have a great impact on the results of evaluation and benchmarking for COVID-19 classification techniques. Accordingly, three steps should be provided to achieve an efficient evaluation and benchmarking process for COIVID-19 classification over AI techniques. To train and test COVID-19 classification techniques, the dataset will be separated into two parts. The first part will be used towards training the set, whereas the second part will be used for testing the set.

7.1.2 Evaluation Criteria Definition

As mentioned before, each classification type has its own evaluation criteria. Accordingly, in this section, the criteria within each classification type are identified, which will involve DMs. As mentioned in the critical review and analysis section, classification tasks are divided into four types, namely, binary, multi-class, multi-labelled and hierarchical. On the basis of each classification task, the evaluation criteria of COVID-19 AI classification techniques are listed in Table 2.

Table 2. Evaluation criteria of binary, multi-class, multi-labelled and hierarchical AI classification techniques

Binary classification		
Evaluation criteria	**Formula**	**Description**
Accuracy	$\frac{tp+tn}{tp+fn+fp+tn}$	Overall effectiveness of a classifier
Precision	$\frac{tp}{tp+fp}$	Class agreement of the data labels with the positive labels given by the classifier
Recall (sensitivity)	$\frac{tp}{tp+fn}$	Effectiveness of a classifier to identify positive labels
F score	$\frac{(\beta^2+1)tp}{(\beta^2+1)tp+\beta^2 fn+fp}$	Relations between data positive labels and those given by a classifier
Specificity	$\frac{tn}{fp+tn}$	How effectively a classifier identifies negative labels
AUC	$\frac{1}{2}(\frac{tp}{tp+fn}+\frac{tn}{tn+fp})$	Classifier's ability to avoid false classification
Multi-class classification		
Evaluation criteria	**Formula**	**Description**
Average accuracy	$\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$	Average per-class effectiveness of a classifier
Error rate	$\frac{\sum_{i=1}^{l}\frac{fp_i+fn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$	Average per-class classification error
Precisionμ	$\frac{\sum_{i=1}^{l}tp_i}{\sum_{i=1}^{l}(tp_i+fp_i)}$	Agreement of the data class labels with those of classifiers if calculated from the sums of per-sample decisions
Recallμ	$\frac{\sum_{i=1}^{l}tp_i}{\sum_{i=1}^{l}(tp_i+fn_i)}$	Effectiveness of a classifier to identify class labels if calculated from the sums of per-sample decisions
F scoreμ	$\frac{(\beta^2+1)\text{Precision}_\mu\text{Recall}_\mu}{\beta^2\text{Precision}_\mu+\text{Recall}_\mu}$	Relations between data positive labels and those given by a classifier based on the sums of per-sample decisions

Precision_M	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$	Average per-class agreement of the data class labels with those of a classifier
Recall_M	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l}$	Average per-class effectiveness of a classifier to identify class labels
Fscore_M	$\frac{(\beta^2+1)\text{Precision}_M\text{Recall}_M}{\beta^2\text{Precision}_M + \text{Recall}_M}$	Relations between data positive labels and those given by a classifier based on a per-class average
Multi-labelled classification		
Evaluation criteria	**Formula**	**Description**
Exact match ratio	$\frac{\sum_{i=1}^{n} I(L_i^d = L_i^d)}{n}$	Average per-sample exact classification
Labelling F score	$\frac{\sum_{i=1}^{n} \frac{2\sum_{j=1}^{l} L_i^c[j]L_i^d[j]}{\sum_{j=1}^{l}(L_i^c[j] + L_i^d[j])}}{n}$	Average per-sample classification with partial matches
Retrieval F score	$\frac{\sum_{j=1}^{l} \frac{2\sum_{i=1}^{n} L_i^c[j]L_i^d[j]}{\sum_{i=1}^{n}(L_i^c[j] + L_i^d[j])}}{l}$	Average per-class classification with partial matches
Hamming loss	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{l} I(L_i^c[j] \neq L_i^d[j])}{nl}$	Average per-example per-class total error
Hierarchical classification		
Evaluation criteria	**Formula**	**Description**
Precision↓	$\frac{\lvert C_\downarrow^c \cap C_\downarrow^d \rvert}{\lvert C_\downarrow^c \rvert}$	Positive agreement on subclass labels with regard to the subclass labels given by a classifier
Recall↓	$\frac{\lvert C_\downarrow^c \cap C_\downarrow^d \rvert}{\lvert C_\downarrow^d \rvert}$	Positive agreement on subclass labels with regard to the subclass labels given by data
Fscore↓	$\frac{(\beta^2+1)\text{Precision}_\downarrow\text{Recall}_\downarrow}{\beta^2\text{Precision}_\downarrow + \text{Recall}_\downarrow}$	Relations between data positive subclass labels and those given by a classifier
Precision↑	$\frac{\lvert C_\uparrow^c \cap C_\uparrow^d \rvert}{\lvert C_\uparrow^c \rvert}$	Positive agreement on superclass labels with regard to the superclass labels given by a classifier
Recall↑	$\frac{\lvert C_\uparrow^c \cap C_\uparrow^d \rvert}{\lvert C_\uparrow^d \rvert}$	Positive agreement on superclass labels with regard to the superclass labels given by data
Fscore ↑	$\frac{(\beta^2+1)\text{Precision}_\uparrow\text{Precision}_\uparrow}{\beta^2\text{Precision}_\uparrow + \text{Precision}_\uparrow}$	Relations between data positive superclass labels and those given by a classifier

tp = true positive, **tn** = true negative, **fp** = false positive, **fn** = false negative, **AUC** = area under the curve, **μ** = micro-averaging, **M** = macro-averaging, **I** = indicator function, $\mathbf{L_i}$ = set of class labels, $\boldsymbol{C_\downarrow^c}$ = **subclasses of class C,** $\boldsymbol{C_\uparrow^c}$ = subclasses assigned by a classifier, $\boldsymbol{C_\downarrow^d}$ = **data class labels**, $\boldsymbol{C_\uparrow^d}$= **data class labels**

A shown in Table 2, the evaluation criteria for COVID-19 classification techniques are different in terms of the number and calculation procedures within each type of classification. For example, binary classification has eight evaluation criteria, multi-class and hierarchical classifications have six criteria each, whereas multi-labelled classification has four criteria. Furthermore, as shown in Table 2, the precision criteria in the binary type are different from the criteria of precisionμ in multi-class type because the formulas for the two types are different, and other criteria belong to a single classification type. The usage of criteria depends on the target of classification (binary, multi-class, multi-labelled and hierarchical). Thus, different numbers and types of criteria will be involved in a particular DM of each classification task.

7.1.3 AI Classification Techniques

In this step, the number and type of COVID-19 AI classification techniques are identified, which will be included in each DM type. In general, different types of COVID-19 classification techniques can be found in the literature. Some studies are based on traditional machine learning classification techniques (e.g. [33]). On the other hand, the majority of classification tasks are based on deep learning techniques (e.g. [7, 34, 36]). However, the classification techniques that belong to a similar type (e.g. traditional machine learning and deep learning techniques) need to be included for the evaluation and benchmarking process. Furthermore, the number of candidate classification techniques should be defined in the evaluation and benchmarking scenario. As mentioned in Section 7.1.1, the dataset is divided into training and testing sets. However, each individual instance is supposed to belong to a predefined class [18, 98, 103, 104]. In the testing portion, if the classification technique performance looks 'acceptable', then the classification technique can be used to classify future data for which the class label is unknown. Ultimately, the classification technique that provides an acceptable result can be considered an 'acceptable technique'. Furthermore, for more reliable classification techniques, the difference ratio between the performance of the technique in the training and validation stages in terms of accuracy and loss function is very important to avoid over-fitting and under-fitting issues.

7.1.4 DM Construction

DM considers the main component in the proposed methodology of evaluation and benchmarking of AI classification techniques used in COVID-19 medical images. DM is composed of decision alternatives and identified criteria. In our case, the classification techniques for COVID-19 are the decision alternatives, and the criteria are identified evaluation criteria based on each classification task. As mentioned earlier, the AI domain has four types of classification tasks (binary, multi-class, multi-labelled and hierarchical). Each type has its own evaluation criteria; thus, each type should have a unique DM based on the distinction of the evaluation criteria. In this study, the DMs of COVID-19 medical image classifications will be constructed based on four different types, namely, binary DM, multi-class DM, multi-labelled DM and hierarchical DM. The DM data of specific classification type are generated from the crossover between the number of COVID-19 AI classification techniques and the number of specific classification type evaluation criteria as follows.

A. Binary DM: This DM is constructed on the basis of the intersection between decision alternatives (i.e. set of COVID-19 AI classification techniques) and six evaluation criteria (i.e. accuracy, precision, recall [sensitivity], F score, specificity, area under the curve) as presented in Table 3.

Table 3. DM of COVID-19 AI binary classification techniques

Evaluation criteria **AI COVID-19 classification techniques**	**Accuracy**	**Precision**	**Recall (sensitivity)**	**F score**	**Specificity**	**Area under the curve**
Technique 1	Av(T1/TS)	Pv (T1/TS)	Rv (T1/TS)	FSv (T1/TS)	Sv (T1/TS)	AUCv (T1/TS)
Technique 2	Av (T2/TS)	Pv (T2/TS)	Rv (T2/TS)	FSv (T2/TS)	S (T2/TS)	AUCv (T2/TS)
	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Technique *n*	Av(Tn/TS)	Pv (Tn/TS)	Rv (Tn/TS)	FSv (Tn/TS)	Sv (Tn/TS)	AUCv (Tn/TS)

T = classification technique; Av = accuracy value; Pv = precision value; Rv = recall (sensitivity) value; FSv = F score value; Sv = specificity value; AUCv = area under the curve value; TS = test samples; n: number of AI classification techniques

B. Multi-class DM: This DM is constructed on the basis of the intersection between decision alternatives (i.e. set of COVID-19 AI classification techniques) and eight evaluation criteria (i.e. average accuracy, error rate, precisionμ, recallμ, F scoreμ, precisionM, recallM, F scoreM) as presented in Table 4.

Table 4. DM of COVID-19 AI multi-class classification techniques

Evaluation criteria **COVID-19 AI classification techniques**	**Average accuracy**	**Error rate**	**Precisionμ**	**Recallμ**	**F scoreμ**	**PrecisionM**	**RecallM**	**F scoreM**
Technique 1	AAv (M1/TS)	ERv (M1/TS)	Pμv (M1/TS)	Rμv (M1/TS)	FSμv (M1/TS)	P_{MV} (M1/TS)	R_{MV} (M1/TS)	FS_{MV} (M1/TS)
Technique 2	AAv (M2/TS)	ERv (M2/TS)	Pμv (M2/TS)	Rμv (M2/TS)	FSμv (M2/TS)	P_{MV} (M2/TS)	R_{MV} (M2/TS)	FS_{MV} (M2/TS)
	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
Technique *n*	AAv (Mn/TS)	ERv (Mn/TS)	Pμv (Mn/TS)	Rμv (Mn/TS)	FSμv (Mn/TS)	P_{MV} (Mn/TS)	R_{MV} (Mn/TS)	FS_{MV} (Mn/TS)

T = classification technique; AAv = average accuracy value; ERv = error rate value; Pμv = precisionμ value; Rμv = recallμ value; FSμv = F scoreμ value; P_{MV} = $precision_M$ value; R_{MV} = $recall_M$ value; FS_{MV} = F $score_M$ value; TS = test samples; n: number of AI classification techniques

C. Multi-labelled DM: This DM is constructed on the basis of the intersection between decision alternatives (i.e. set of COVID-19 AI classification techniques) and four evaluation criteria (i.e. exact match ratio, labelling F score, retrieval F score and Hamming loss) as presented in Table 5.

Table 5. DM of COVID-19 AI multi-labelled classification techniques

Evaluation criteria				
COVID-19 AI classification techniques	**Exact match ratio**	**Labelling F score**	**Retrieval F score**	**Hamming loss**
Technique 1	EMv (M1/TS)	LFv (M1/TS)	RFv (M1/TS)	HLv (M1/TS)
Technique 2	EMv (M2/TS)	LFv (M2/TS)	RFv (M2/TS)	HLv (M2/TS)
	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Technique *n*	EMv (Mn/TS)	LFv (Mn/TS)	RFv (Mn/TS)	HLv (Mn/TS)

T = classification technique; **EMv** = exact match ratio value; **LFv** = labelling F score value; **RFv** = retrieval F score value; **HLv** = Hamming loss value; **TS** = test samples; **n**: number of AI classification techniques

D. Hierarchical DM: This DM is constructed on the basis of the intersection between decision alternatives (i.e. set of COVID-19 AI classification techniques) and six evaluation criteria (i.e. precision↓, recall↓, F score↓, precision↑, recall↑ and F score ↑) as presented in Table 6.

Table 6. DM of COVID-19 AI hierarchical classification techniques

Evaluation criteria						
COVID-19 AI classification techniques	**Precision↓**	**Recall↓**	**F score↓**	**Precision↑**	**Recall↑**	**F score ↑**
M1	P↓v (M1/TS)	R↓v (M1/TS)	FS↓v (M1/TS)	P↑v (M1/TS)	P↑v (M1/TS)	FS↑v (M1/TS)
M2	P↓v (M2/TS)	R↓v (M2/TS)	FS↓v (M2/TS)	P↑v (M2/TS)	P↑v (M2/TS)	FS↑v (M2/TS)
	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Mn	P↓v (Mn/TS)	R↓v (Mn/TS)	FS↓v (Mn/TS)	P↑v (Mn/TS)	P↑v (Mn/TS)	FS↑v (Mn/TS)

T = classification technique; P↓v = precision↓ value; R↓v = recall↓ value; FS↓v = F score↓value; P↑v = precision↑ value; R↑v = recall↑ value; FS↑v = F score ↑value; TS = test samples; n: number of AI classification techniques

However, the data within the four DMs represent the values of the evaluation of each COVID-19 AI classification technique based on the identified evaluation criteria of each classification task. Practically, on the basis of these constructed DMs, three evaluation and benchmarking challenges will be generated and encountered in the future (i.e. multi-criteria, trade-off amongst the criteria and important criteria), as highlighted in Section 5. The evaluation and benchmarking of AI classification techniques used in COVID-19 medical images is considered a complex MCDA problem. To this end, the development of decision-making approach is important to preclude the evaluation and benchmarking problem complexity.

7.2 Development Phase

To develop a methodology of evaluation and benchmarking of AI classification techniques used in COVID-19 medical image detection, integration of MCDA methods is presented. Such development is based on AHP method for subjective weighting of identified evaluation criteria within each constructed DM as presented in Section 7.2.1 and VIKOR method for benchmarking and selecting best alternatives (i.e. COVID-19 AI classification techniques) in the constructed DMs as presented in Section 7.2.2.

7.2.1 AHP Weighting Method

This stage presents the process of assigning suitable weights to the multi-evaluation criteria within each DM subjectively based on the AHP method. The AHP approach involves several steps, which are applicable for any AI classification type of COVID-19 medical image detection. The procedure of AHP includes the following steps [56].

Step 1:

The problem is modelled as a hierarchy to start the AHP approach. The hierarchy contains the decision goal and the criteria that must be designed. Pairwise comparison amongst the criteria in the DM of each classification

type is conducted to obtain the weights subjectively. Examples of pairwise comparison for three criteria are illustrated in Figure 5.

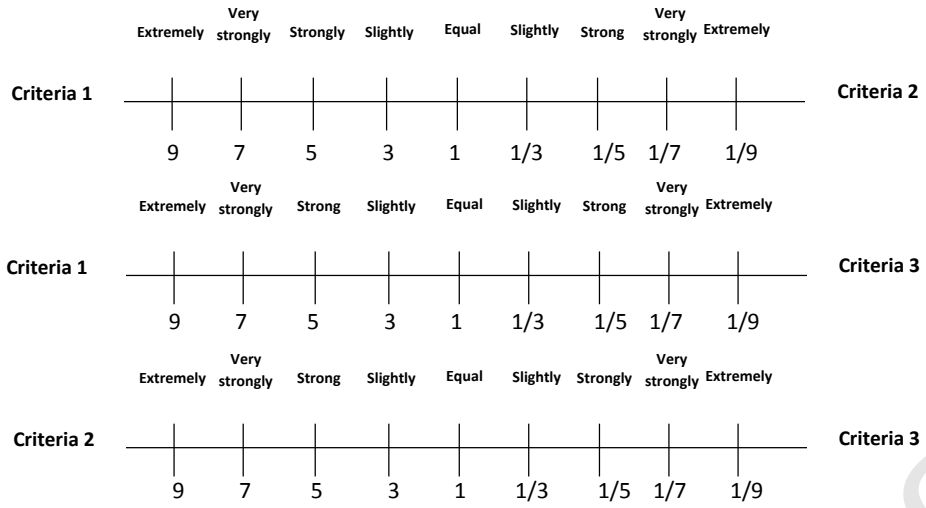

Figure 5. Pairwise comparison example

Step 2:

The AHP builds pairwise matrix comparison in Equation (1) to determine a weighting decision:

$$A = \begin{pmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & \cdots & x_{nn} \end{pmatrix}, \qquad (1)$$

where $x_{ii} = 1, x_{ji} = \frac{1}{x_{ij}}$.

Step 3:

This stage involves the design of a pairwise comparison questionnaire within each type of classification and distributes it to the experts. However, in this step, the number of experts included in the questionnaire should be defined. The target experts are those who have relevant experience with a case study, besides enough period of experience in the same domain. Their preferences and judgments on the evaluation criteria of each classification type used in AHP were evaluated.

Step 4:

In this step, each element in matrix A (1) is normalised to construct the normalised matrix A_{norm}, A_{norm} (aij) as follows:

$$a_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} xij}, \qquad (2)$$

$$A_{norm} = \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & a_{nn} \end{pmatrix}, \qquad (3)$$

where A(x_{ij}) is given by Equation (2).

Step 5:

This step includes AHP pairwise comparison to utilise mathematical calculations, convert judgments and assign weights for each criterion of each AI classification type. The weights of the decision criterion can be calculated using Equation (4):

$$W_i = \sum_{j=1}^{n} a_{ij} \,/n \text{ and } \sum_{j=1}^{n} W_i = 1, \qquad (4)$$

where n is the number of compared evaluation criteria of each COVID-19 AI classification type.

Step 6:
In this step, Equation (5) is utilised to check the consistency ratio (CR) to the pairwise comparison matrix as follows:

$$CR = \frac{CI}{RI}. \quad (5)$$

The consistency index (CI) is calculated using Equation (6) as follows:

$$CI = \frac{\lambda\, max - n}{n-1}, \quad (6)$$

where $\lambda\, max$ is the maximum eigenvalue of the judgement matrix. Random CI (RI) is calculated using Equation (7) as follows:

$$RI = \frac{1.98(n-1)}{n}.CI. \quad (7)$$

A pairwise comparison matrix with a corresponding CR of no more than 10% or 0.1 is acceptable; otherwise it will be ignored.

7.2.2 VIKOR Benchmarking Method
To start with the benchmarking of COVID-19 AI classification techniques, the VIKOR method is utilised considering its suitability for such purpose. In addition, it can provide rapid results and determine which option is the most appropriate one. The COVID-19 AI classification techniques can be benchmarked and ranked according to the VIKOR method using the obtained criteria weights from the AHP method. The VIKOR approach involves several steps [105, 106].

Step 1:
Identify the best $f^{*}i$ and worst $f^{-}i$ values of all criteria within each DM, i = 1; 2; ...; n. If the ith function represents:

A benefit criterion (the larger the better):

$$f_i^{*}=\max_j f_{ij,}\ f_i^{-}=\min_j f_{ij,} \quad (8)$$

A cost criterion (the smaller the better):

$$f_i^{*}=\min_j f_{ij,}\ f_i^{-}=\max_j f_{ij,}. \quad (9)$$

Step 2:
AHP is considered for the computation of each criterion weight. A set of weights $w = w_1, w_2, w_3, \cdots, w_j, \cdots, w_n$ from the decision-maker is accommodated in the DM; this set is equal to 1. The resulting matrix can also be computed as demonstrated in the following equation:

$$WM = wi * \frac{f_i^{*}-fij}{f_i^{*}-f_i^{-}}. \quad (10)$$

A weighted matrix is generated as follows:

$$\begin{bmatrix} \frac{w_1(f_i^{*}-f11\ )}{f_i^{*}-f_i^{-}} & \cdots & \frac{w_i(f_i^{*}-fij\ )}{f_i^{*}-f_i^{-}} \\ \frac{w_1(f_i^{*}-f21\ )}{f_i^{*}-f_i^{-}} & \cdots & \frac{w_i(f_i^{*}-fij\ )}{f_i^{*}-f_i^{-}} \\ \vdots & \vdots & \vdots \\ \frac{w_1(f_i^{*}-f31\ )}{f_i^{*}-f_i^{-}} & \cdots & \frac{w_i(f_i^{*}-fij\ )}{f_i^{*}-f_i^{-}} \end{bmatrix}. \quad (11)$$

Step 3:

Compute the *Sj* and *Rj* values, *j=1,2,3,….,J, i=1,2,3,…,n* by using the following equations:

$$Sj = \sum_{i=1}^{n} wi * \frac{f_i^* - fij}{f_i^* - f_i^-}, \quad (12)$$

$$Rj = \max_i wi * \frac{f_i^* - fij}{f_i^* - f_i^-}, \quad (13)$$

where wi is the weight of criteria expressing their relative importance.

Step 4: Compute the values of Qj, $j = (1,2,\cdots,J)$ using the following relation:

$$Q_j = \frac{v(S_j - S^*)}{S^- - S^*} + \frac{(1-v)(R_j - R^*)}{R^- - R^*}, \quad (14)$$

where

$S^* = min_j S_j, S^- = max_j S_j,$

$R^* = min_j R_j, R^- = max_j R_j.$

v is introduced as the weight of the strategy of 'the majority of criteria' (or 'the maximum group utility'); here, v = 0.5.

Step 5:
Now the alternative set (i.e. COVID-19 AI classification techniques) can be benchmarked. This process is accomplished by sorting the R and Q values in ascending order. The lowest value indicates the optimal performance.

Step 6:
Propose the alternative (a') as a compromise solution. It ranks the best by the measure Q (minimum) if two conditions are satisfied. The conditions are as follows:

R1. 'Acceptable advantage'

$$Q(a'') - Q(a') \geq DQ, \quad (15)$$

where (a'') is the alternative in the second position in the ranking list by Q, DQ = 1/(J−1) and J is the number of alternatives.

R2. 'Stability' is acceptable with the decision-making context. Alternative a' should also be the best as ranked by S and/or R. This compromise solution is stable within the decision-making process, which could be a 'voting by majority rule' (v > 0:5), 'by consensus' (v ≅ 0.5) or 'with veto' (v < 0.5). Here, v is the decision-making strategy weight of 'the majority of criteria' (or 'the maximum group utility').

7.3 Validation Phase

This phase presents the process of objective (Section 7.3.1) and subjective (Section 7.3.2) validations for the results of benchmarking COVID-19 AI classification techniques. Further details are explained in the following subsections.

7.3.1 Objective Validation

The results of the proposed methodology will be validated by utilising an objective approach as similar to [107]. To validate the results of the ranking with the use of the previous test, the COVID-19 AI classification techniques will be divided into (n) groups on the basis of the ranking results, which were acquired from the proposed methodology. Every group consists of a number of selected COVID-19 AI classification techniques. The number of techniques within each group varies depending on various scenarios. The validation result will not be influenced by the number of groups or AI classification techniques within each group.
To make sure that the benchmarking results of COVID-19 AI classification techniques are valid, this study utilises two statistical approaches: mean and standard deviation. The mean ± standard deviation can be calculated for each group of data and is used to ensure that the set of COVID-19 AI classification techniques is subjected to systematic ordering.

The mean is the average result. It is calculated by performing a deviation of the sum of the observed results over the result numbers with the use of the following equation:

$$mean = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad (16)$$

Standard deviation is used to determine the dispersion or variation amount in the set of values and is calculated as follows:

$$Standard\ deviation = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}. \qquad (17)$$

For example, let us consider that we have four groups with (n) number of COVID-19 AI classification techniques for each group. In this scenario, the first group must reach the best value, and that has to be proven when the standard deviation and the mean are measured. We assumed that the first group acquired the best in both standard deviation and the mean compared with the other three groups. However, for the second group, its results for the mean and standard deviation have to be poorer than those in the first group and better than those in the third and fourth groups or have to be equal to those in the third group. Accordingly, for the systematic ranking results, the first group must prove that it is the best compared with the other groups.

7.3.2 Subjective Validation

This section describes the subjective validation process. The COVID-19 AI classification techniques will be evaluated by specialist experts in AI classification of medical cases. The experts can prove the effectiveness of the benchmarking results of COVID-19 AI classification techniques obtained by our proposed decision-making approach by examining the values of all evaluation criteria used.

8. Conclusion

The COVID-19 pandemic has a tremendous impact on the life of people around the world, and the number of infected patients has considerably increased. COVID-19 quickly gained a foothold, and nations, governments and scholars are attempting to address this worldwide crisis. Different medical tests are used in the detection of COVID-19. Several studies have used X-rays and CT scans to support and reveal anomalies indicative of COVID-19. CT scan and X-ray tests are utilised as initial detection tools to evaluate the severity of COVID-19, monitor the emergency conditions of patients and predict disease progression. The growing developments of AI techniques have led to the challenges of choosing evaluation and benchmarking AI techniques and which technique is suitable for the diagnosis and classification of COVID-19 medical images. Thus, this study presented a systematic review of AI techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking. The results showed that only 11 studies utilised AI techniques in detecting and classifying COVID-19 with different case studies. However, this study proved that the process of evaluating and benchmarking of AI classification techniques (i.e. binary, multi-class, multi-labelled and hierarchical classifications), which could be used in the detection and diagnosis of COVID-19 medical image, is a critical gap of related literature. The challenges of such gap are discussed, and the process of evaluation and benchmarking of COVID-19 AI classification techniques is considered a multi-complex attribute problem. Thus, using MCDA is essential. As a potential future research direction, this study provided a detailed methodology for the evaluation and benchmarking of AI classification techniques used in the detection of COVID-19 medical images. Such methodology is presented on the basis of three sequential phases (i.e. identification, development and validation).

References

[1] S. Kooraki, M. Hosseiny, L. Myers, and A. J. J. o. t. A. c. o. r. Gholamrezanezhad, "Coronavirus (COVID-19) outbreak: what the department of radiology should know," 2020.

[2] Y. Wang, Y. Wang, Y. Chen, and Q. J. J. o. m. v. Qin, "Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures," vol. 92, no. 6, pp. 568-576, 2020.

[3] T. Ai *et al.*, "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases," p. 200642, 2020.

[4] Y. Fang *et al.*, "Sensitivity of chest CT for COVID-19: comparison to RT-PCR," p. 200432, 2020.
[5] H. Zeng *et al.*, "Antibodies in infants born to mothers with COVID-19 pneumonia," 2020.
[6] D. A. J. A. o. p. Schwartz and l. medicine, "An analysis of 38 pregnant women with COVID-19, their newborn infants, and maternal-fetal transmission of SARS-CoV-2: maternal coronavirus infections and pregnancy outcomes," 2020.
[7] T. Ozturk *et al.*, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," p. 103792, 2020.
[8] M. Li *et al.*, "Coronavirus disease (COVID-19): spectrum of CT findings and temporal progression of the disease," 2020.
[9] F. Shi *et al.*, "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19," 2020.
[10] H. S. Maghdid, A. T. Asaad, K. Z. Ghafoor, A. S. Sadiq, and M. K. J. a. p. a. Khan, "Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms," 2020.
[11] W. H. Organization, "Coronavirus disease 2019 (COVID-19): situation report, 72," 2020.
[12] D. D. Miller and E. W. J. T. A. j. o. m. Brown, "Artificial intelligence in medical practice: the question to the answer?," vol. 131, no. 2, pp. 129-133, 2018.
[13] K.-H. Yu, A. L. Beam, and I. S. J. N. b. e. Kohane, "Artificial intelligence in healthcare," vol. 2, no. 10, pp. 719-731, 2018.
[14] A. S. Albahri *et al.*, "Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review," *Journal of Medical Systems,* vol. 44, no. 7, p. 122, 2020/05/25 2020.
[15] Z. Yang *et al.*, "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," vol. 12, no. 3, p. 165, 2020.
[16] M. Alsalem *et al.*, "Systematic review of an automated multiclass detection and classification system for acute Leukaemia in terms of evaluation and benchmarking, open challenges, issues and methodological aspects," vol. 42, no. 11, p. 204, 2018.
[17] M. Alsalem *et al.*, "Multiclass benchmarking framework for automated acute Leukaemia detection and classification based on BWM and group-VIKOR," *Journal of medical systems,* vol. 43, no. 7, p. 212, 2019.
[18] A. Zaidan *et al.*, "Multi-agent learning neural network and Bayesian model for real-time IoT skin detectors: a new evaluation and benchmarking methodology," pp. 1-52, 2019.
[19] A. Zaidan *et al.*, "A review on smartphone skin cancer diagnosis apps in evaluation and benchmarking: coherent taxonomy, open issues and recommendation pathway solution," vol. 8, no. 4, pp. 223-238, 2018.
[20] Q. M. Yas, A. Zaidan, B. Zaidan, B. Rahmatullah, and H. A. J. M. Karim, "Comprehensive insights into evaluation and benchmarking of real-time skin detectors: Review, open issues & challenges, and recommended solutions," vol. 114, pp. 243-260, 2018.
[21] M. Talal *et al.*, "Smart home-based IoT for real-time and secure remote health monitoring of triage and priority system using body sensors: Multi-driven systematic review," vol. 43, no. 3, p. 42, 2019.
[22] A. Mohsin *et al.*, "Blockchain authentication of network applications: Taxonomy, classification, capabilities, open challenges, motivations, recommendations and future directions," vol. 64, pp. 41-60, 2019.
[24] O. Albahri *et al.*, "Systematic review of real-time remote health monitoring system in triage and priority-based sensor technology: Taxonomy, open challenges, motivation and recommendations," vol. 42, no. 5, p. 80, 2018.
[25] A. Mohsin *et al.*, "Real-time remote health monitoring systems using body sensor information and finger vein biometric verification: A multi-layer systematic review," vol. 42, no. 12, p. 238, 2018.

[26] O. Zughoul *et al.*, "Comprehensive insights into the criteria of student performance in various educational domains," vol. 6, pp. 73245-73264, 2018.
[27] A. Mohsin *et al.*, "Real-time medical systems based on human biometric steganography: A systematic review," vol. 42, no. 12, p. 245, 2018.
[28] A. A. Zaidan *et al.*, "A survey on communication components for IoT-based technologies in smart homes," vol. 69, no. 1, pp. 1-25, 2018.
[29] A. Mohsin *et al.*, "Based medical systems for patient's authentication: Towards a new verification secure framework using CIA standard," vol. 43, no. 7, p. 192, 2019.
[30] A. Mohsin *et al.*, "Finger Vein Biometrics: Taxonomy Analysis, Open Challenges, Future Directions, and Recommended Solution for Decentralised Network Architectures," vol. 8, pp. 9821-9845, 2020.
[31] M. L. Shuwandy, B. Zaidan, A. Zaidan, and A. J. J. o. m. s. Albahri, "Sensor-based mHealth authentication for real-time remote healthcare monitoring system: A multilayer systematic review," vol. 43, no. 2, p. 33, 2019.
[32] D. Li *et al.*, "False-negative results of real-time reverse-transcriptase polymerase chain reaction for severe acute respiratory syndrome coronavirus 2: role of deep-learning-based CT diagnosis and insights from two cases," vol. 21, no. 4, pp. 505-508, 2020.
[33] L. A. Wallis, "COVID-19 Severity Scoring Tool for low resourced settings," *African Journal of Emergency Medicine,* 2020.
[34] F. Ucar and D. Korkmaz, "COVIDiagnosis-Net: Deep Bayes-SqueezeNet based Diagnostic of the Coronavirus Disease 2019 (COVID-19) from X-Ray Images," *Medical Hypotheses,* p. 109761, 2020.
[35] Y. Oh, S. Park, and J. C. Ye, "Deep Learning COVID-19 Features on CXR using Limited Training Data Sets," *arXiv preprint arXiv:2004.05758,* 2020.
[36] M. Toğaçar, B. Ergen, Z. J. C. i. B. Cömert, and Medicine, "COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches," p. 103805, 2020.
[37] A. J. T. L. D. H. Laghi, "Cautions about radiologic diagnosis of COVID-19 infection driven by artificial intelligence," vol. 2, no. 5, p. e225, 2020.
[38] B. J. T. L. D. H. McCall, "COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread," vol. 2, no. 4, pp. e166-e167, 2020.
[39] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management,* vol. 45, no. 4, pp. 427-437, 2009.
[40] R. M. Pereira, D. Bertolini, L. O. Teixeira, C. N. Silla Jr, Y. M. J. C. M. Costa, and P. i. Biomedicine, "COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios," p. 105532, 2020.
[41] M. Abdel-Basset, R. Mohamed, M. Elhoseny, R. K. Chakrabortty, and M. J. I. A. Ryan, "A hybrid COVID-19 detection model using an improved marine predators algorithm and a ranking-based diversity reduction strategy," 2020.
[42] L. A. J. A. J. o. E. M. Wallis, "COVID-19 Severity Scoring Tool for low resourced settings," 2020.
[43] F. Ucar and D. J. M. H. Korkmaz, "COVIDiagnosis-Net: Deep Bayes-SqueezeNet based Diagnostic of the Coronavirus Disease 2019 (COVID-19) from X-Ray Images," p. 109761, 2020.
[44] Y. Oh, S. Park, and J. C. J. I. T. o. M. I. Ye, "Deep learning covid-19 features on cxr using limited training data sets," 2020.
[45] S. Y. Yerima, S. Sezer, and I. Muttik, "High accuracy android malware detection using ensemble learning," *IET Information Security,* vol. 9, no. 6, pp. 313-320, 2015.

[46] M. Lindorfer, M. Neugschwandtner, and C. Platzer, "MARVIN: Efficient and Comprehensive Mobile App Classification through Static and Dynamic Analysis," in *2015 IEEE 39th Annual Computer Software and Applications Conference*, 2015, vol. 2, pp. 422-433.

[47] A. Shastry, M. Kantarcioglu, Y. Zhou, and B. Thuraisingham, "Randomizing Smartphone Malware Profiles against Statistical Mining Techniques," Berlin, Heidelberg, 2012, pp. 239-254: Springer Berlin Heidelberg.

[48] H. Kurniawan, Y. Rosmansyah, and B. Dabarsyah, "Android anomaly detection system using machine learning classification," in *2015 International Conference on Electrical Engineering and Informatics (ICEEI)*, 2015, pp. 288-293.

[49] Y. Wei, Z. Hanlin, G. Linqiang, and R. Hardy, "On behavior-based detection of malware on Android platform," in *2013 IEEE Global Communications Conference (GLOBECOM)*, 2013, pp. 814-819.

[50] M. Ilangkumaran, V. Sasirekha, L. Anojkumar, and M. Boopathi Raja, "Machine tool selection using AHP and VIKOR methodologies under fuzzy environment," *International Journal of Modelling in Operations Management,* vol. 2, no. 4, pp. 409-436, 2012.

[51] H. E. Aktan and P. K. Samut, "Agricultural performance evaluation by integrating fuzzy AHP and VIKOR methods," *International Journal of Applied Decision Sciences,* vol. 6, no. 4, pp. 324-344, 2013.

[52] R. L. Keeney and H. Raiffa, *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press, 1993.

[53] M. Oliveira, D. B. Fontes, and T. Pereira, "Multicriteria decision making: a case study in the automobile industry," 2013.

[54] A. Jadhav and R. Sonar, "Analytic hierarchy process (AHP), weighted scoring method (WSM), and hybrid knowledge based system (HKBS) for software selection: a comparative study," in *2009 Second International Conference on Emerging Trends in Engineering & Technology*, 2009, pp. 991-997: IEEE.

[55] K. Mohammed *et al.*, "Real-time remote-health monitoring systems: a review on patients prioritisation for multiple-chronic diseases, taxonomy analysis, concerns and solution procedure," vol. 43, no. 7, p. 223, 2019.

[56] O. Albahri *et al.*, "Fault-tolerant mHealth framework in the context of IoT-based real-time wearable health data sensors," vol. 7, pp. 50052-50080, 2019.

[57] J. Malczewski, *GIS and multicriteria decision analysis*. John Wiley & Sons, 1999.

[58] A. Albahri *et al.*, "Based multiple heterogeneous wearable sensors: A smart real-time health monitoring structured for hospitals distributor," vol. 7, pp. 37269-37323, 2019.

[59] M. Talal *et al.*, "Comprehensive review and analysis of anti-malware apps for smartphones," vol. 72, no. 2, pp. 285-337, 2019.

[60] S. Zionts, "MCDM-If not a Roman Numeral, then what?," *Interfaces,* vol. 9, no. 4, pp. 94-101, 1979.

[61] M. Khatari, A. Zaidan, B. Zaidan, O. Albahri, and M. J. I. J. I. T. D. M. Alsalem, "Multi-criteria evaluation and benchmarking for active queue management methods: Open issues challenges and recommended pathway solutions," vol. 18, no. 4, pp. 1187-1242, 2019.

[62] E. Almahdi, A. Zaidan, B. Zaidan, M. Alsalem, O. Albahri, and A. J. J. o. m. s. Albahri, "Mobile patient monitoring systems from a benchmarking aspect: Challenges, open issues and recommended solutions," vol. 43, no. 7, p. 207, 2019.

[63] E. Almahdi, A. Zaidan, B. Zaidan, M. Alsalem, O. Albahri, and A. J. J. o. m. s. Albahri, "Mobile-based patient monitoring systems: A prioritisation framework using multi-criteria decision-making techniques," vol. 43, no. 7, p. 219, 2019.

[64] N. Napi *et al.*, "Medical emergency triage and patient prioritisation in a telemedicine environment: a systematic review," pp. 1-22, 2019.

[65] M. Whaiduzzaman, A. Gani, N. B. Anuar, M. Shiraz, M. N. Haque, and I. T. Haque, "Cloud service selection using multicriteria decision analysis," *The Scientific World Journal,* vol. 2014, 2014.

[66] N. Ibrahim *et al.*, "Multi-Criteria Evaluation and Benchmarking for Young Learners' English Language Mobile Applications in Terms of LSRW Skills," vol. 7, pp. 146620-146651, 2019.

[67] M. Alaa *et al.*, "Assessment and ranking framework for the English skills of pre-service teachers based on fuzzy Delphi and TOPSIS methods," vol. 7, pp. 126201-126223, 2019.

[68] C. Lim, K. Tan, A. Zaidan, B. J. M. T. Zaidan, and Applications, "A proposed methodology of bringing past life in digital cultural heritage through crowd simulation: a case study in George Town, Malaysia," vol. 79, no. 5, pp. 3387-3423, 2020.

[69] K. Mohammed *et al.*, "Novel technique for reorganisation of opinion order to interval levels for solving several instances representing prioritisation in patients with multiple chronic diseases," vol. 185, p. 105151, 2020.

[70] N. Kalid, A. Zaidan, B. Zaidan, O. H. Salman, M. Hashim, and H. J. J. o. m. s. Muzammil, "Based real time remote health monitoring systems: A review on patients prioritization and related" big data" using body sensors information and communication technology," vol. 42, no. 2, p. 30, 2018.

[71] O. Albahri, A. Zaidan, B. Zaidan, M. Hashim, A. Albahri, and M. J. J. o. m. s. Alsalem, "Real-time remote health-monitoring Systems in a Medical Centre: A review of the provision of healthcare services-based body sensor information, open challenges and methodological aspects," vol. 42, no. 9, p. 164, 2018.

[72] F. Jumaah, A. Zadain, B. Zaidan, A. Hamzah, and R. J. M. Bahbibi, "Decision-making solution based multi-measurement design parameter for optimization of GPS receiver tracking channels in static and dynamic real-time positioning multipath environment," vol. 118, pp. 83-95, 2018.

[73] N. Kalid *et al.*, "Based on real time remote health monitoring systems: a new approach for prioritization "large scales data" patients with chronic heart diseases using body sensors and communication technology," vol. 42, no. 4, p. 69, 2018.

[74] A. Albahri, A. Zaidan, O. Albahri, B. Zaidan, and M. J. J. o. m. s. Alsalem, "Real-time fault-tolerant mHealth system: Comprehensive review of healthcare services, opens issues, challenges and methodological aspects," vol. 42, no. 8, p. 137, 2018.

[75] O. Enaizan *et al.*, "Electronic medical record systems: Decision support examination framework for individual, security and privacy concerns using multi-perspective analysis," pp. 1-28, 2018.

[76] M. Aruldoss, T. M. Lakshmi, and V. P. Venkatesan, "A survey on multi criteria decision making methods and its applications," *American Journal of Information Systems,* vol. 1, no. 1, pp. 31-43, 2013.

[77] H. AlSattar *et al.*, "MOGSABAT: A metaheuristic hybrid algorithm for solving multi-objective optimisation problems," pp. 1-15, 2018.

[78] B. N. Abdullateef, N. F. Elias, H. Mohamed, A. Zaidan, and B. J. S. Zaidan, "An evaluation and selection problems of OSS-LMS packages," vol. 5, no. 1, p. 248, 2016.

[79] Q. M. Yas, A. Zadain, B. Zaidan, M. Lakulu, B. J. I. J. o. P. R. Rahmatullah, and A. Intelligence, "Towards on develop a framework for the evaluation and benchmarking of skin detectors based on artificial intelligent models using multi-criteria decision-making techniques," vol. 31, no. 03, p. 1759002, 2017.

[80] B. Zaidan, A. Zaidan, H. A. Karim, N. J. S. P. Ahmad, and Experience, "A new digital watermarking evaluation and benchmarking methodology using an external group of evaluators and multi-criteria analysis based on 'large-scale data'," vol. 47, no. 10, pp. 1365-1392, 2017.

[81] B. Zaidan, A. J. J. o. C. Zaidan, Systems, and Computers, "Software and hardware FPGA-based digital watermarking and steganography approaches: Toward new methodology for

evaluation and benchmarking using multi-criteria decision-making techniques," vol. 26, no. 07, p. 1750116, 2017.

[82] B. Zaidan, A. Zaidan, H. Abdul Karim, N. J. I. J. o. I. T. Ahmad, and D. Making, "A new approach based on multi-dimensional evaluation and benchmarking for data hiding techniques," pp. 1-42, 2017.

[83] A. Zaidan, B. Zaidan, A. Al-Haiqi, M. L. M. Kiah, M. Hussain, and M. Abdulnabi, "Evaluation and selection of open-source EMR software packages based on integrated AHP and TOPSIS," *Journal of biomedical informatics,* vol. 53, pp. 390-404, 2015.

[84] A. Zaidan, B. Zaidan, M. Hussain, A. Haiqi, M. M. Kiah, and M. J. D. S. S. Abdulnabi, "Multi-criteria analysis for OS-EMR software selection problem: A comparative study," vol. 78, pp. 15-27, 2015.

[85] O. H. Salman, A. Zaidan, B. Zaidan, Naserkalid, M. J. I. J. o. I. T. Hashim, and D. Making, "Novel methodology for triage and prioritizing using “big data” patients with chronic heart diseases through telemedicine environmental," vol. 16, no. 05, pp. 1211-1245, 2017.

[86] F. Jumaah, A. Zaidan, B. Zaidan, R. Bahbibi, M. Qahtan, and A. J. T. S. Sali, "Technique for order performance by similarity to ideal solution for solving complex situations in multi-criteria optimization of the tracking channels of GPS baseband telecommunication receivers," vol. 68, no. 3, pp. 425-443, 2018.

[87] B. Rahmatullah, A. Zaidan, F. Mohamed, and A. Sali, "Multi-complex attributes analysis for optimum GPS baseband receiver tracking channels selection," in *2017 4th international conference on control, decision and information technologies (CoDIT)*, 2017, pp. 1084-1088: IEEE.

[88] B. Zaidan and A. J. M. Zaidan, "Comparative study on the evaluation and benchmarking information hiding approaches based multi-measurement analysis using TOPSIS method with different normalisation, separation and context techniques," vol. 117, pp. 277-294, 2018.

[89] K. I. Mohammed, et al., "A Uniform Intelligent Prioritisation for Solving Diverse and Big Data Generated from Multiple Chronic Diseases Patients based on Hybrid Decision-Making and Voting Method " *IEEE Access,* 16-5-2020 2020.

[90] B. N. Abdullateef, N. F. Elias, H. Mohamed, A. Zaidan, and B. Zaidan, "An evaluation and selection problems of OSS-LMS packages," *SpringerPlus,* vol. 5, no. 1, p. 248, 2016.

[91] İ. Kaya, M. Çolak, and F. Terzi, "Use of MCDM techniques for energy policy and decision-making problems: A review," *International Journal of Energy Research,* vol. 42, no. 7, pp. 2344-2372, 2018.

[92] S. Opricovic and G.-H. Tzeng, "Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS," *European journal of operational research,* vol. 156, no. 2, pp. 445-455, 2004.

[93] H. Nilsson, E.-M. Nordström, and K. Öhman, "Decision support for participatory forest planning using AHP and TOPSIS," *Forests,* vol. 7, no. 5, p. 100, 2016.

[94] T. L. Saaty and M. S. Ozdemir, "Why the magic number seven plus or minus two," *Mathematical and computer modelling,* vol. 38, no. 3-4, pp. 233-244, 2003.

[95] A. A. Zaidan, H. A. Karim, N. N. Ahmad, B. B. Zaidan, and A. Sali, "An automated anti-pornography system using a skin detector based on artificial intelligence: A review," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 27, no. 04, p. 1350012, 2013.

[96] A. Zaidan, H. Abdul Karim, N. N. Ahmad, B. Zaidan, A. J. I. J. o. P. R. Sali, and A. Intelligence, "A four-phases methodology to propose anti-pornography system based on neural and Bayesian methods of artificial intelligence," vol. 28, no. 01, p. 1459001, 2014.

[97] A. Zaidan, N. N. Ahmad, H. A. Karim, M. Larbani, B. Zaidan, and A. J. N. Sali, "On the multi-agent learning neural and Bayesian methods in skin detector and pornography classifier: An automated anti-pornography system," vol. 131, pp. 397-418, 2014.

[98] A. Zaidan, N. N. Ahmad, H. A. Karim, M. Larbani, B. Zaidan, and A. J. E. A. o. A. I. Sali, "Image skin segmentation based on multi-agent learning Bayesian and neural network," vol. 32, pp. 136-150, 2014.

[99] A. A. Zaidan, "Anti-pornography algorithm based on multi-agent learning in skin detector and pornography classifier," Multimedia University (Malaysia), 2013.

[100] M. Alsalem *et al.*, "Systematic review of an automated multiclass detection and classification system for acute Leukaemia in terms of evaluation and benchmarking, open challenges, issues and methodological aspects," *Journal of medical systems,* vol. 42, no. 11, p. 204, 2018.

[101] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis,* vol. 42, pp. 60-88, 2017.

[102] D. ping Tian, "A review on image feature extraction and representation techniques," *International Journal of Multimedia and Ubiquitous Engineering,* vol. 8, no. 4, pp. 385-396, 2013.

[103] A. Zaidan, H. A. Karim, N. Ahmad, B. Zaidan, M. M. J. J. o. C. Kiah, Systems, and Computers, "Robust pornography classification solving the image size variation problem based on multi-agent learning," vol. 24, no. 02, p. 1550023, 2015.

[104] A. Zaidan and B. J. A. I. R. Zaidan, "A review on intelligent process for smart home applications based on IoT: coherent taxonomy, motivation, open challenges, and recommendations," vol. 53, no. 1, pp. 141-165, 2020.

[105] K. H. Abdulkareem, et al. , "A Novel Multi-Perspective Benchmarking Framework for Selecting Image Dehazing Intelligent Algorithms Based on BWM and Group VIKOR Techniques," *International Journal of Information Technology & Decision Making,* 2020.

[106] K. H. Abdulkareem, et al., "A new standardisation and selection framework for real-time image dehazing algorithms from multi-foggy scenes based on fuzzy Delphi and hybrid multi-criteria decision analysis methods," *Neural Computing and Applications,* 2020.

[107] M. Qader, B. Zaidan, A. Zaidan, S. Ali, M. Kamaluddin, and W. J. M. Radzi, "A methodology for football players selection problem based on multi-measurements criteria analysis," vol. 111, pp. 38-50, 2017.

[108] A. S. Albahri et al ., Multi-biological Laboratory Examination Framework for the Prioritisation of Patients with COVID-19 Based on Integrated AHP and Group VIKOR Methods,International Journal of Information Technology & Decision Making, 2020.

[109] Fayiz Faiez et al., Novel Multi-perspective Hiring Framework for the Selection of Software Programmer Applicants Based on AHP and Group TOPSIS Techniques, International Journal of Information Technology & Decision Making (IJITDM), 2020.

[110] R.T.Mohammed et al., Review of the Research Landscape of Multi-criteria Evaluation and Benchmarking Processes for Many-objective Optimisation Methods: Coherent Taxonomy, Challenges and Recommended Solution, International Journal of Information Technology & Decision Making (IJITDM), 2020.